Universality beyond the classical asymptotic regime

Kevin Han Huang 黄瀚

Gatsby Computational Neuroscience Unit

University College London

A dissertation submitted in partial fulfilment

of the requirements for the degree of

Doctor of Philosophy

of

University College London

11 April 2025

Declaration and funding support

I, Kevin Han Huang, confirm that the work presented in my thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

I thank the Gatsby Charitable foundation for its generous financial support throughout my PhD.

Abstract

A typical learning problem involves training an estimator $f(X_1, \ldots, X_n)$ on some data set $X_1, \ldots, X_n$. Gaussian universality is the observation that, for many potentially complicated estimators, properties of the estimator are preserved if the training data are substituted by appropriately chosen Gaussian distributions. This unlocks a wide range of empirical and theoretical tools for analysing the trained estimator, since Gaussian distributions are both analytically tractable and computationally fast to simulate. Universality results have been observed in statistical physics, random matrix theory and other branches of probability; in recent papers, they have been theoretically and/or empirically established for several high-dimensional models across statistics and machine learning (ML). One crucial question is the extent to which universality may hold under high dimensionality and dependence.

To address this, this thesis develops Gaussian universality results for a general class of estimators of high-dimensional data, with nearly matching upper and lower bounds. The results cover any f well-approximated by strictly monotone functions of polynomials, whose degree grows not too fast with respect to the sample size n. No explicit requirements are imposed on the number of data dimensions with respect to n. Together with the fourth moment phenomenon of Nualart and Peccati (2005), our results imply necessary and sufficient conditions for the asymptotic normality of approximately polynomial estimators.

The remainder of this thesis focuses on how universality results can recover, extend and establish new high-dimensional analyses across statistics and machine learning. These include: (i) a complete distributional characterisation of high-dimensional U-statistics used for kernel-based testing via a moment ratio; (ii) a high-dimensional delta method; (iii) a finite-sample approximation of subgraph count statistics that recover known geometric conditions; (iv) characterising the unexpected effects of dependence under the popular ML practice of data augmentation; (v) analysis of optimisation algorithms found in ML and AI for Science.

Impact Statement

This thesis focuses on the theory of Gaussian universality, which extends the classical probability result of central limit theorem to more general algorithms and estimation methods. The main contributions include the development of universality results for high-dimensional data and for block dependence, a tight characterisation of a range of settings to which universality applies, and the applications of universality results to provide theoretical and practical intuitions on various statistical and machine learning algorithms.

In probability, statistics and machine learning theory, Gaussian approximations are routinely used for simplifying analysis of complex models, understanding hyperparameter choices and providing consistency and uncertainty guarantees for algorithms. One direct impact of this work is the provision of several general tools and recipes, which extend known results, that can be used to obtain new understanding for various practical models and algorithms. This thesis also contains concrete applications where such analysis is performed, which offer additional understanding on the effects of data augmentation, the behaviour of high-dimensional hypothesis testing, the prediction performance of an optimisation problem, the stability of training algorithms and so on. These provide additional mathematical insights that may aid practitioners in their day-to-day choice of algorithms and models.

In critical real-world applications such as finance, medicine, scientific discovery and social policy research, classical statistical results such as the central limit theorem are pivotal in providing uncertainty quantification, robustness guarantees and bias controls. This thesis provides several extensions of such results to the setting with high-dimensional and dependent data, which commonly arise in the modern era of "big data". Such extensions are vital, both for ensuring the validity and safety of the algorithms used in these domains, and for identifying and rectifying possible points of failure with a rigorous mathematical guidance.

Acknowledgements

This thesis would not have been possible without the support of many incredible people.

I would like to extend my most sincere gratitude to my advisors, Peter Orbanz and Morgane Austern, for their guidance, patience, trust, encouragement and unwavering support throughout this journey. Starting a PhD in the middle of a covid lockdown could be much harder than it turned out to be – and I would have not enjoyed Zoom as much – if not for the weekly dosage of banters, maths, and banters again with my two advisors. Peter never fails to dig out something interesting in the mumble jumble I throw on his board, and Morgane always spots where an inequality can get loose before I even finish my first sentence. Intellectually, Peter and Morgane have stretched me in many different ways that I could not possibly enumerate. On a more personal level, both of them have been my constant source of support and courage in the many ups and downs throughout my PhD. It was truly inspiring to be guided by people who care deeply about maths and science, and I am beyond grateful to have them as my advisors.

I would also like to express my gratitude towards my thesis examiners, Gesine Reinert and Sam Livingstone. The many useful and interesting discussions during the viva process have both been very helpful for the final form of this thesis and for inspiring further intellectual conversations hopefully to be explored.

I would like to thank everyone at the Gatsby Unit and the Gatsby Charitable Foundation for the greatest research environment one could possibly ask for. I was once worried about joining a theoretical neuroscience and machine learning department as someone who primarily does maths and statistics, and I could not be more wrong. Among the many people, I would like to thank Arthur Gretton for the many fun corridor chats, his helpful insights on U-statistics and kernels, and the various other forms of support he has kindly given me. I am also very grateful to everyone in Arthur's group who have provided me with an immense amount of support, especially in the earlier years of my PhD as the only non-Arthur ML student in the building. I would also like to thank Maneesh Sahani for his patience and guidance during my neuroscience rotation, despite my evident lack of knowledge about the brain. I am also grateful to Peter Latham for reassuring me that everything in the world should be Gaussian and for repeatedly affirming me of my research. A special thanks to I-Chun, Barry and Mike for keeping the Gatsby family together and for being a reliable source of support throughout my PhD.

My PhD life would have been very different without my amazing office mates and year mates. At the intersection of these two sets are Pierre Glaser and Antonin Schrab: Pierre essentially introduced me to stochastic optimisation, and Antonin taught me a lot about testing. Above all, they have filled the past couple of years with fun, wisdom, rapport and quite a fair amount of Frenchness, and I am extremely fortunate to have walked the PhD journey with them. I would also like to thank Heishiro Kanagawa, Hugh Dance, Dimitri Meunier, Wenkai Xu, Antoine Moulin, Clémentine Dominé, Will Dorrell, Rodrigo Carrasco-Davis, Tom George, Hudson Chen, Michael Li and Liang Zhou for the constant support and the many fun and deep conversations, and for making me at home at Gatsby. I am also very grateful to Max Hird for his friendship and world-class fritters, for introducing me to the fun community that is the statistics department, and for spending an incredible amount of time nerding out about probability theory together with me.

I am deeply indebted to my collaborators and the people who have supported me in research groups within and beyond Gatsby. I would like to thank especially Ryan P. Adams for hosting me in his group at Princeton, and for getting me excited about aspects of computer science and physics that I would have never considered otherwise. Among the many other people, I would like to thank: Lee Gunderson, Gecia Bravo-Hermsdorff, Hugh, Vasco Portilheiro and Vince Velkey in Peter's group for the fun math discussions, rapport and camaraderie; Peter Vincent, Samo Hrodmaka, Marcel Nonnenmacher, Cong Sun and many others in Gatsby and the Sainsbury Wellcome Centre, for the insightful conversations and importantly for semi-successfully educating me about neuroscience; Xing Liu for being one of my dearest friends and amazing collaborators; Terry Soo, Alex Watson, Sam Livingstone, Codina Cotar and many others for making me feel welcome in UCL Statistics; Elif Ertekin and Jenny Ni Zhan for the patience with my non-existent materials science knowledge and for teaching me everything about electrons; Cindy Zhang, Jenny, Alex Guerra and many others in Ryan's group, as well as (and especially) Atharv Joshi, for going out of their way to make me feel welcome at Princeton; Matthew Esmaili, Haoyu Ye, Tianle Liu and Qizhao Chen in Morgane's group for making my one physical visit and many virtual visits to Harvard incredibly enjoyable. I am also indebted to Yudong Chen, Mengchu Li, Yi Yu, Sergio Bacallado, Richard Samworth, and many others I had the fortune to know in Cambridge, for the kindness, support and advice I have constantly received since my undergraduate years and throughout my PhD journey.

I am forever grateful to my family and friends for always being there for me during the ups and downs of my academic and personal life. My deepest gratitude goes to my family in Shantou, 曼华, 建生, 华, 泓 and 乐乐, and my extended family in Singapore, 蓓蓓, 德硕, 可乐 and 谷雨, for their unconditional love and trust that have supported me for more than a decade's time away from home. I am beyond grateful to Kangyu Wang, Xinpeng Wang, Yiqing Zhao, Richard Zhipeng Wang, Huiyao Zheng, Albert Qiaochu

Jiang, Kristen Dilan Yang, Siyang Fu, Xi Li, Ziyu Wu, Vera Ren and Jason Li, among the many other people I had the fortune to be with in Cambridge and London, for making my life in the UK immeasurably enjoyable. They are always the ones to ensure that I stay well-fed and that I am promptly removed from my pile of math drafts to “get a life”, the importance of which cannot be overstated in the production of this thesis.

UCL research paper declaration

Chapters 3 to 7 are based on the works completed over the course of this thesis, for which I have retained the copyrights. I have contributed to every aspect of the content that is included in this thesis. The works with shared contributions are clarified with a note below and only aspects of the work that I have contributed to are included.

Sections 3.1, 3.2 and 3.4 of Chapter 3 are based on the publication

> **K. H. Huang**, X. Liu, A. Duncan, and A. Gandy. A high-dimensional convergence theorem for U-statistics with applications to kernel-based testing. In The Thirty Sixth Annual Conference on Learning Theory (COLT), pages 3827–3918. PMLR, 2023. Paper URL.

In this work, XL has contributed to moment computations for MMD and KSD under Gaussian mean-shifts, which are excluded from this thesis, and the experiments, which are briefly discussed in this thesis. Section 3.3 of Chapter 3 is based on the work

> **K. H. Huang** and P. Orbanz. Slow rates of approximation of U-statistics and V-statistics by quadratic forms of Gaussians. arXiv:2406.12437, 2024. Paper URL.

Chapters 4 and 5 are based on the work

> **K. H. Huang**, M. Austern and P. Orbanz. Gaussian universality for approximately polynomial functions of high-dimensional data. arXiv:2403.10711, 2024. Paper URL.

Chapter 6 is based on the work

> **K. H. Huang**, P. Orbanz and M. Austern. Gaussian and non-Gaussian universality of data augmentation. arXiv:2202.09134, 2022. Paper URL.

Section 7.1 of Chapter 7 is based on the work

> M. E. Mallory*, **K. H. Huang*** and M. Austern. Universality of high-dimensional logistic regression and a novel CGMT under dependence with applications to data augmentation. arXiv:2502.15752, 2025. (*equal contribution). Paper URL.

In this work, MEM has contributed to the universality results, which are excluded from this thesis. Section 7.2.1 of Chapter 7 is based on the publication

> P. Glaser, **K. H. Huang** and A. Gretton. Near-optimality of contrastive divergence algorithms. The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS), 2024. Paper URL.

In this work, PG has contributed to the ideation and most results, which are excluded from this thesis. KHH has contributed to the offline contrastive divergence results, which are only briefly discussed in this thesis via their connections to universality. Section 7.2.2 of Chapter 7 is based on the work

> **K. H. Huang**, N. Zhan, E. Ertekin, P. Orbanz and R. P. Adams. Diagonal symmetrization of neural network solvers for the many-electron Schrödinger equation. arXiv:2502.05318, 2025. Paper URL.

In this work, NZ has contributed to the experiments on lithium hydrides, which are excluded from this thesis.

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor / senior author unless this is not appropriate, e.g. if the paper was a single-author work):

Candidate: Kevin Han Huang

Date: 11 Apr 2025

Supervisor / Senior Author signature (where appropriate): Peter Orbanz

Date: 11 Apr 2025

Contents

Chapter 1

Introduction

A central premise of statistics and machine learning is the ability to learn from data. Given a finite set of observed data from some unknown mechanism μ, one seeks the best estimators and models, trained on these observations, that are capable of inferring various properties of μ and of making predictions about future observations from μ. Over the last decade, the complexity and scale of these learning algorithms and datasets have grown at an unprecedented pace, and have led to remarkable empirical successes across many fields of machine learning (ML) and applied statistics.

A solid theoretical understanding of these algorithms, however, remains difficult. This is in part due to the complex data and training regimes in which they operate, and in part due to the many heuristics they require to run effectively. Many well-established tools of statistical and ML theory are developed under the classical regime of assumptions, where data are typically low-dimensional and independently sampled, and where estimators take simple closed forms with a few learnable parameters. In contrast, modern statistical estimation and ML algorithms are usually defined implicitly through complex optimisation algorithms, operate on data and parameters that live in a high-dimensional space, and are heavily influenced by the engineering heuristics employed during training.

The gap between theory and practice leaves many practically important questions unsolved: How sensitive is my algorithm to specific hyperparameter choices? Do specific training heuristics help or hurt my models? How confident should I be about the answers returned by my models? These questions underline a broader set of theoretical properties we desire for our algorithms and models, such as uncertainty quantification, robustness and bias control (Abdar et al., 2021; Mehrabi et al., 2021; Freiesleben and Grote, 2023). The ability to obtain such theoretical guarantees is critical in high-stake applications e.g. finance, medicine and social policy research, where verifying the correctness of model outputs is expensive, difficult or ethically and legally challenging (Grimmer et al., 2021; Giovanola and Tiribelli, 2023; Blasco et al., 2024).

One of the many attempts to address this gap is the theoretical framework of *universality*, which has gained substantial interests across statistics and machine learning over the last decade. This thesis constitutes a modest effort to contribute to its development.

Specifically, the goal of this thesis is to develop theoretical characterisations and applications of universality in the modern regime of high-dimensional and dependent data, for estimators at various levels of complexity, and in pursuit of partial answers to some of the aforementioned practical questions.

To give a high-level introduction of universality, from now on, we treat estimators and algorithms as functions of the inherent randomness in the observed data. To be more concrete, we study objects of the form

$$f(X) \;\coloneqq\; f(X_1, \ldots, X_n)\,. \tag{1.1}$$

$X = (X_1, \ldots, X_n)$ are random observations taking values in a high-dimensional Euclidean space $\mathbb{R}^d$, where $d = d(n)$ is typically of a comparable size to n. The function $f : (\mathbb{R}^d)^n \to \mathbb{R}^q$ describes a chosen property of interest of a learning algorithm trained with X. Examples in this thesis range from estimators used in high-dimensional hypothesis testing, prediction risk of high-dimensional estimators, to stability estimate of gradients of large neural network models used in AI for physics.

Universality is the probabilistic phenomenon that, for many potentially complicated functions f, properties of (1.1) resemble those of the surrogate estimate

$$f(Z) \;\coloneqq\; f(Z_1, \ldots, Z_n)\,.$$

$Z = (Z_1, \ldots, Z_n)$ are random vectors that take substantially simpler forms than X. Such results are "universal" in that it typically holds for a wide class of data distributions of X, allowing for the influence of complicated datasets to be reduced to simple surrogates. When Z consists of Gaussian random vectors, we refer to this phenomenon as *Gaussian universality*. Most of this thesis concerns Gaussian universality (except for parts of Chapter 6), and we use the two terms interchangeably unless otherwise specified.

A special case of Gaussian universality is the celebrated central limit theorem (CLT). For i.i.d. univariate random variables $X_1, \ldots, X_n$, the CLT can be viewed as a universality result by taking f to be a rescaled empirical average:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \mathbb{E}[X_1]) \;\approx\; \frac{1}{\sqrt{n}}\sum_{i=1}^n (Z_i - \mathbb{E}[X_1]) \qquad \text{in distribution as } n \to \infty\,.$$

Z_i's are i.i.d. Gaussian variables with the same mean and variance as X_1, and one concludes that the above sum has a normal limit by noting that an average of Gaussians is again a Gaussian. The CLT approximation error can be further quantified at finite n by the celebrated Berry-Esseen theorem, which gives a finite number C such that

$$\left|\mathbb{P}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \mathbb{E}[X_1]) \le t\Big) - \mathbb{P}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n (Z_i - \mathbb{E}[X_1]) \le t\Big)\right| \le \frac{C}{n^{1/2}}\frac{\mathbb{E}|X_1|^3}{\mathrm{Var}[X_1]^{3/2}}\,.$$

The CLT was first developed by de Moivre (1733) for Bernoulli random variables X_i's

and extended to general X_i's by a series of work throughout the 19th and early 20th centuries. When first established, it confirmed the many numerical findings at its time about the emergence of Gaussian distribution as a "common error curve" for an average; see Fischer (2011) for a historical account. To date, the CLT has served as one of the most fundamental results used in statistics, probability and machine learning in both theoretical and applied domains. Normal approximations of averages are now routinely applied in hypothesis testing, confidence intervals, regression analysis, statistical modelling, optimisation algorithm analyses and, indeed, any general setting where one expects a sum of weakly dependent, low-dimensional and mildly well-behaved random variables.

One of the first efforts to develop universality results beyond averages are the works of Rotar (1976), Rotar et al. (1979), Mossel, O'Donnell, and Oleszkiewicz (2005, 2010) and Chatterjee (2006). A shared observation was that Lindeberg's swapping technique — used for proving the CLT in Lindeberg (1922); Trotter (1959) — can be extended to well-behaved multilinear polynomials as well as general functions with suitable stability properties. Meanwhile, the Malliavin-Stein method has also been shown as a powerful alternative to Lindeberg's technique for establishing universality in Gaussian variables (see e.g. Nourdin et al. (2010)). These immediately led to a body of fruitful universality results for spectral properties of large random matrices (Chatterjee, 2006; Tao and Vu, 2011, 2015; Wang and Paul, 2014; Wood, 2016; Basak et al., 2018). At the same time, the notion of universality has been developed and widely applied in many problems in statistical physics; see Kadanoff (1990) for a survey of the area. Building on these understandings, a wave of universality results started emerging for estimators found in communications and statistical learning (Korada and Montanari, 2011; Wen et al., 2012), statistical physics (Bayati et al., 2015; Caravenna et al., 2016) and high-dimensional statistics (Chernozhukov et al., 2013, 2017; Montanari and Nguyen, 2017; Dobriban and Liu, 2019). In parallel to the course of this thesis, universality results are being rapidly established theoretically and/or empirically for estimators found in machine learning: A non-exhaustive list includes random feature models (Hu and Lu, 2022), regularised regression (Han and Shen, 2023), generalised linear models (Dandi et al., 2023), perceptron models (Gerace et al., 2024), max-margin classifiers (Montanari et al., 2023), general classes of empirical risk minimisers (Montanari and Saeed, 2022), representations of data generated by generative adversarial networks (Seddik et al., 2020), student-teacher models (Loureiro et al., 2021; Pesce et al., 2023) and diffusion models (Ghane et al., 2025).

In almost all of the aforementioned results or their subsequent extensions, universality has proved to be successful in characterising various properties of the estimator $f(X)$. This is often no mean feat: Unlike the case of an empirical average, $f(Z)$ may still follow an intractable distribution. Numerically, $f(Z)$ is typically analysed by simulations with Gaussian variables, which are computationally fast to generate. Theoretically, further

analyses are made possible thanks to the wide range of tools developed specifically for addressing Gaussian data, such as the cavity method (Opper et al., 2001), approximate message passing method (Donoho et al., 2009), the replica method (Mézard et al., 1987) and the convex Gaussian min-max theorem (Gordon, 1985; Thrampoulidis et al., 2014). These techniques are grounded on the well-understood properties of Gaussian processes and Gaussian matrices, and are constantly evolving to adapt to new complicated settings.

In view of these developments, this thesis sets out to address the following questions:

(i) For what class of functions f can universality results of the form $f(X) \approx f(Z)$ be established, and are there examples where universality ceases to hold?

(ii) Is the finite-sample upper bound on the universality approximation, typically given by the Lindeberg method, improveable?

(iii) How do universality results behave under high-dimensionality, i.e. when dimension d of the data is comparable or large compared to n?

(iv) How do universality results behave under dependence, i.e. when $X_1, \ldots, X_n$ are not i.i.d.?

(v) How may we use universality to gain theoretical and practical insights on statistical and machine learning applications?

The core theoretical results of this thesis, which seek to address (i), (ii) and (iii), are presented in Chapter 4. Informally those results imply that, if $f \approx h \circ q$ is well-approximated by some strictly monotonic function h of some low-degree polynomial function q, universality applies, even when the dimension $d = d(n)$ of data is large relative to the number of samples n.

The rest of the thesis investigates (iii)–(v) across a diverse range of examples and applications in Chapters 3, 5, 6 and 7. These include estimators trained with the ML technique of data augmentation, U-statistics found in high-dimensional kernel-based testing, subgraph count statistics, plug-in estimators, ridge and ridgeless regressions, softmax ensemble estimator, the contrastive divergence algorithm (typically used for energy-based model training) and the variational Monte Carlo algorithm (used in large-scale neural network solvers to the many-body Schrödinger equation). Examples of the properties we analyse include variance and stability properties, consistency, training and test risk behaviours e.g. the double-descent risk curve of high-dimensional estimators, and other effects of hyperparameter choices.

1.1 Thesis outline and relation to the author's works

In the rest of this chapter, Section 1.2 clarifies notations and terminology. Chapter 2 reviews the Lindeberg method, a core proof technique of universality, and briefly discusses its comparison to other distributional approximation techniques. Chapter 8 concludes the thesis by discussing some ongoing developments in the literature and future directions.

The remaining chapters are based on works completed over the course of this thesis. Note that the chapters below are organised in the order of ease of presentation, rather than in the order of completion of the corresponding works.

Chapter 3 motivates the applicability of universality in high-dimensional analysis by studying the distributional approximation of a degree-two U-statistic. As a consequence, we establish how commonly used kernel-based test statistics can exhibit different asymptotics as a result of their hyperparameter choices. Most of Chapter 3 is based on the publication

> **K. H. Huang**, X. Liu, A. Duncan, and A. Gandy. A high-dimensional convergence theorem for U-statistics with applications to kernel-based testing. In The Thirty Sixth Annual Conference on Learning Theory (COLT), pages 3827–3918. PMLR, 2023,

except for tightness results on the error of approximation, which is based on the work

> **K. H. Huang** and P. Orbanz. Slow rates of approximation of U-statistics and V-statistics by quadratic forms of Gaussians. arXiv:2406.12437, 2024.

Chapter 4 develops a set of general universality results that characterise the class of functions for which universality holds. As direct applications, Chapter 5 presents a higher-order delta method with possibly non-Gaussian limits, and generalise a number of known results on high-dimensional and infinite-order U-statistics, and on fluctuations of subgraph counts. Both chapters are based on the work

> **K. H. Huang**, M. Austern and P. Orbanz. Gaussian universality for approximately polynomial functions of high-dimensional data. arXiv:2403.10711, 2024.

Chapter 6 analyses the effects of data augmentation, a ubiquitous machine learning technique, by developing universality results under block dependence and for estimators beyond polynomials. This enables us to show that variance reduction and regularisation, two effects commonly associated with data augmentation, can be nuanced and hyperparameter-dependent. This is based on the work

> **K. H. Huang**, P. Orbanz and M. Austern. Gaussian and Non-Gaussian universality of data augmentation. arXiv:2202.09134, 2022.

Chapter 7 considers the role of Gaussian universality in optimisation analysis, especially in the case where the object of interest may not admit a closed-form formula in terms of the data. Section 7.1 presents a convex Gaussian min-max theorem under dependence, which is useful for analysing the risks of a high-dimensional optimisation. This is developed as part of the joint work

> M. E. Mallory*, **K. H. Huang*** and M. Austern. Universality of high-dimensional logistic regression and a novel CGMT under dependence with applications to data augmentation. arXiv:2502.15752, 2025. (*equal contribution)

Section 7.2 discusses the implications of Gaussian universality in one-step and multi-step stability analyses of training algorithms employed in practical machine learning problems. The results presented are partially used in the publication

> P. Glaser, **K. H. Huang** and A. Gretton. Near-optimality of contrastive divergence algorithms. The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS), 2024,

as well as in the work

> **K. H. Huang**, N. Zhan, E. Ertekin, P. Orbanz and R. P. Adams. Diagonal symmetrization of neural network solvers for the many-electron Schrödinger equation. arXiv:2502.05318, 2025.

We emphasise that universality is not the key message of either work, and our focus will primarily be on highlighting the connection of universality to their analyses.

1.2 Notation and terminology

Asymptotics. Throughout this thesis, we use the asymptotic notations $o, O, \Theta, \omega, \Omega$ defined in the usual way (see e.g. Chapter 3 of Cormen et al. (2009)) under the limit $n \to \infty$. The dimension parameter $d = d(n)$ is always allowed to depend on n, and we omit the dependence on n for simplicity. With an abuse of notation, we write $n, d \to \infty$ to mean that the n-dependent sequence $d(n) \to \infty$ as $n \to \infty$. We also use the terminology *finite-sample bounds* to mean error bounds that hold for finite n and d, without taking asymptotics. The error bounds often involve unspecified numerical constants that do not depend on n, d or any other properties of the data or the estimators. These are referred to as *absolute constants* in our results.

Norms. We use $\|\bullet\|$ for the Euclidean norm, $\|\bullet\|_{op}$ for the matrix operator norm, $\|\bullet\|_{L_\nu} := \mathbb{E}[|\bullet|^\nu]^{1/\nu}$ for the L_ν norm and $\|\bullet\|_\infty$ as the infinity norm (typically for a function).

Terminology: Gaussian universality. We use the term Gaussian universality to refer to the approximation of a function of $(X_i)_{i\leq n}$ by the same function of Gaussian vectors $(Z_i)_{i\leq n}$; this matches the nomenclature of most of the literature surveyed in the introduction. Some texts—for example, in the context of Gaussian approximation of Wiener chaos—use the term instead to indicate that the overall function is asymptotically normal, see e.g. Chapter 11 of Nourdin and Peccati (2012). Since our approximation $f(Z)$ still involves n-dependent quantities, it may have Gaussian or non-Gaussian limits (Section 4.4).

Chapter 2

Brief review on the Lindeberg method

> This phenomenon, when the final asymptotic result proves to be insensitive to the fine details of the original problem, is known as universality.
>
> *Andrei Okounkov*
> *Symmetric functions and random partitions, 2003*

In this chapter, we briefly review the Lindeberg method (also known in the literature as Lindeberg's swapping technique or Lindeberg's principle) for proving upper bounds on a universality approximation. Some of the earliest developments of these results trace back to Rotar (1976), Rotar et al. (1979), Mossel, O'Donnell, and Oleszkiewicz (2005, 2010) and Chatterjee (2006), and we also refer interested readers to Van Handel (2014) for a more comprehensive introduction to the method.

The chapter is organised as follows. Section 2.1 presents a full walk-through of the Lindeberg method in the independent univariate case. Section 2.2 discusses possible modifications to the Lindeberg proof, which are formalised and used in the rest of the thesis for various applications.

2.1 The Lindeberg method for functions of independent univariate random variables

Let $X_1, X_2, \ldots$ be independent (not necessarily identically distributed) and mean-zero univariate random variables with finite third absolute moments. The surrogate variables are independent random variables $Z_1, Z_2, \ldots$ such that $Z_i \sim \mathcal{N}(0, \mathrm{Var}[X_i])$. Fix a thrice-differentiable function $f : \mathbb{R}^n \to \mathbb{R}$. The objective is to show that the two random variables

$$f(X) = f(X_1, \ldots, X_n) \qquad \text{and} \qquad f(Z) = f(Z_1, \ldots, Z_n)$$

are close to each other in distribution. This can be measured by, for example, the Kolmogorov distance, where the universality approximation result concerns a difference in cumulative distribution functions (c.d.f.):

$$\sup_{t\in\mathbb{R}} |\mathbb{P}(f(X) \le t) - \mathbb{P}(f(Z) \le t)| \xrightarrow{n\to\infty} 0 \,. \tag{2.1}$$

The Lindeberg method typically provides universality approximation with respect to an integral probability metric (Müller, 1997) on a class of smooth functions. Specifically, consider the following class of thrice-differentiable functions with bounded derivatives

$$\mathcal{H} \coloneqq \left\{ h : \mathbb{R} \to \mathbb{R} \,\middle|\, \|\partial h\|_\infty \le 1 \,,\, \|\partial^2 h\|_\infty \le 1 \,,\, \|\partial^3 h\|_\infty \le 1 \right\} ,$$

and consider the induced metric $d_{\mathcal{H}}$ on two probability measure μ and ν in $\mathbb{R}$ as

$$d_{\mathcal{H}}(\mu, \nu) \coloneqq \sup_{h\in\mathcal{H}} |\mathbb{E}_{U\sim\mu}[h(U)] - \mathbb{E}_{V\sim\nu}[h(V)]| \,.$$

For the rest of the thesis, with an abuse of notation, we also write the above interchangeably as

$$d_{\mathcal{H}}(U, V) \,,$$

which is taken as the probability metric $d_{\mathcal{H}}$ evaluated on the laws of U and V. Since appropriately rescaled elements of $\mathcal{H}$ approximate the indicator functions, one can show that convergence in $d_{\mathcal{H}}$ implies convergence in the Kolmogorov metric*. It therefore suffices to prove

$$d_{\mathcal{H}}(f(X), f(Z)) = \sup_{h\in\mathcal{H}} |\mathbb{E}[h \circ f(X)] - \mathbb{E}[h \circ f(Z)]| \xrightarrow{n\to\infty} 0 \,.$$

The key step of the Lindeberg method rests on building a discrete interpolation path from $h \circ f(X)$ to $h \circ f(Z)$†. This is done by considering the telescoping sum

$$h \circ f(X) - h \circ f(Z) = \sum_{i=1}^n \left(h \circ f \circ W_{-i}(X_i) - h \circ f \circ W_{-i}(Z_i) \right) , \tag{2.2}$$

where we have denoted the random function $W_{-i} : \mathbb{R} \to \mathbb{R}^n$ as

$$W_{-i}(y) \coloneqq (X_1, \ldots, X_{i-1}, y, Z_{i+1}, \ldots, Z_n) \,.$$

Each summand is a difference of functions that differs only in the i-th data point. Since $h \circ f \circ W_{-i}$ is thrice-differentiable by construction, we may perform a Taylor expansion with respect to X_i around $\mathbb{E}[X_i] = 0$ to obtain

$$\begin{aligned} &\left| \mathbb{E}\left[h \circ f \circ W_{-i}(X_i) - \partial_i (h \circ f)(W_{-i}(0))\, X_i - \frac{1}{2} \partial_i^2 (h \circ f)(W_{-i}(0))\, X_i^2 \right] \right| \\ &\le \frac{1}{6} \|\partial_i^3 (h \circ f)\|_\infty \, \mathbb{E}|X_i|^3 \,. \end{aligned} \tag{2.3}$$

*More details can be found in Section 6.3 as well as the proof of Theorem 4.1 in Appendix C.2, with the exact choice of h given in Lemma A.10. Also see Section 2.2 for a brief discussion

†For Lindeberg method with a continuous interpolation path, see e.g. Montanari and Saeed (2022).

The same argument also applies to Z_i. This allows us to substitute $h \circ f \circ W_{-i}(X_i)$ and $h \circ f \circ W_{-i}(Z_i)$ by their second-order Taylor approximations, and obtain

$$\begin{aligned}
&\big|\mathbb{E}[h \circ f \circ W_{-i}(X_i) - h \circ f \circ W_{-i}(Z_i)]\big| \\
&\leq \big|\mathbb{E}[\partial_i(h \circ f)(W_{-i}(0)) + \frac{1}{2}\partial_i^2(h \circ f)(W_{-i}(0))\, X_i^2 \\
&\qquad - \partial_i(h \circ f)(W_{-i}(0))\, Z_i - \frac{1}{2}\partial_i^2(h \circ f)(W_{-i}(0))\, Z_i^2]\big| \\
&\quad + \frac{1}{6}\|\partial_i^3(h \circ f)\|_\infty\, \mathbb{E}|X_i|^3 + \frac{1}{6}\|\partial_i^3(h \circ f)\|_\infty\, \mathbb{E}|Z_i|^3 \,.
\end{aligned}$$

Since X_i and Z_i match in mean and variance, and $W_{-i}(0), W_{-i}(0)$ are independent of X_i, Z_i, the above difference in second-order Taylor approximations vanish. Therefore for all $1 \leq i \leq n$,

$$\big|\mathbb{E}[h \circ f \circ W_{-i}(X_i) - h \circ f \circ W_{-i}(Z_i)]\big| \leq \frac{1}{6}\|\partial_i^3(h \circ f)\|_\infty \left(\mathbb{E}|X_i|^3 + \mathbb{E}|Z_i|^3\right).$$

Substituting this control into (2.2) gives us the approximation bound

$$\begin{aligned}
d_{\mathcal{H}}(f(X), f(Z)) \;&=\; \sup_{h\in\mathcal{H}} \Big|\textstyle\sum_{i=1}^n \mathbb{E}\big[h \circ f \circ W_{-i}(X_i) - h \circ f \circ W_{-i}(Z_i)\big]\Big| \\
&\leq\; \frac{1}{6}\textstyle\sum_{i=1}^n \sup_{h\in\mathcal{H}} \|\partial_i^3(h \circ f)\|_\infty \left(\mathbb{E}|X_i|^3 + \mathbb{E}|Z_i|^3\right) \\
&\leq\; \frac{1}{6}\textstyle\sum_{i=1}^n \left(\|\partial_i f\|_\infty^3 + 3\|\partial_i f\|_\infty \|\partial_i^2 f\|_\infty + \|\partial_i^3 f\|_\infty\right) \left(\mathbb{E}|X_i|^3 + \mathbb{E}|Z_i|^3\right).
\end{aligned}$$

In the last line, we have applied a higher-order chain rule and used that the first three derivatives of h are bounded from above by 1. In summary, we obtain:

Lemma 2.1. *Let $X = (X_i)_{i\leq n}$ be a set of independent, mean-zero univariate random variables. Let $Z = (Z_i)_{i\leq n}$ be independent normal variables with $Z_i \sim \mathcal{N}(0, \mathrm{Var}[X_i])$. Fix a thrice-differentiable function $f : \mathbb{R}^n \to \mathbb{R}$. Then*

$$d_{\mathcal{H}}(f(X), f(Z)) \;\leq\; \frac{1}{6}\sum_{i=1}^n \left(\|\partial_i f\|_\infty^3 + 3\|\partial_i f\|_\infty \|\partial_i^2 f\|_\infty + \|\partial_i^3 f\|_\infty\right) \left(\mathbb{E}|X_i|^3 + \mathbb{E}|Z_i|^3\right).$$

Remark 2.1. Note that we have omitted the condition $\mathbb{E}|X_i|^3 < \infty$, as in the case where $\mathbb{E}|X_i|^3$ is unbounded, the above bound is interpreted as a vacuous bound. This convention is assumed throughout this thesis.

Lemma 2.1 is a typical universality approximation bound obtained by the Lindeberg method. To see how it may be small as $n \to \infty$, consider the case where X_i's are i.i.d. mean-zero and $f(X) = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$. In this case, the partial derivatives of f can be evaluated as

$$\|\partial_i f\|_\infty \;=\; \frac{1}{\sqrt{n}} \qquad \text{and} \qquad \|\partial_i^2 f\|_\infty \;=\; \|\partial_i^3 f\|_\infty \;=\; 0$$

for all $1 \leq i \leq n$. Provided that $\mathbb{E}|X_1|^3 < \infty$, Lemma 2.1 implies

$$d_{\mathcal{H}}\Big(\frac{1}{\sqrt{n}}\textstyle\sum_{i=1}^n X_i\,,\;\; \frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i\Big) \;=\; O(n^{-1/2}) \,.$$

For a general f, in order for the bound in Lemma 2.1 to imply convergence, we require that for all $1 \leq i \leq n$,

$$\|\partial_i f\|_\infty \;=\; o(n^{-1/3})\;, \quad \|\partial_i^2 f\|_\infty \;=\; O(n^{-2/3})\;, \quad \|\partial_i^3 f\|_\infty \;=\; o(n^{-1})\;.$$

Since each partial derivative measures the influence of the i-th input to the function, this says that the universality approximation is valid if the contribution of the i-th data point to the overall statistic $f(X)$ is vanishingly small. Conditions of this nature are called *stability conditions* (Mossel et al., 2005, 2010). One example of an unstable estimator is

$$f(X) \;=\; \max\{X_1, \ldots, X_n\}\;. \tag{2.4}$$

In this case, the generalised extreme value distribution is a suitable surrogate for X_i's and approximation results are found in the extreme value theory literature (Haan and Ferreira, 2006). We do not consider this case and focus only on stable estimators in this thesis.

Comparison to other proof techniques for distributional approximation. The Fourier method, Stein's method and Edgeworth expansion method (see e.g. Tao and Vu (2011); Chen et al. (2011); Ross (2011); Hall (2013)) are all techniques that have been routinely used for proving central limit theorems and their variants, as well as distributional approximations beyond normal variables. Beyond CLT for empirical averages, a wealth of results are available in the Stein's method literature for Gaussian universality approximations in the Wiener space and random matrix ensembles among others (Nourdin et al., 2009; Nourdin and Peccati, 2010). These methods typically require the knowledge of the limiting distribution family of $f(X)$: For example, Stein's method relies on the availability of Stein's kernel for the distributional approximation of $f(X)$, and techniques to control the approximation are typically specific to the different Stein kernel choice; extensions do exist in certain cases, see e.g. Gaunt (2020); Gaunt and Sutcliffe (2023). In comparison, the Lindeberg method directly targets the approximation of the input variable by Gaussians without knowing what the limit of $f(X)$ or $f(Z)$ may be, which is natural for our universality-type approximations. The Lindeberg method is also more flexible with rather mild assumptions on the function f. This flexibility is known to come at a price: The Lindeberg method is known to result in sub-optimal error rates in the Kolmogorov distance (2.1) for a range of specific problems (see the discussion after Theorem 3.3 in Chen et al. (2011) for empirical averages, the remark in Section 4 of Brailovskaya and van Handel (2022) for random matrices, and the example at the start of Section 4.5). At a high level, this sub-optimality is due to too much smoothing: the approximation of the Kolmogorov metric by $d_{\mathcal{H}}$ with thrice-differentiable test functions prevents one from obtaining finer controls for well-behaved estimators like the empirical averages. One important finding in this thesis, Theorem 4.7, is that this sub-optimality is a necessary price of generality: If the data are generated by some general n-dependent

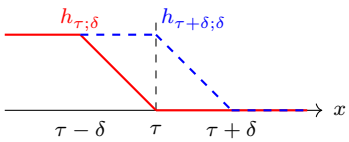
probability measures (which occurs, for example, when data are $\mathbb{R}^{d(n)}$-valued and the dimension $d = d(n)$ are of comparable size to n), the bounds obtained by the Lindeberg method are, in fact, near-optimal.

2.2 Modifications and extensions to the Lindeberg proof

In this section, we provide an informal discussion on how various aspects of the proof in Section 2.1 may be adapted to accommodate a more complicated setup. For each adaptation, we also provide pointers to the relevant sections of this thesis that takes advantage of the adaptation.

Finite-sample bounds on the Kolmogorov distance. For practical purposes such as hypothesis testing and confidence intervals, it is desirable to obtain distributional controls on the Kolmogorov distance (2.1). To relate the $d_{\mathcal{H}}$-control in Lemma 2.1 to (2.1), one needs to approximate the indicator $\mathbb{I}\{f(X) \le t\}$ by $h_t(f(X))$ for some smooth function h_t. If $\mathcal{H}$ is just the class of bounded Lipschitz functions, one example of h_t is $h_t(x) := \frac{1}{\delta} h_{t;\delta}(x)$ for some sufficiently small $\delta > 0$, where

$$h_{\tau;\delta}(x) := \begin{cases} 1 & \text{if } x < \tau - \delta \,, \\ \frac{\tau - x}{\delta} & \text{if } x \in [\tau - \delta, \tau) \,, \\ 0 & \text{if } x \ge \tau \,. \end{cases}$$



For the class $\mathcal{H}$ of thrice differentiable functions considered in the Lindeberg method, we adapt this construction to obtain a thrice differentiable approximation (Lemma A.10 used in Chapters 3 to 5). Asymptotically, the approximation of $\mathbb{I}\{\bullet \le t\}$ by h_t allows one to prove that convergence in $d_{\mathcal{H}}$ implies convergence in Kolmogorov metric (Corollary 6.4 used in Chapter 6). To obtain a finite-sample bound in addition, one needs to account explicitly for the error made in the interval $[\tau - \delta, \tau + \delta]$. This can be achieved if an anti-concentration bound for $f(Z)$ is known, i.e. if one can control how fast $\mathbb{P}(f(Z) \in [\tau - \delta, \tau + \delta])$ decays in δ as $\delta \to 0^+$. In Chapters 3 to 5, we follow Mossel et al. (2010)'s approach to utilise the Carbery-Wright inequality (Carbery and Wright, 2001), which gives an anti-concentration control for degree-m polynomials of Gaussians. This typically yields a sub-optimal rate, but as we show in Chapter 4, the rate can be nearly optimal in a general class of approximately polynomial functions, when the data probability measure also depends on n (e.g. by $d = d(n)$). The lower bound cannot be proved by the Lindeberg method; see Section 4.5 for our proof technique.

Beyond multilinear functions. Rotar (1976); Rotar et al. (1979); Mossel et al. (2010) use the Lindeberg method to prove universality results for multilinear polynomials. This

is the most natural case for the Lindeberg method, as multilinearity would immediately yield $\|\partial_i^2 f\|_\infty = \|\partial_i^3 f\|_\infty = 0$ in the bound of Lemma 2.1. To extend this beyond multilinear functions, two common approaches are as follows:

(i) We may approximate f by a suitable multilinear function. This is considered in Section 5.1 for approximating V-statistics by U-statistics.

(ii) We may obtain explicit bounds on $\partial_i^2 f$ and $\partial_i^3 f$. This is considered in Section 6.4.5 for ridge regression in moderate dimensions, where universality is shown to hold but the departure from linearity is the source of an unexpected observation.

Two additional preprocessing techniques on X and f are also useful:

(i) The dependence of $f(X)$ on X_i may be completely described by some feature vectors $Y_i := \phi(X_i)$, and $f(X)$ may be linear in Y_i even though it is not linear in X_i. This is used in Chapters 3 and 5 for rewriting U-statistics in terms of degree-two polynomials in appropriately transformed versions of the original data.

(ii) Since the Kolmogorov distance is invariant under a strictly monotonic transformation $\tilde{\tau}$, it suffices for us to have $f = \tilde{\tau} \circ \tilde{f}$, where $\tilde{f}$ is a multilinear function. We comment on this observation in the remarks after Theorem 4.1.

Unbounded derivatives of f. Bounding terms like $\|\partial_i f\|_\infty$ in Lemma 2.1 requires a uniform control on the derivatives of f, which is too strong for most practical estimators. Notice that the $\|\bullet\|_\infty$ arises from a crude bound on the third-order Taylor remainder in (2.3). Instead, it suffices to control $\partial_i f(\bullet)$ only on the intervals $[\mathbb{E}[X_i], X_i]$ and $[\mathbb{E}[X_i], Z_i]$, where the Taylor approximation was performed. This notably implies only local control of the derivatives in the two intervals of most interest to our problem. All Lindeberg-based proofs in the thesis use these local controls, which are also a staple in the universality literature (e.g. Montanari and Saeed (2022), Han and Shen (2023)).

ν-th moment for $\nu \in (2, 3]$. If X_i's only have bounded ν-th moments for $\nu \in (2, 3]$, we can choose h to be a twice-differentiable test function with a $(\nu - 2)$-Hölder second derivative rather than a thrice-differentiable function. The only modification in the proof is that, instead of controlling the Taylor remainder in (2.3) by the third-order Taylor remainder, we use the Hölder condition. This is used in Chapters 3 to 5 and an explicit construction of our choice of h is also in Lemma A.10.

High-dimensionality. Suppose X_i's are $\mathbb{R}^d$-valued instead of $\mathbb{R}$-valued, where the dimension $d = d(n)$ may grow in n. The step in (2.3) can introduce rather crude dimension dependence, if we use the Cauchy-Schwarz inequality to obtain quantities such as $\|X_i\|^3$ and $\|\partial_i f(X)\|$, where $\|\bullet\|$ is the Euclidean norm. To obtain a more careful control, we may keep the vector product $\partial_i f(X)^\top X_i$ and exploit concentration properties of this product. For example, in the case of a simple degree-two U-statistic

$f(X) = \frac{1}{\sqrt{n(n-1)}} \sum_{i \neq j} X_i^\top X_j$, where X_i's are $\mathbb{R}^p$-valued, mean-zero and i.i.d., this product evaluates to

$$\partial_i f(X)^\top X_i \;=\; \frac{1}{\sqrt{n(n-1)}} \sum_{j \neq i} X_j^\top X_i \;\approx\; \frac{1}{\sqrt{n}} \Big(\frac{1}{\sqrt{n}} \sum_{j \neq i} X_j \Big)^\top X_i \,. \tag{2.5}$$

Suppose for $j \neq i$, $\mathbb{E}[X_j^\top X_i | X_i]$ is $O(1)$ with high probability. Then by the central limit theorem, (2.5) is $O(n^{-1/2})$ with high probability, regardless of how large the dimension d is with respect to n. This is heavily exploited in Chapters 3 to 5 for removing the dependence on the dimension d in the universality approximation bound.

However, this approach does not help with getting rid of dimension dependence in the ridge regression and ridgeless regression examples in Chapter 6. There, the high-dimensionality manifests through the pseudo-inverse of a large sample covariance matrix $(\frac{1}{n} \sum_{i=1}^n X_i X_i^\top)^\dagger$, which necessitates a careful control on the smallest non-zero eigenvalue of a large $\mathbb{R}^{d \times d}$ random matrix. In those settings, we need to combine the Lindeberg proof with concentration results from random matrix theory, and impose the condition that $d = O(n)$, i.e. the dimension grows at most proportionally to n.

Dependence. Say n is divisible by k and write $n = mk$. Suppose $(X_i)_{i \leq n}$ satisfies block dependence, i.e. we can form the blocks

$$\mathbf{X}_1 \coloneqq (X_1, \ldots, X_k) \,, \qquad \ldots \,, \qquad \mathbf{X}_m \coloneqq (X_{(m-1)k+1}, \ldots, X_{mk}) \,,$$

such that $\mathbf{X}_l$'s are independent with each other but arbitrary dependence is allowed within each block $\mathbf{X}_l$. Provided that $m \to \infty$, the Lindeberg method of Lemma 2.1 still applies for approximating $f(X)$ by $f(Z)$; the only differences are that $(Z_i)_{i \leq n}$ are also block-dependent to match the dependence structure of $(X_i)_{i \leq n}$, and that the moment boundedness condition on $\mathbb{E}|X_i|^3$ needs to be replaced by the stronger control on $\mathbb{E}\|\mathbf{X}_i\|^3$. The remarks after Theorem 4.1 discuss the implication of requiring $\mathbb{E}\|\mathbf{X}_i\|^3$ to be bounded and the implicit condition it may impose on the size of k. Chapter 6 develops a universality result for block dependence and applies it to study the effects of data augmentation. Lahiry and Sur (2024) also builds universality results for high-dimensional regularised linear models under block-dependent coordinates.

While not covered by this thesis, we remark that once an approximation result is established under block dependence, one may extend this result to m-dependence and mixing (see definitions and applications in e.g. Cryer (1986); Brock et al. (1992); Schweinberger and Handcock (2015); Wackernagel (2003); Billingsley (1995); Bradley (2005)). This is achieved by the classical big-block-small-block technique: One represents the data $(X_i)_{i \leq n}$ as an alternating sequence of big blocks and small blocks of random vectors, where the big blocks become approximately independent and the small blocks have negligible contributions (Bernstein, 1927; Ibragimov, 1975; Davidson, 1992). In a recent

joint work of Mallory, Huang, and Austern (2025), universality results are provided for high-dimensional logistic regression models for block dependence, m-dependence and specific mixing processes. We mention this work in Section 7.1, but shall focus only on how an exact risk analysis may be performed under dependence *after* Gaussian universality is established.

Non-smoothness. In many practical cases, such as estimators arising from an optimisation problem, f may not be differentiable due to e.g. the presence of $\max$ and $\min$. These points of non-differentiability can be neglected, if the quantities $f(X)$ and $f(Z)$ of interest do not take values on those points with high probability. This is used in the consideration of a maximum of a high-dimensional average in Section 6.6.1 with an example use case in Section 7.2.2.

Random components with negligible contributions. Suppose we may express $f(X) = f_1(X) + f_2(X)$ such that $f_1(X)$ is amenable to the Lindeberg method, whereas $f_2(X)$ is not (e.g. lands on points of non-differentiability with high probability). WLOG suppose that $f_1(X)$ and $f_2(X)$ both have zero means. If $f_2(X)$ has negligible contributions to the overall asymptotic distribution, one may expect to ignore $f_2(X)$ and obtain a universality approximation of $f(X)$ by only $f_1(Z)$. It turns out that a sufficient condition for "ignoring" $f_2(X)$ is that $\mathrm{Var}[f_2(X)] \ll \mathrm{Var}[f_1(X)]$. This is formalised under a concept called "variance domination" in Section 4.3. We use this technique to extend a universality result on degree-m polynomials (Theorem 4.1) to a result for approximately polynomial functions, where the approximation is made in the L_2 sense. As applications, this idea is also exploited to obtain different limiting approximations for U-statistics and a high-dimensional delta method (Chapters 3 and 5).

Chapter 3

Distribution approximations of degree-two U-statistics in large dimensions

In this section, we focus on a simple yet illustrative application of universality in high-dimensional analysis. It will become clear that the main results here are special cases of the general results in Chapter 4. Specifically, we consider the distributional approximation of a one-dimensional U-statistic of degree two, given by

$$D_n \coloneqq u_2(Y) = \frac{1}{n(n-1)} \sum_{i \neq j} u(Y_i, Y_j) \,, \tag{3.1}$$

where $Y \coloneqq (Y_i)_{i \leq n}$ is a collection of i.i.d. random vectors in $\mathbb{R}^d$, $n \geq 2$ and $u : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a symmetric measurable function. Here, we use the notation D_n to emphasise the role of the U-statistic as a measure of discrepancy in our applications in Section 3.4, but also highlight that it is exactly a special case of the degree-m U-statistic $u_m(Y)$ to be considered in Section 5.3. We also write the quantity to be estimated by D_n as

$$D \coloneqq \mathbb{E}[u_2(Y)] = \mathbb{E}[u(Y_1, Y_2)] \,. \tag{3.2}$$

Numerous estimators can be formulated as a U-statistic: Modern applications include gene-set testing (Chen and Qin, 2010), high-dimensional change-point detection (Wang et al., 2022), convergence guarantees for random forests (Peng et al., 2022) and kernel-based tests in machine learning (Gretton et al., 2012).

The asymptotic theory of D_n is well-established in the classical setting, where d is fixed and small relative to n (e.g. Chapter 5 of Serfling (1980)). Yet, those results fail to apply to the modern context of high-dimensional data, where d is of a comparable size to n, and where such U-statistics are empirically observed to exhibit pathological behaviours (Reddi et al., 2015; Ramdas et al., 2015). In the context of high-dimensional testing, many theoretical works do study the limiting distributions of specific forms of D_n (Chen and Qin, 2010; Wang et al., 2015; Yan and Zhang, 2022), but with efforts mostly focused on obtaining Gaussian limits. A related line of work, building on the seminal work of De Jong (1990), has investigated criteria for the Gaussianity of D_n regardless of degeneracy (Döbler and Peccati, 2017, 2019); these results complement ours via the fourth moment theorem of Nualart and Peccati (2005), as we shall discuss

in Section 3.2.3. Recent works (Döbler et al., 2022; Bhattacharya et al., 2022) have also obtained the asymptotics of D_n beyond the classical notion of degeneracy, with focus on quadratic forms with varying weights and time-indexed sequences of U-processes. As our primary application in this section is kernel-based testing in high-dimensions, we focus on deriving bounds in Kolmogorov distance that are valid for any fixed n and d, which allow us to understand how dimension d plays a role in the distributional approximation.

In this chapter, we show how universality results can be applied to obtain Gaussian and non-Gaussian approximations of D_n, when dimension d is allowed to grow at an *arbitrary* rate relative to n under a suitable assumption. As a byproduct, we show that the effect of dimension d on the limit of D_n is captured completely by a variance ratio ρ_d. This ratio is a high-dimensional analogue of the classical notion of degeneracy in U-statistics: Depending on the ratio, the limiting distribution of U-statistics can take either the non-degenerate Gaussian limit, the degenerate limit or an intermediate distribution.

The rest of the chapter is organised as follows: Section 3.1 sketches the intuition of the main result for the linear kernel and how unexpected asymptotic limits can arise in large dimensions. Section 3.2 presents the formal result with a finite-sample, dimension-independent error bound, established via universality. Section 3.3 shows that this bound is nearly tight by a pair of matching upper and lower bounds for a specific U-statistic. Section 3.4 presents practical implications of these results in the context of high-dimensional distributional tests with Maximum Mean Discrepancy (MMD) and Kernel Stein Discrepancy (KSD). All proofs are included in Appendix A.

3.1 Intuition via the example of a linear kernel

Loosely speaking, our main result in the upcoming Section 3.2 says that as $n, d \to \infty$, the statistic D_n converges in distribution to a quadratic form of Gaussians:

$$D_n \overset{d}{\approx} W_2 + Z_2 + D \,, \tag{3.3}$$

where W_2 is some infinite sum of weighted and centred chi-squares, Z_2 is some Gaussian, and the two variables are correlated. D is the population version of D_n defined in (3.2). $W_2 + D$ is closely related to the classical degenerate limit, whereas $Z_2 + D$ gives exactly the classical non-degenerate limit.

To understand how W_2 and Z_2 arise, it is instructive to consider a decomposition of D_n for the simple case of the linear kernel $u(y_1, y_2) := y_1^\top y_2$. Denoting the centred random vectors $\bar{Y}_i = Y_i - \mathbb{E}[Y_1]$, we have

$$\frac{1}{n(n-1)} \sum_{i \neq j} Y_i^\top Y_j \;=\; \frac{1}{n(n-1)} \sum_{i \neq j} (\bar{Y}_i + \mathbb{E}[Y_1])^\top (\bar{Y}_j + \mathbb{E}[Y_1])$$

$$= \frac{1}{n(n-1)} \sum_{i \neq j} \bar{Y}_i^\top \bar{Y}_j + \frac{2}{n} \sum_{i=1}^n \bar{Y}_i^\top \mathbb{E}[Y_1] + \mathbb{E}[Y_1]^\top \mathbb{E}[Y_1]$$
$$= \underbrace{\frac{1}{n-1}\left(\left\| \frac{1}{\sqrt{n}} \sum_{i \leq n} \bar{Y}_i \right\|^2 - \frac{1}{n} \sum_{i=1}^n \bar{Y}_1^\top \bar{Y}_1 \right)}_{(\star)_W} + \underbrace{\frac{2}{n} \sum_{i=1}^n \bar{Y}_i^\top \mathbb{E}[Y_1]}_{(\star)_Z} + \mathbb{E}[Y_1^\top Y_2] \,. \quad (3.4)$$

Here, D_n decomposes into a sum of three terms: $(\star)_W$ corresponds to W_2 and, under the CLT for d fixed, behaves like a centred chi-squared variable at the scale $\frac{1}{n-1}$; $(\star)_Z$ corresponds to Z_2 and, under the CLT for d fixed, behaves like a normal variable at the scale $\frac{1}{\sqrt{n}}$; the third term is $\mathbb{E}[D_n]$. Notably in the case of fixed dimension, the variance of $(\star)_W$ is always smaller than that of $(\star)_Z$ unless $(\star)_Z = 0$ almost surely.

Suppose d is fixed. Classical limit theorems on U-statistics say that the asymptotic distribution of D_n (upon appropriate rescaling) depends on the notion of *degeneracy*: D_n is degenerate if $\sigma_{\text{cond}} = 0$, where

$$\sigma_{\text{cond}} \coloneqq \sqrt{\text{Var}\mathbb{E}[u(Y_1, Y_2)|Y_1]} \,. \quad (3.5)$$

When d is fixed, upon rescaling, a non-degenerate D_n has a Gaussian limit, whereas a degenerate D_n has a limit described by an infinite sum of weighted and centred chi-squares. In the case of a linear kernel, $\sigma_{\text{cond}} = \sqrt{\mathbb{E}[Y_1]^\top \text{Var}[Y_2] \mathbb{E}[Y_1]}$ is exactly the variance of the linear term $(\star)_Z$ of (3.4). Therefore in the case of (3.4), one way to interpret degeneracy is that if $(\star)_Z$ does not vanish, D_n is asymptotically close to Z_2, and if $(\star)_Z$ vanishes, D_n is asymptotically close to W_2.

The two key arguments in the fixed d case are (i) applying CLT to approximate averages by Gaussians and (ii) determining which of $(\star)_W$ and $(\star)_Z$ dominates based on degeneracy. In the case of a growing $d = d(n) \to \infty$, Gaussian universality effectively substitutes (i) and allows us to replace Y_i's by Gaussians. The high-dimensional analogue of (ii), on the other hand, is more subtle, as the variances of $(\star)_W$ and $(\star)_Z$ are affected by both the growing n and d. More concretely, observe that in the general case, the variance of D_n decomposes in a similar manner to (3.4) as

$$\begin{aligned} \text{Var}[D_n] &= O\Big(\frac{\mathbb{E}[(u(Y_1, Y_2) - D)(u(Y_1, Y_2) - D)]}{n(n-1)} + \frac{\mathbb{E}[(u(Y_1, Y_2) - D)(u(Y_1, Y_3) - D)]}{n} \Big) \\ &= O\Big(\frac{\sigma_{\text{full}}^2}{n(n-1)} + \frac{\sigma_{\text{cond}}^2}{n} \Big) \,, \end{aligned}$$

where σ_{cond}^2 is defined as above and corresponds to the variance of $\sqrt{n(n-1)}\,(\star)_W$, and

$$\sigma_{\text{full}} \coloneqq \sqrt{\text{Var}[u(X_1, X_2)]} \quad (3.6)$$

corresponds to the variance of $\sqrt{n}\,(\star)_Z$. When d is large, σ_{full} and σ_{cond} can scale with d at different rates, and how the variance of $(\star)_W$ compares with $(\star)_Z$ depends on the ratio

$$\rho_d \coloneqq \frac{\sigma_{\text{full}}}{\sigma_{\text{cond}}} \,.$$

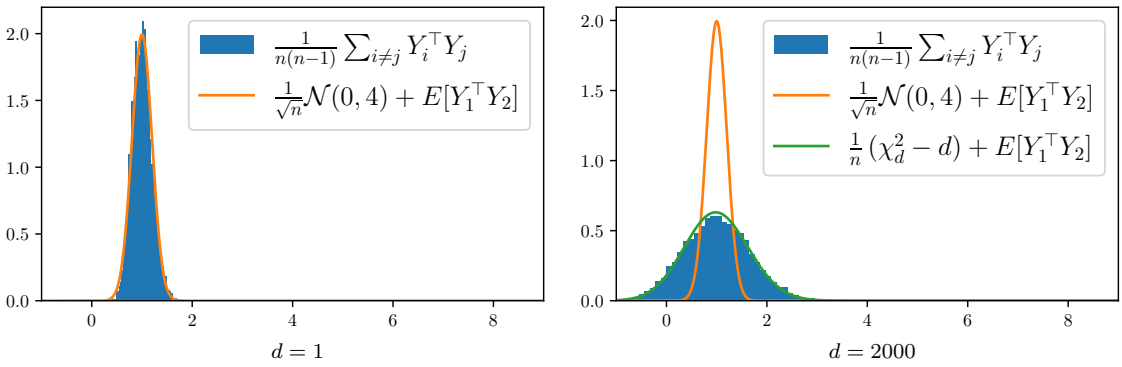


Figure 3.1: Probability density plots of D_n under the linear kernel (3.1) with different dimension parameters d. In both plots, $n = 200$ data are drawn with $Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\frac{1}{\sqrt{d}}\mathbf{1}_d, I_d)$, and one can compute $\sigma_{\text{cond}} = 1$.

As our result in Section 3.2 reveals, this comparison of variances is generally sufficient to determine whether the asymptotic behaviour of D_n is driven by $(\star)_W$, $(\star)_Z$ or both:

$\rho_d \lesssim n^{1/2}$	$\rho_d \sim n^{1/2}$	$\rho_d \gtrsim n^{1/2}$
Non-degenerate limit	Intermediate limit	Degenerate limit
Gaussian	Quadratic form of Gaussian	∞-sum of weighted and centred chi-squares

Table 3.1: Possible asymptotics of a degree-two U-statistic

In other words, the condition $\rho_d \gtrsim n^{1/2}$ is the high-dimensional analogue of degeneracy. This reveals the first unexpected asymptotic in the high-dimensional regime: Even for a non-degenerate U-statistic D_n with $\sigma_{\text{cond}} \neq 0$, ρ_d can become asymptotically larger than $n^{1/2}$ as d grows, causing D_n to behave like a degenerate U-statistic. This is illustrated in Figure 3.1 for the case of the linear kernel (3.4), where the same U-statistic transitions from a non-degenerate limit to a degenerate limit as d increases from $d = 1$ to $d = 2000$. This degenerate behaviour is further demonstrated in Figure 3.2 for non-degenerate U-statistics that naturally arise in the setting of distribution tests.

An additional observation from Figure 3.1 is that, as d becomes large, the chi-squared variable $\frac{1}{n}\chi^2_d$ in the degenerate limit becomes asymptotically Gaussian, as demonstrated by the symmetry of the density plot. This reveals another unexpected asymptotic limit in the high-dimensional regime. While the degenerate limit is described by an infinite sum of chi-squares, the infinite sum is additionally affected by a growing dimension d. As a result, the degenerate limit can itself become asymptotically Gaussian, albeit at a different variance and scale compared to that of the non-degenerate limit. Figure 3.2 demonstrates two cases, one where the c.d.f. of the degenerate approximation is symmetric and one where asymmetry arises. This effect is not a result of ρ_d, but instead of the fourth moment theorem of Nualart and Peccati (2005): We shall discuss this briefly in Section 3.2.3 in

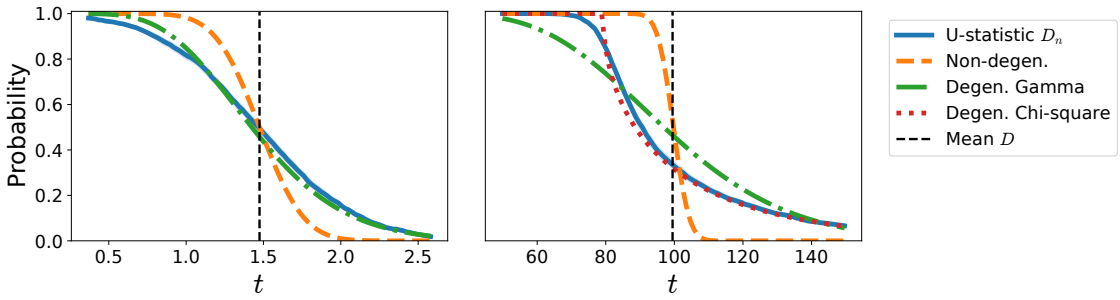


Figure 3.2: Behaviour of $\mathbb{P}(X > t)$ for $X = D_n$, a particular *non-degenerate* degree-two high-dimensional U-statistic, versus X as the non-degenerate Gaussian approximations and the degenerate approximations. The two plots correspond to different setups detailed in Section 3.4. Both show the departure of D_n from the classical non-degenerate limit, with the right plot additionally showing asymmetry.

the context of D_n and defer a formal statement to Section 4.4 in the context of more general estimators.

3.2 Distributional approximations with dimension-free error bounds

To formalise and extend the observations in Section 3.1 beyond linear kernels, the main technical hurdle is to establish the approximation (3.3) in the regime where $n, d \to \infty$. In this section, we establish the asymptotics in Table 3.1 through a finite-sample bound for (3.3), which is dimension-independent under a mild condition. The main technique is the application of Gaussian universality to a degree-two polynomial.

L_ν moment terms. As with classical Berry-Esseen bounds for empirical averages, finite-sample bounds typically require moment controls that are slightly more than the second moment. For $\nu \in (2, 3]$, our bounds will involve the L_ν-analogue of σ_{cond} and σ_{full} from (3.5) and (3.6) as

$$M_{\text{cond};\nu} := \big\|\mathbb{E}[u(Y_1, Y_2)|Y_2] - \mathbb{E}[u(Y_1, Y_2)]\big\|_{L_\nu}, \quad M_{\text{full};\nu} := \big\|u(Y_1, Y_2) - \mathbb{E}[u(Y_1, Y_2)]\big\|_{L_\nu}.$$

They are respectively related to the non-degenerate and degenerate approximations Z and W in (3.3), and also scale as d grows.

3.2.1. Non-degenerate approximation when $\rho_d = o(n^{1/2})$

The Berry-Esseen bound for non-degenerate Gaussian approximation is well-known from classical results (see e.g. Theorem 10.3 of Chen et al. (2011)). We restate it here for completeness and for motivating our assumptions for the general case. If $\sigma_{\text{cond}} > 0$, then for a normal random variable $Z \sim \mathcal{N}(\mathbb{E}[D_n], 4n^{-1}\sigma_{\text{cond}}^2)$ and $\nu \in (2, 3]$, we have

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}\Big(\frac{\sqrt{n}}{\sigma_{\text{cond}}} D_n < t\Big) - \mathbb{P}\Big(\frac{\sqrt{n}}{\sigma_{\text{cond}}} Z < t\Big)\right| \le \frac{6.1 M_{\text{cond};\nu}^\nu}{n^{(\nu-2)/2}\sigma_{\text{cond}}^\nu} + \frac{(1+\sqrt{2})\rho_d}{2(n-1)^{1/2}}. \tag{3.7}$$

This establishes the non-degenerate approximation in Table 3.1, which is only valid when $\rho_d = o(n^{1/2})$. The ratio $M_{\text{cond};\nu}/\sigma_{\text{cond}}$ also appears in the bound (3.7); however, we do not focus on how this ratio scales, since it appears in the Berry-Esseen bound even for sample averages. Error bounds in our main theorem will depend on similar ratios, and for our theorem to imply asymptotic convergence, the following assumption is required:

Assumption 3.1. There exists some $\nu \in (2, 3]$ and some absolute constant $C < \infty$ such that $\frac{M_{\text{full};\nu}}{\sigma_{\text{full}}} \leq C$ and $\frac{M_{\text{cond};\nu}}{\sigma_{\text{cond}}} \leq C$.

Remark. (i) If Assumption 3.1 holds for $\nu > 3$, it also holds for all $\nu \in (2, 3]$. We restrict our attention to $\nu \in (2, 3]$ for simplicity.
(ii) To see an example of when Assumption 3.1 can be violated, consider the linear kernel $u(Y_1, Y_2) = Y_1^\top Y_2$. Also denote $\mu = \mathbb{E}[Y_1]$ and $\bar{Y}_i = Y_i - \mu$. Then the first moment ratio computes as

$$\frac{M_{\text{full};\nu}}{\sigma_{\text{full}}} = \frac{\|\bar{Y}_1^\top \bar{Y}_2\|_{L_\nu}}{\|\bar{Y}_1^\top \bar{Y}_2\|_{L_2}} = \frac{\|\sum_{l\leq d}(\bar{Y}_1)_l(\bar{Y}_2)_l\|_{L_\nu}}{\|\sum_{l\leq d}(\bar{Y}_1)_l(\bar{Y}_2)_l\|_{L_2}} = \frac{\|\sum_{l\leq d}(\bar{Y}_1)_l(\bar{Y}_2)_l\|_{L_\nu}}{\sqrt{\sum_{l,l'\leq d}\mathbb{E}[(\bar{Y}_1)_l(\bar{Y}_1)_{l'}]\mathbb{E}[(\bar{Y}_2)_l(\bar{Y}_2)_{l'}]}\|_{L_2}} .$$

Suppose that all coordinates of $\bar{Y}_1$ have unit variance. If the different coordinates of $\bar{Y}_1$ are uncorrelated, the denominator computes as $\Theta(\sqrt{d})$. Meanwhile, the numerator is on the order $O(d)$, and is not guaranteed to be $\Theta(\sqrt{d})$ due to potential dependencies across the coordinates that do not show up in the linear correlations. This will cause Assumption 3.1 to be violated. Note that a similar argument also applies to the other moment ratio

$$\frac{M_{\text{cond};\nu}}{\sigma_{\text{cond}}} = \frac{\|\mu^\top \bar{Y}_1\|_{L_\nu}}{\|\mu^\top \bar{Y}_1\|_{L_2}} = \frac{\|\sum_{l\leq d}\mu_l(\bar{Y}_1)_l\|_{L_\nu}}{\sqrt{\sum_{l,l'\leq d}\mathbb{E}[(\bar{Y}_1)_l(\bar{Y}_1)_{l'}]\mu_l\mu_{l'}}\|_{L_2}} .$$

3.2.2. The general case

Our general approximation relies on a functional decomposition assumption. For a triangular array of $\mathbb{R}^d \to \mathbb{R}$ functions $\{\phi_k^{(K)}\}_{k\leq K, K\in\mathbb{N}}$ and a triangular array of real values $\{\lambda_k^{(K)}\}_{k\leq K, K\in\mathbb{N}}$, we define the L_ν approximation error for $\nu \geq 1$ and a given $K \in \mathbb{N}$ as

$$\varepsilon_{K;\nu} := \left\| \sum_{k=1}^K \lambda_k^{(K)}\phi_k^{(K)}(Y_1)\phi_k^{(K)}(Y_2) - u(Y_1, Y_2)\right\|_{L_\nu} .$$

Assumption 3.2. There exists some $\nu \in (2, 3]$ such that, for any given n and d, there exists some (n, d)-dependent choices of $(\phi_k^{(K)})$ and $(\lambda_k^{(K)})$ such that, as $K \to \infty$, the L_ν approximation error $\varepsilon_{K;\nu} \to 0$.

Remark 3.1. Assumption 3.2 always holds for $\nu = 2$ by the spectral decomposition of the Hilbert-Schmidt operator $f(\bullet) \mapsto \mathbb{E}[u(\bullet, Y_1)f(Y_1)]$ on the space $L_2(\mathbb{R}^d, \mu_{Y_1})$, where μ_{Y_1} is the law of Y_1. For degenerate U-statistics with d fixed, the corresponding orthonormal eigenbasis of functions and eigenvalues are used to prove asymptotic results (see Section 5.5.2 of Serfling (1980)) and finite-sample bounds (Bentkus and Götze,

1999; Götze and Tikhomirov, 2005; Yanushkevichiene, 2012). In fact, these finite-sample bounds are dependent on the specific $\lambda_k^{(K)}$'s, making the results hard to apply. Instead, we forgo orthonormality at the cost of a convergence slightly stronger than L_2. This allows for a much more flexible choice of $(\phi_k^{(K)}, \lambda_k^{(K)})$ and is particularly well-suited for a kernel-based setting; see the discussion after Lemma 3.11 in Section 3.4.2. We also defer to Assumption 5.1 in Section 5.3 for a similar assumption for a degree-m U-statistic and a discussion on how it can be easily verified for any u well-approximated by a Taylor expansion.

Assumption 3.2 allows us to approximate each $u(Y_i, Y_j)$ by an inner product of two independent, high-dimensional random vectors in $\mathbb{R}^K$. This reduces the study of D_n to a degree-two polynomial of high-dimensional random vectors, which is essentially the linear kernel case in (3.4). The distributional approximation thus depends on the structure of the inner product, described by

$$\Lambda^K \coloneqq \operatorname{diag}\{\lambda_1^{(K)}, \dots, \lambda_K^{(K)}\} \in \mathbb{R}^{K\times K}, \quad \phi^K(x) \coloneqq (\phi_1^{(K)}(x), \dots, \phi_K^{(K)}(x))^\top \in \mathbb{R}^K.$$

Loosely speaking, they can be viewed as a diagonal matrix of the first K "eigenvalues" and a concatenation of the first K "eigenfunctions", although we emphasise that these values are not necessarily associated with a spectral decomposition in view of Remark 3.1. We also denote the mean and variance of $\phi^K(X_1)$ by

$$\mu^K \coloneqq \mathbb{E}[\phi^K(Y_1)] \qquad \text{and} \qquad \Sigma^K \coloneqq \operatorname{Cov}[\phi^K(Y_1)].$$

We seek to apply Gaussian universality to replace the vectors $\phi^K(Y_i)$. To this end, let η_i^K, with $i, K \in \mathbb{N}$, be i.i.d. standard Gaussian vectors in $\mathbb{R}^K$. The approximation is given by a quadratic form of Gaussians, defined as

$$\begin{aligned} U_n^K \coloneqq{}& \frac{1}{n(n-1)} \sum\nolimits_{1\le i\ne j\le n} (\eta_i^K)^\top (\Sigma^K)^{1/2} \Lambda^K (\Sigma^K)^{1/2} \eta_j^K \\ &+ \frac{2}{n} \sum\nolimits_{i=1}^n (\mu^K)^\top \Lambda^K (\Sigma^K)^{1/2} \eta_i^K + \mathbb{E}[D_n]. \end{aligned} \tag{3.8}$$

The three components respectively correspond to W_2, Z_2 and D in (3.3). We also denote the dominating moment terms from W_2 and Z_2 by

$$\sigma_{\max} \coloneqq \max\{\sigma_{\text{full}}, (n-1)^{1/2}\sigma_{\text{cond}}\}, \quad M_{\max;\nu} \coloneqq \max\{M_{\text{full};\nu}, (n-1)^{1/2} M_{\text{cond};\nu}\}.$$

We are ready to state the main result.

Theorem 3.1. *There exists an absolute constant $C > 0$ such that, if $\nu \in (2, 3]$ satisfies Assumption 3.2, then the following holds:*

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\max}} D_n > t \Big) - \lim_{K\to\infty} \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\max}} U_n^K > t \Big) \right| \le C n^{-\frac{\nu-2}{4\nu+2}} \Big(\frac{M_{\max;\nu}}{\sigma_{\max}} \Big)^{\frac{\nu}{2\nu+1}}.$$

Theorem 3.1 turns out to be a direct application of the general universality result

(Theorem 4.1) in Section 4.2, proved via the Lindeberg method. We include the proof of Theorem 3.1 in Appendix A.3, defer a discussion of the key ideas to Section 4.2, and present a degree-m U-statistic generalisation in Section 5.3. A few observations on the results in Theorem 3.1:

(i) The bounds are independent of specific choices of $\lambda_k^{(K)}$ and $\phi_k^{(K)}$ in Assumption 3.2. It therefore suffices to verify Assumption 3.2 for any choice of $(\phi_k^{(K)}, \lambda_k^{(K)})$, which is non-unique in general;

(ii) If $\nu = 3$, the RHS is given by $Cn^{-\frac{1}{14}}\big(\frac{M_{\max;3}}{\sigma_{\max}}\big)^{3/7}$. If Assumption 3.1 holds for ν, the RHS can be replaced by $C'n^{-\frac{\nu-2}{4\nu+2}}$ for some constant C' and is *dimension-independent*;

(iii) One may be tempted to move $\lim_{K\to\infty}$ inside $\mathbb{P}$ such that, instead of the cumbersome expression of U_n^K with finite K, one may deal with random quantities in a Hilbert space. The reason to stick with U_n^K is that in Assumption 3.2, convergence of the infinite sum is required only in L_ν and not almost surely. This makes verification of the assumption substantially simpler in practice: In Appendix A.1, we illustrate how this assumption holds via a simple Taylor-expansion argument coupled with suitable tail behaviour of the data to control error terms. The same argument is not applicable if we instead require an almost sure convergence.

Theorem 3.1 immediately implies a convergence theorem. In the next result and subsequent results in the section, with a slight abuse of notation, we use $n, d \to \infty$ to denote the asymptotic regime as $n \to \infty$ and $d = d(n)$ is some positive integer variable dependent on n.

Corollary 3.2. *Suppose Assumptions 3.1 and 3.2 hold for some $\nu \in (2, 3]$ and the sequential weak limit $\bar{U} = \lim_{n\to\infty}\lim_{K\to\infty}\frac{\sqrt{n(n-1)}}{\sigma_{\max}}(U_n^K - D)$ exists. Then*

$$\frac{\sqrt{n(n-1)}}{\sigma_{\max}}(D_n - D) \xrightarrow{d} \bar{U} \qquad \text{as } n, d \to \infty .$$

Since Theorem 3.1 and Corollary 3.2 do not impose any restriction on ρ_d, they cover all three cases in Table 3.1. Yet, analysing the quadratic form approximation U_n^K can appear challenging. Indeed, U_n^K is a quadratic form of Gaussians, which does not admit a closed-form c.d.f. in general and whose limiting behaviour depends heavily on $\lambda_k^{(K)}$ and $\phi_k^{(K)}$. Nevertheless, the presence of Gaussianity still allows us to obtain crude bounds on the c.d.f. of U_n^K. Together with Theorem 3.1, this allows us to provide direct controls on the c.d.f. of the original U-statistic D_n, in a way that is independent of K and specific choices of ϕ_k and λ_k.

Proposition 3.3. *Suppose Assumption 3.2 holds for some $\nu \in (2, 3]$. Then there exist*

constants $C_1, C_2, C_3 > 0$ *such that for all* $\epsilon > 0$,

$$\mathbb{P}(|D_n - D| > \epsilon) \geq 1 - C_1\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\max}}\Big)^{1/2}\epsilon^{1/2} - C_2\, n^{-\frac{\nu-2}{4\nu+2}}\Big(\frac{M_{\max;\nu}}{\sigma_{\max}}\Big)^{\frac{\nu}{2\nu+1}},$$
$$\mathbb{P}(|D_n - D| > \epsilon) \leq C_3\epsilon^{-2}\Big(\frac{\sigma_{\max}}{\sqrt{n(n-1)}}\Big)^2 .$$

Remark. (i) The second line is a concentration inequality directly available via Markov's inequality, whereas the first bound is an anti-concentration result. Anti-concentration results are generally available only for random variables from known distribution families, and we obtain such a result by comparing D_n to U_n^K via universality. (ii) In the anti-concentration bound, the trailing error term involving $M_{\max;\nu}/\sigma_{\max}$ is inherited from Theorem 3.1 and is negligible. (iii) The dependence on ϵ in the concentration inequality is only ϵ^{-2}, since the approximation of Assumption 3.2 holds in L_ν for some $\nu \in (2, 3]$. If a stronger version of Assumption 3.2 is assumed, e.g. if the approximation holds almost surely, the result is improvable to a sub-exponential concentration bound.

Proposition 3.3 implies that the deviation of D_n from its mean is on the order $\frac{\sigma_{\max}}{n}$:

Corollary 3.4. *Fix* $\epsilon > 0$. *If Assumptions 3.1 and 3.2 hold for some* $\nu \in (2, 3]$, *then*

$$\mathbb{P}(|D_n - D| > \epsilon) \;\to\; \begin{cases} 1 & \text{if } \sigma_{\max} = \omega(n) \\ 0 & \text{if } \sigma_{\max} = o(n) \end{cases} \qquad \text{as } n, d \to \infty\ .$$

Since $\sigma_{\max} = \max\{\sigma_{\text{full}}, (n-1)^{1/2}\sigma_{\text{cond}}\}$, the case $\sigma_{\max} = \omega(n)$ happens only in the high-dimensional regime, in which case D_n fails to be a consistent estimator of D.

3.2.3. Degenerate approximation when $\rho_d = \omega(n^{1/2})$

Recall that the stochasticity of U_n^K in (3.8) comes from a linear term and a quadratic term. It turns out that, unless we are at the boundary case where $\rho_d = \Theta(n^{1/2})$, we can always approximate U_n^K further by keeping only one of these two terms. We have seen in Section 3.2.1 that keeping the linear term yields the non-degenerate limit, which is valid when $\rho_d = o(n^{1/2})$. Here, we show that keeping the quadratic term yields the degenerate limit, which is valid when $\rho_d = \omega(n^{1/2})$. Note that in this case, $\sigma_{\max} = \sigma_{\text{full}}$.

To state the result, let $\{\xi_k\}_{k=1}^\infty$ be a sequence of i.i.d. standard Gaussians in 1d, and for $K \in \mathbb{N}$, let $\{\tau_k^{(K)}\}_{k=1}^K$ be the eigenvalues of $(\Sigma^K)^{1/2}\Lambda^K(\Sigma^K)^{1/2}$. The limiting distribution we consider is given in terms of

$$W_n^K \;:=\; \frac{1}{\sqrt{n(n-1)}}\sum_{k=1}^K \tau_k^{(K)}(\xi_k^2 - 1) + D\ . \tag{3.9}$$

The next result adapts Theorem 3.1 by replacing U_n^K with W_n^K:

Proposition 3.5. *Suppose Assumption 3.2 holds for some $\nu \in (2, 3]$. There exists an absolute constant $C > 0$ such that*

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} D_n > t \Big) - \lim_{K\to\infty} \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} W_n^K > t \Big) \right|$$
$$\leq C\Big(\frac{1}{(n-1)^{1/5}} + \Big(\frac{\sqrt{n-1}\,\sigma_{\text{cond}}}{\sigma_{\text{full}}} \Big)^{2/5} + n^{-\frac{\nu-2}{4\nu+2}} \Big(\frac{(M_{\text{full};\nu})^\nu}{\sigma_{\text{full}}^\nu} + \frac{((n-1)^{1/2} M_{\text{cond};\nu})^\nu}{\sigma_{\text{full}}^\nu} \Big)^{\frac{1}{2\nu+1}} \Big) .$$

Remark 3.2. In the case $\nu = 3$, the error term above becomes

$$C\Big(\frac{1}{(n-1)^{1/5}} + \Big(\frac{\sqrt{n-1}\,\sigma_{\text{cond}}}{\sigma_{\text{full}}} \Big)^{2/5} + n^{-\frac{1}{14}} \Big(\frac{(M_{\text{full};3})^3}{\sigma_{\text{full}}^3} + \frac{\big((n-1)^{1/2} M_{\text{cond};3}\big)^3}{\sigma_{\text{full}}^3} \Big)^{\frac{1}{7}} \Big) .$$

In the case when Assumption 3.1 holds for ν, the error term is $\Theta\big(\big(\frac{n-1}{\rho_d^2}\big)^{1/5} + n^{-\frac{\nu-2}{4\nu+2}} \big)$.

Proposition 3.5 agrees with the classical results for degenerate U-statistics. In those results, $(\phi_k^{(K)})$ are chosen such that they are orthonormal in $L_2(\mathbb{R}^d, R)$ and $\mathbb{E}[\phi_k^{(K)}(Y_1)] = 0$. This corresponds to Σ^K being a diagonal matrix and the expression for $\tau_k^{(K)}$ can be simplified. In the high-dimensional regime, Proposition 3.5 says that the degenerate approximation holds so long as $\rho_d = \omega(n^{1/2})$.

Proposition 3.5 allows us to obtain a better understanding of the asymptotic behavior of D_n in the case $\rho_d = \omega(n^{1/2})$. To see this, write $W_2 := \lim_{K\to\infty} W_n^K$ as the distributional limit of W_n^K as $K \to \infty$ (for fixed n and d). Provided that W_2 exists, Proposition 3.5 says that we may approximate D_n by W_2 in the Kolmogorov metric. The next proposition guarantees the existence of W_2.

Proposition 3.6. *Fix n, d. Suppose Assumption 3.2 holds for some $\nu \geq 2$ and that $|D|, \sigma_{\text{full}} < \infty$. Then W_2 exists.*

It now suffices to analyse W_2. In Section 3.1, we have seen that in the case of a linear kernel, the degenerate limit may become asymptotically Gaussian as $d \to \infty$. This is connected to the fourth moment theorem of Nualart and Peccati (2005): As a special case, their result implies that a sequence of polynomials of Gaussians is asymptotically Gaussian if and only if its limiting excess kurtosis is zero (see Section 4.4 for a formal statement). Since W_n^K is a degree-two polynomial of Gaussians parameterised by K, n and d, their result applies to W_n^K. Moreover, the limiting moments of W_n^K can be computed easily when Assumption 3.2 holds for $\nu \geq 4$, since they depend only on moments of the original U-statistic D_n and *not* on specific values of the intractable weights $\tau_k^{(K)}$. Lemma A.9 in the appendix shows that

(i) $\mathbb{E}[W_n^K] = D$ for every $K \in \mathbb{N}$,
(ii) $\lim_{K\to\infty} \text{Var}[W_n^K] = \frac{2}{n(n-1)}\sigma_{\text{full}}^2$, and
(iii) $\lim_{K\to\infty} \mathbb{E}\big[(W_n^K - D)^4\big] = \frac{12(4\mathbb{E}[u(X_1,X_2)u(X_2,X_3)u(X_3,X_4)u(X_4,X_1)] + \sigma_{\text{full}}^4)}{n^2(n-1)^2}$,

provided that Assumption 3.2 holds for $\nu \geq 1$, $\nu \geq 2$ and $\nu \geq 4$ respectively. Upon taking the additional asymptotic as $n \to \infty$, if the excess kurtosis is indeed zero, Gaussian is still the correct limiting distribution for D_n, but now with a *larger* variance (described by σ^2_{full}) than what one may have predicted by the Gaussian CLT limit for non-degenerate U-statistics (described by $(n-1)^{1/2}\sigma_{\text{cond}}$ in Section 3.2.1).

Meanwhile, when the limiting excess kurtosis is not zero, the limiting distribution is an infinite sum of weighted chi-squares. A naive example is the following:

Lemma 3.7. *Suppose $\lambda_k^{(K)} = \lambda_k$ is independent of K and there exists a finite K_* such that $\lambda_k = 0$ for all $k > K_*$. Then $W_2^{K^*}$ converges weakly to a weighted sum of independent chi-squares as $K \to \infty$.*

A weighted sum of chi-squares does not admit a closed-form distribution function. Fortunately in the case when $\tau_k^{(K)} \geq 0$ for all k, many numerical approximation schemes are available and used widely. These methods generally rely on matching the moments of W_n, which can be computed easily due to Proposition 3.6. The simplest example is the Welch-Satterthwaite method, which approximates the distribution of W_n by a gamma distribution with the same mean and variance, and is employed in our Figure 3.2 to demonstrate the degenerate limit. We refer readers to Bodenham and Adams (2016) and Duchesne and De Micheaux (2010) for a review of other moment-matching methods.

3.3 Matching upper and lower bounds for specific U-statistics

Despite their general applicability, Theorem 3.1 and Proposition 3.5 both yield an error bound on the order $n^{-1/14}$ (provided that a third moment exists), in contrast with the $O(n^{-1/2})$ error for non-degenerate approximation in Section 3.2.1. This calls into question whether the $n^{-1/14}$ bound is improvable. It turns out that the error bound for quadratic form approximations of D_n is nuanced even in the classical case: Known upper bounds depend on the number of non-zero eigenvalues of the Hilbert-Schmidt operator associated with the kernel u of the U-statistic, ranging from $O(n^{-1})$ for five non-zero eigenvalues (Götze and Zaitsev, 2014), $O(n^{-1/12})$ for one non-zero eigenvalue (Yanushkevichiene, 2012) (and control of an $18/5$th moment), to our $O(n^{-1/14})$ bound in Theorem 3.1 with no eigenvalue assumptions. Yanushkevichiene (2012) also conjectures that the $n^{-1/12}$ rate is unimprovable for degree-two U-statistics in view of a construction by Senatov (1998).

Our next result shows that, for every $\gamma \in (0, \frac{1}{12}]$, there exists a degree-two U-statistic $u_n(X) = \frac{1}{n(n-1)} \sum_{i \neq j} k_u(X_i, X_j)$ on i.i.d. $\mathbb{R}^{d(n)}$-valued vectors $(X_i)_{i \leq n}$, such that the approximation error by a quadratic form of Gaussians is $\Theta(n^{-\gamma})$.

Theorem 3.8. *Fix $\nu \in (2, 3]$. Let χ_1^2 be a chi-squared random variable with 1 degree of freedom, $\xi \sim \mathcal{N}(0, 1)$ be independent of χ_1^2 and $\overline{\chi_1^2} = \chi_1^2 - 1$. There exist some absolute constants $c, C > 0$, $N \in \mathbb{N}$ and a sequence $(\sigma_n)_{n\in\mathbb{N}}$ with $\sigma_n \to 0$, as well as some random vectors (X_i) and a symmetric function k_u that depends on σ_n, such that for all $n > N$ and $d(n) \in \mathbb{N}$,*

$$cn^{-\frac{\nu-2}{4\nu}} \leq \sup_{t\in\mathbb{R}} \left|\mathbb{P}\left(\sqrt{n(n-1)}\, u_n(X) \leq t\right) - \mathbb{P}\left(\sigma_n\xi + \overline{\chi_1^2} \leq t\right)\right| \leq Cn^{-\frac{\nu-2}{4\nu}}.$$

Remark. (i) In the construction, we choose σ_n to decay as $\Theta\left(n^{-\frac{\nu-2}{2\nu}}\right)$. The approximation becomes a chi-squared approximation in the limit $n \to \infty$, but at a very slow rate. (ii) Since ξ can be obtained as the limiting distribution of a partial sum of weighted and centred chi-squared variables, Theorem 3.8 can be read as a result on the approximation of $u_n(X)$ by infinite sums of weighted and shifted chi-squares.

In the case when a third moment exists ($\nu = 3$), the approximation error is exactly $\Theta(n^{-1/12})$, which answers the question from Yanushkevichiene (2012) in the affirmative. An implication is that, without additional structural assumptions on the data distribution or the function u used in the U-statistic, the slow $n^{-1/12}$ rate of quadratic-form-of-Gaussian approximation for U-statistics is not improvable, and the $n^{-1/14}$ rate of Theorem 3.1 and Proposition 3.5 are not too far from being worst-case optimal.

We conclude with a few comments on the proofs, which are included in Appendix B.3. The U-statistic used in Theorem 3.8 is a special case of a construction developed for more generic polynomials in Section 4.5, inspired by a result of Senatov (1998). As such, we defer the full construction and the proof technique for the lower bound to Section 4.5.1. The upper bound of Theorem 3.8 improves upon the more general result of Theorem 3.1 by using an argument specific to the construction, instead of applying the Lindeberg method. Due to the similarity of proof techniques, an analogous result as Theorem 3.8 holds for V-statistics; we include this as Theorem B.1 in the appendix.

3.4 Distribution tests with Maximum Mean Discrepancy and Kernel Stein Discrepancy

In this section, we study the implications of the universality results (Section 3.2) and how the different asymptotic limits manifest in high-dimensional distribution tests. Given two probability measures P and Q on $\mathbb{R}^d$, we consider the problem of testing $H_0 : P = Q$ against $H_1 : P \neq Q$ through some measure of discrepancy between P and Q. We focus on *Maximum Mean Discrepancy* (MMD) and *(Langevin) Kernelized Stein Discrepancy* (KSD), two kernel-based methods that use a U-statistic D_n as the test statistic.

For MMD and KSD, it is well-known that $\sigma_{\text{cond}} = 0$ under H_0 and the limit of D_n is an infinite sum of weighted and centred chi-squares (see Gretton et al. (2012) for MMD and Liu et al. (2016) for KSD). As discussed in Sections 3.1 and 3.2.3, the infinite sum itself may have a Gaussian limit depending on the limiting excess kurtosis of the infinite sum, which in turn depends on the weights. Since the weights and hence the limiting distribution is intractable in general, a common practice is to simulate D_n under H_0 by distribution-agnostic methods such as a permutation test or a wild bootstrap (Schrab et al., 2023). As such, we do not focus on the distribution of D_n under H_0 here.

Instead, we are interested in quantifying the power of D_n given as $\mathbb{P}_{H_1}(D_n > t)$. The test threshold t is often chosen adaptively in practice, but we assume t to be fixed for simplicity of analysis. Classically for dimension d fixed, it has been shown that for MMD and KSD, D_n has $\sigma_{\text{cond}} > 0$ under H_1 and its limiting distribution is typically taken as a Gaussian (Gretton et al., 2012; Liu et al., 2016), which is used to characterize the asymptotic power. Those results cease to hold in the high-dimensional regime, and our results in Section 3.2 offer two insights to this problem:

(i) Depending on the variance ratio ρ_d, D_n may not always have the non-degenerate Gaussian distribution as its limit. In the non-Gaussian case, the confidence interval and thereby the distribution curve can be wider than what a Berry-Esseen bound predicts, and there may be potential asymmetry;

(ii) We can completely characterise the high-dimensional behaviour of the power in terms of ρ_d, which in turn depends on the hyperparameters and the set of alternatives considered.

In what follows, we introduce additional notation in Section 3.4.1 and show, in Section 3.4.2, that our results naturally apply to MMD and KSD. We then investigate their high-dimensional behaviours in an example of Gaussian mean-shift under simple kernels in Section 3.4.3.

3.4.1. Notation

We follow the kernel definition from Steinwart and Scovel (2012) as below:

Definition 3.9. A function $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is called a *kernel* on $\mathbb{R}^d$ if there exists a Hilbert space $(\mathcal{H}, \langle \bullet, \bullet \rangle_{\mathcal{H}})$ and a map $\phi : \mathbb{R}^d \to \mathcal{H}$ such that $\kappa(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$ for all $x, x' \in \mathcal{H}$.

We give the minimal definitions of MMD and KSD, and refer interested readers to Gretton et al. (2012) and Gorham and Mackey (2017) for further reading. Throughout, we let $\{Y_j\}_{j=1}^n$ be i.i.d. samples from P and $\{X_i\}_{i=1}^n$ be i.i.d. samples from Q. We also

write $Z_i := (X_i, Y_i)$ and assume that κ is measurable. MMD with respect to κ is defined by

$$D^{\mathrm{MMD}}(Q, P) \;:=\; \mathbb{E}_{Y,Y' \sim P}[\kappa(Y, Y')] - 2\mathbb{E}_{Y \sim P, X \sim Q}[\kappa(Y, X)] + \mathbb{E}_{X,X' \sim Q}[\kappa(X, X')] \; .$$

A popular unbiased estimator for D^{MMD} is exactly a U-statistic:

$$D_n^{\mathrm{MMD}} \;:=\; \frac{1}{n(n-1)} \sum\nolimits_{1 \le i \ne j \le n} u^{\mathrm{MMD}}(Z_i, Z_j) \; ,$$

where the summand is given by $u^{\mathrm{MMD}}\big((x, y), (x', y')\big) := \kappa(x, x') + \kappa(y, y') - \kappa(x, y') - \kappa(x', y)$. To define KSD, we assume that κ is continuously differentiable with respect to both arguments, and P admits a continuously differentiable, positive Lebesgue density p. The following formulation of KSD is due to Theorem 2.1 of Chwialkowski et al. (2016):

$$D^{\mathrm{KSD}}(Q, P) \;:=\; \mathbb{E}_{X,X' \sim Q}[u_P^{\mathrm{KSD}}(X, X')] \; ,$$

where we assume $\mathbb{E}_{X \sim Q}[u_P^{\mathrm{KSD}}(X, X)] < \infty$ and the function $u_P^{\mathrm{KSD}} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is given by

$$\begin{aligned} u_P^{\mathrm{KSD}}(x, x') = \big(\nabla \log p(x)\big)^\top \big(\nabla \log p(x')\big) \kappa(x, x') \;+\; \big(\nabla \log p(x)\big)^\top \nabla_2 \kappa(x, x') \\ + \big(\nabla \log p(x')\big)^\top \nabla_1 \kappa(x, x') \;+\; \mathrm{Tr}(\nabla_1 \nabla_2 \kappa(x, x')) \; . \end{aligned}$$

∇_1 and ∇_2 are the differential operators with respect to the first and second arguments of κ respectively. The estimator is again a U-statistic, given by

$$D_n^{\mathrm{KSD}} \;:=\; \frac{1}{n(n-1)} \sum\nolimits_{1 \le i \ne j \le n} u_P^{\mathrm{KSD}}(X_i, X_j) \; .$$

3.4.2. Verification of Assumption 3.2 for MMD and KSD

It turns out that a kernel structure allows Assumption 3.2 to be fulfilled under some natural conditions. Let $V_1, V_2 \overset{i.i.d.}{\sim} R$ for some probability measure R on $\mathbb{R}^b$ and κ^* be a measurable kernel on $\mathbb{R}^b$. A sequence of functions $\{\phi_k\}_{k=1}^\infty$ in $L_2(\mathbb{R}^b, R)$ and a sequence of non-negative values $\{\lambda_k\}_{k=1}^\infty$ with $\lim_{k \to \infty} \lambda_k = 0$ is called a *weak Mercer representation* if

$$\Big| \sum\nolimits_{k=1}^K \lambda_k \phi_k(V_1) \phi_k(V_2) - \kappa^*(V_1, V_2) \Big| \to 0 \quad \text{almost surely} \qquad \text{as } K \to \infty \; .$$

Steinwart and Scovel (2012) show that such a representation exists if $\mathbb{E}[\kappa^*(V_1, V_1)] < \infty$, whose result is summarised in Lemma A.13 in the appendix. To deduce from this the L_ν convergence of Assumption 3.2, we need the following assumptions on the kernel κ^*:

Assumption 3.3. Fix $\nu > 2$. Assume $\mathbb{E}[\kappa^*(V_1, V_1)] < \infty$ and let $\{\lambda_k\}_{k=1}^\infty$ and $\{\phi_k\}_{k=1}^\infty$ be a weak Mercer representation of κ^* under R. Also assume that for some $\nu^* > \nu$, $\|\kappa^*(V_1, V_2)\|_{L_{\nu^*}} < \infty$ and $\sup_{K \ge 1} \| \sum_{k=1}^K \lambda_k \phi_k(V_1) \phi_k(V_2) \|_{L_{\nu^*}} < \infty$.

For MMD, we can use the weak Mercer representation of u^{MMD} to show that our

results apply:

Lemma 3.10. *u^{MMD} defines a kernel on $\mathbb{R}^{2d}$. Moreover, if Assumption 3.3 holds for $\kappa^* = u^{\mathrm{MMD}}$ under $P \otimes Q$ for some $\nu > 2$, then Assumption 3.2 holds for $\min\{\nu, 3\}$ with $u = u^{\mathrm{MMD}}$ and $R = P \otimes Q$.*

In the case of KSD, we use the representation of κ directly. We require some additional assumptions for the score function $\nabla \log p(x)$ to be well-behaved and the differential operation on κ to behave well under the representation.

Assumption 3.4. Fix n, d and $\nu > 2$. Assume that Assumption 3.3 holds with ν for κ under Q, with $\{\lambda_k\}_{k=1}^{\infty}$ and $\{\phi_k\}_{k=1}^{\infty}$ as the weak Mercer representation of κ under Q and ν^* being defined as in Assumption 3.3. Further assume that

(i) $\| \|\nabla \log p(X_1)\|_2 \|_{L_{2\nu^{**}}} < \infty$ for $\nu^{**} = \frac{\nu(\nu+\nu^*)}{\nu^*-\nu}$;

(ii) $\sup_{k\in\mathbb{N}} \|\phi_k(X_1)\|_{L_{2\nu}} < \infty$;

(iii) ϕ_k's are differentiable with $\sup_{k\in\mathbb{N}} \| \|\nabla\phi_k(X_1)\|_2 \|_{L_\nu} < \infty$;

(iv) As $K \to \infty$, we have the convergence

$$\left\| \left\| \sum_{k=1}^{K} \lambda_k (\nabla\phi_k(X_1))\phi_k(X_2) - \nabla_1\kappa(X_1, X_2) \right\|_2 \right\|_{L_{2\nu}} \to 0 \,,$$
$$\left\| \sum_{k=1}^{K} \lambda_k (\nabla\phi_k(X_1))^\top (\nabla\phi_k(X_2)) - \mathrm{Tr}(\nabla_1\nabla_2\kappa(X_1, X_2)) \right\|_{L_\nu} \to 0 \,.$$

We can now form a decomposition of u_P^{KSD}. Given $\{\lambda_k\}_{k=1}^{\infty}$ and $\{\phi_k\}_{k=1}^{\infty}$ from Assumption 3.4 and any fixed $d \in \mathbb{N}$, define the sequences $\{\alpha_k\}_{k=1}^{\infty}$ and $\{\psi_k\}_{k=1}^{\infty}$ as, for $1 \le l \le d$ and $k' \in \mathbb{N}$,

$$\alpha_{(k'-1)d+l} := \lambda_{k'} \quad \text{and} \quad \psi_{(k'-1)d+l}(x) := (\partial_{x_l} \log p(x))\phi_{k'}(x) + \partial_{x_l}\phi_{k'}(x) \,. \tag{3.10}$$

Lemma 3.11. *If Assumption 3.4 holds for some $\nu > 2$, then Assumption 3.2 holds for $\min\{\nu, 3\}$ with $u = u_P^{\mathrm{KSD}}$, $R = Q$, $\lambda_k^{(K)} = \alpha_k$ and $\phi_k^{(K)} = \psi_k$.*

Remark. We do remark that Assumption 3.4, in particular (iv), can be difficult to verify for specific kernels. We present it here only to illustrate how our Assumption 3.2 can be related to the use of Mercer representation in KSD analysis; as we discuss in Appendix A.1, it can be much more straightforward to verify Assumption 3.2 directly.

The benefits of formulating our results in terms of Assumption 3.2 are now clear: By forgoing orthonormality, we can choose a functional decomposition e.g. in terms of the Mercer representation of a kernel, which is already widely considered in this literature. The non-negative eigenvalues from the Mercer representation (Lemma A.13) also allow moment-matching methods discussed in Section 3.2.3 to be considered. In fact, a Mercer representation is not even necessary, as there are generally many non-unique choices of

$(\phi_k^{(K)}, \lambda_k^{(K)})$ in Assumption 3.2. In Appendix A.1.1 in the appendix, we show that for the setup with RBF kernel in Section 3.4.3, we can verify Assumption 3.2 easily on a decomposition obtained by Taylor expansions.

3.4.3. Gaussian mean-shift examples

We study KSD and MMD under Gaussian mean-shift, where $P = \mathcal{N}(0, \Sigma)$ and $Q = \mathcal{N}(\mu, \Sigma)$ with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d\times d}$ to be specified. Two simple kernels are considered in this section, namely the *RBF* kernel and the *linear* kernel. The mathematical results in this section can be found in Section 4.3 in the joint work (Huang, Liu, Duncan, and Gandy, 2023). As Propositions 3.12 and 3.13 are obtained mainly via cumbersome moment computations in each of the special cases, we refer interested readers to Huang, Liu, Duncan, and Gandy (2023) for their proofs.

RBF kernel. We consider the RBF kernel $\kappa(x, x') = \exp(-\|x - x'\|_2^2/(2\gamma))$, where $\gamma = \gamma(d)$ is a bandwidth potentially depending on d. A common strategy to choose γ is the *median heuristic*:

$$\gamma_{\text{med}} \coloneqq \text{Median}\left\{\|V - V'\|_2^2 : V, V' \in \mathcal{V},\ V \neq V'\right\},$$

where the samples $\mathcal{V} = \{X_i\}_{i=1}^n$ for KSD and $\mathcal{V} = \{X_i\}_{i=1}^n \cup \{Y_i\}_{i=1}^n$ for MMD. In Appendix A.1, we include a further discussion of this setup as well as the verification of Assumption 3.2. We refer interested readers to Appendix A of Huang et al. (2023) for a discussion and results on the verification of Assumption 3.1.

We focus on $\Sigma = I_d$, where the dimension dependence of the moment ratio ρ_d can be explicitly studied for both KSD and MMD. Importantly, we give bounds in terms of the bandwidth γ and the scale of mean shift $\|\mu\|_2^2$, which reveal their effects on ρ_d and thereby on the behaviour of the test power. The assumptions on γ and $\|\mu\|_2^2$ in both propositions are for simplicity rather than necessity.

Proposition 3.12 (KSD-RBF moment ratio)**.** *Assume $\gamma = \omega(1)$ and $\|\mu\|_2^2 = \Omega(1)$. Under the Gaussian mean-shift setup with $\Sigma = I_d$, the KSD U-statistic satisfies that*

(i) If $\gamma = o(d^{1/2})$, then $\rho_d = \exp\left(\frac{3d}{4\gamma^2} + o\left(\frac{d}{\gamma^2}\right)\right)\Theta\left(\frac{d}{\gamma\|\mu\|_2^2} + \frac{d^{1/2}}{\gamma^{1/2}\|\mu\|_2} + 1\right)$;

(ii) If $\gamma = \omega(d^{1/2})$, then $\rho_d = \Theta\left(\frac{d^{1/2}(1+\gamma^{-1/2}\|\mu\|_2)}{\|\mu\|_2\,(1+\gamma^{-1}d^{1/2}\|\mu\|_2)} + 1\right)$;

(iii) If $\gamma = \Theta(d^{1/2})$, then $\rho_d = \Theta\left(\frac{d^{1/2}}{\|\mu\|_2^2} + \frac{d^{1/4}}{\|\mu\|_2} + 1\right)$.

Proposition 3.13 (MMD-RBF moment ratio)**.** *Consider the Gaussian mean-shift setup with $\Sigma = I_d$ and assume $\gamma = \omega(1)$ and $\|\mu\|_2^2 = \Omega(1)$. For the MMD U-statistic, if*

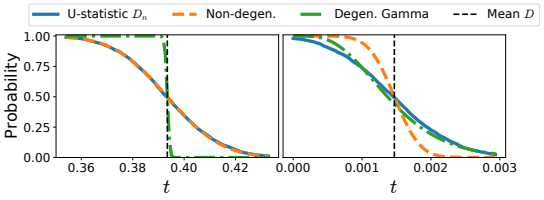


Figure 3.3: Behaviour of $\mathbb{P}(X > t)$ for $X = D_n^{\mathrm{MMD}}$ with the RBF kernel versus X being the theoretical limits. *Left.* $n = 1000$ and $d = 2$. *Right.* $n = 50$ and $d = 1000$.

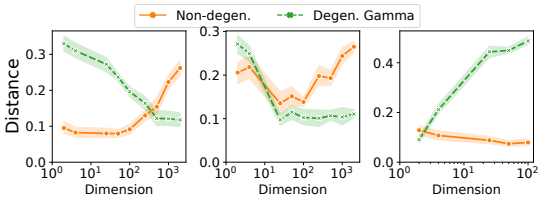


Figure 3.4: L_∞ distance between the c.d.f. of D_n^{KSD} with RBF and those of the theoretical limits as d varies. *Left.* $n = 50$ fixed (high dimensions). *Middle.* $n = \Theta(d^{1/2})$ (high dimensions). *Right:* $n = \Theta(d^2)$ (low dimensions).

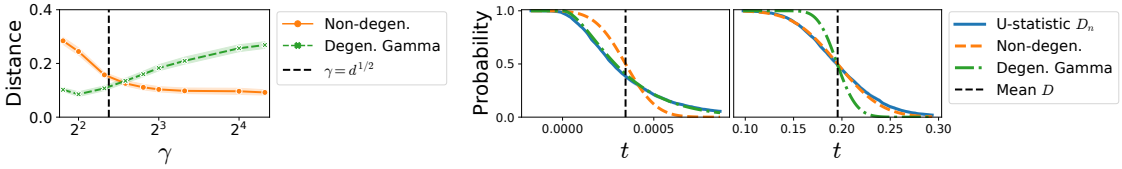


Figure 3.5: Behaviour of $\mathbb{P}(D_n^{\mathrm{KSD}} > t)$ with RBF as γ varies for $n = 50$ and $d = 27$. *Left.* L_∞ distance between the c.d.f. of D_n^{KSD} and the theoretical limits. *Middle.* Distribution curves at $\gamma = 4$. *Right.* Distribution curves at $\gamma = 16$.

$\gamma = o(\|\mu\|_2^2)$ and $\gamma = o(d^{1/2})$, then $\rho_d = \Theta\big(\exp\big(\frac{3d}{4\gamma^2} + o\big(\frac{d}{\gamma^2}\big)\big)\big)$. If instead $\gamma = \omega(\|\mu\|_2^2)$, then

(i) For $\gamma = o(d^{1/2})$, we have $\rho_d = \Theta\Big(\frac{\gamma}{\|\mu\|_2^2} \exp\Big(\frac{3d}{4\gamma^2} + o\big(\frac{d}{\gamma^2}\big)\Big)\Big)$;

(ii) For $\gamma = \omega(d^{1/2})$, we have $\rho_d = \Theta\Big(\frac{\|\mu\|_2 + d^{1/2}}{\|\mu\|_2 + \gamma^{-1} d^{1/2} \|\mu\|_2^2}\Big)$;

(iii) For $\gamma = \Theta(d^{1/2})$, we have $\rho_d = O\Big(\frac{d^{1/2}}{\|\mu\|_2^2}\Big)$.

The case $\|\mu\|_2 = \Omega(\|\Sigma\|_2) = \Omega(d^{1/2})$ is not very interesting, as it means that the signal-to-noise ratio (SNR) is high and can even increase with d. WLOG we focus on a low SNR setting with $\|\mu\|_2 = \Theta(1)$. In this case, it has been shown that the median-heuristic bandwith scales as $\gamma_{\mathrm{med}} = \Theta(d)$ (Reddi et al., 2015; Ramdas et al., 2015; Wynne and Duncan, 2022). While Propositions 3.12 and 3.13 do not directly address the case $\gamma = \gamma_{\mathrm{med}}$ due to its data dependence, they do show that $\rho_d = \Theta(d^{1/2})$ for both KSD and MMD with a data-*independent* bandwidth $\gamma = \Theta(d)$[†]. In this case, the asymptotic distributions of D_n^{KSD} and D_n^{MMD} are *(i)* the non-degenerate Gaussian limit from Section 3.2.1 when $d = o(n)$ and *(ii)* the degenerate limit from Proposition 3.5 when $d = \omega(n)$.

Intriguingly, in both results, different regimes arise based on how γ compares with the noise scale $\|\Sigma\|_2 = d^{1/2}$. In fact, a change from one asymptotic regime to the other as γ drops from $\omega(d^{1/2})$ to $o(d^{1/2})$ has been reported in Ramdas et al. (2015) but with

[†]In our experiments, the data-independent choice $\gamma = d$ and the data-dependent $\gamma = \gamma_{\mathrm{med}}$ yield almost identical plots.

no further comments[‡§]. Our results offer one explanation: Such transitions may happen due to a change in the dependence of ρ_d on γ, $\|\mu\|_2$ and d. Figure 3.5 shows a transition across different limits as γ varies, where the transition occurs at around $\gamma \sim d^{1/2}$.

Linear kernel. Section 3.2.3 discussed that the limit of D_n can be non-Gaussian. This is true for MMD with a *linear* kernel $\kappa(x, x') = x^\top x'$ (which, notably, is different from the U-statistic with the linear kernel in Section 3.1). In this case, D_n satisfies Lemma 3.7 with $K_* = d$ and the limit is a shifted-and-rescaled chi-square. Figure 3.2 verifies this for some $\Sigma \neq I_d$ by showing an asymmetric distribution curve close to the chi-square limit. We remark that a linear kernel, while not commonly used, is a valid choice here since $D^{\mathrm{MMD}} = 0$ iff $P = Q$ under our setup.

Simulations. We set $\mu = (2, 0, \ldots, 0)^\top \in \mathbb{R}^d$, $\Sigma = I_d$ and $\gamma = \gamma_{\mathrm{med}}$ for KSD with RBF and MMD with RBF. The exact setup for MMD with linear kernel is described in Appendix A.1.4. The limits for comparison are the non-degenerate Gaussian limit in (3.7) ("Non-degen.") and Gamma / shifted-and-rescaled chi-square ("Degen. Gamma" / "Degen. Chi-square") distributions that match the degenerate limit in Proposition 3.5 by mean and variance. Figure 3.2 plots the distribution curves for KSD with RBF and MMD with linear kernel. Figure 3.3 plots the same quantity for MMD with RBF. Figure 3.4 and Figure 3.5 examine the behaviour of KSD with RBF as d or γ varies (as a data-independent function of d, similar to Ramdas et al. (2015)). Results involving D_n are averaged over 30 random seeds, and shaded regions are 95% confidence intervals[¶]. See Huang, Liu, Duncan, and Gandy (2023) for further experiment details and code.

[‡]Their bandwidth γ_{Ramdas} is defined to equal our $\sqrt{2\gamma}$. The change in asymptotic regime occurs at $\gamma_{\mathrm{Ramdas}} = d^{1/4}$ in their Figure 1. While their figure is for MMD with threshold chosen by a permutation test, ours is for KSD with a fixed threshold.

[§]This was investigated in Ramdas (2015, Section 10.4) in a special case when $\gamma = \omega(\|\mu\|_2^2 + d)$ (case *(ii)* of Proposition 3.13) and $n = o(d^{5/2})$, where the author derived the test power of the RBF-kernel MMD for different SNRs.

[¶]The shaded regions are not visible for $\mathbb{P}(D_n > t)$ in Figure 3.2, 3.3 and 3.5 as the confidence intervals are very narrow.

Chapter 4

General results on universality

The main technical tool behind the degree-two U-statistics results in Chapter 3 is a Gaussian universality result for degree-two polynomial of high-dimensional random vectors, which already leads to several unexpected observations. This leads to the question whether these results can be extended to degree-m polynomials, especially when $m = m(n)$ is also allowed to grow in n in addition to the dimension $d = d(n)$.

In this chapter, we first establish universality results for the case where the function f in (1.1) is a degree m-polynomial. A key implication of the results in this chapter is that the degree m, instead of the dimension d, presents a fundamental barrier for Gaussian universality results. This provides an answer raised in Chapter 1 on the exact characterisation of a class of functions for which universality holds, and complements the rich body of works discussed in Chapter 1 that establish universality results on a case-by-case basis. Additionally, we provide results or discussions that

(i) extend universality to approximately polynomial functions, as well as strictly monotonic functions of these approximate polynomials;

(ii) discuss implications of block dependence on universality;

(iii) show that, contrary to the intuition in the classical setting (Chapter 2), the rate provided by the Lindeberg method is not improvable in a general setup.

The generality of the results in this section is an immediate consequence of (i): Informally, given an estimator $f(X_1, \dots, X_n)$, if one can find an appropriate strictly monotonic function h such that $f(X_1, \dots, X_n) \approx h(q(X_1, \dots, X_n))$ and that q can is well-approximated by some low-degree Taylor expansion p_m, then Gaussian universality applies, i.e. we can replace X_i's by Gaussians. We make the notion of approximation precise in Theorem 4.2. A special case of this heuristic, taking h to be the identity, is the delta method; we defer to Section 5.2 to show that an application of our result gives a generalisation of the delta method for high-dimensional data.

The rest of the chapter is organised as follows. Section 4.1 introduces the setup and notation. Section 4.2 provides upper bounds for an exact polynomial (Theorem 4.1) and an approximately polynomial function (Theorem 4.2). Section 4.3 introduces variance

domination. This is a set of bounds that formalises the idea in Chapter 3 of "ignoring some additive part of the random variable if its variance is sufficiently small", and is used in many of our proofs including that of Theorem 4.2. Section 4.5 provides the lower bounds as well as the constructions used in their proofs.

4.1 Setup and additional notation

We first introduce our setup. Throughout this chapter and Chapter 5, we suppress dependence on n, and use the abbreviations

$$d \;=\; d(n) \qquad m \;=\; m(n) \qquad X_i \;=\; X_{ni} \qquad X \;:=\; (X_1, \ldots, X_n)\,.$$

The variables $X_1, \ldots, X_n$ are independent (but not necessarily identically distributed) random elements of $\mathbb{R}^d$, and p_m is a polynomial $\mathbb{R}^{nd} \to \mathbb{R}$ of degree m. The Gaussian surrogates are a collection Z of independent Gaussian vectors

$$Z \;:=\; (Z_1, \ldots, Z_n) \qquad \text{where} \quad Z_i \sim \mathcal{N}(\mathbb{E}[X_i], \mathrm{Var}[X_i])\,.$$

Our results rely on the Lindeberg method. As discussed in Section 2.2, the Lindeberg method is directly applicable if f_n is multilinear, but not if it is a polynomial involving higher powers of X_i. We reduce the polynomial to the multilinear case by augmenting the original random vectors: For each $i \leq n$, consider the centred tensor powers of X_i,

$$\bar{X}_i \;:=\; X_i - \mathbb{E}X_i\,, \quad \overline{X_i^{\otimes 2}} := X_i^{\otimes 2} - \mathbb{E}\big[X_i^{\otimes 2}\big]\,, \quad \ldots\,, \quad \overline{X_i^{\otimes m}} := X_i^{\otimes m} - \mathbb{E}\big[X_i^{\otimes m}\big]\,,$$

where we defer the precise notation of the tensor power to (4.4). Define the concatenated random tensor

$$\mathbf{X}_i \;:=\; \Big(\bar{X}_i\,,\, \overline{X_i^{\otimes 2}}\,,\, \ldots\,,\, \overline{X_i^{\otimes m}}\Big) \tag{4.1}$$

with dimension $D = d + d^2 + \ldots + d^m$. For a suitable tensor T_m of polynomial coefficients, we then have

$$p_m(X) - \mathbb{E}[p_m(X)] \;=\; \langle T_m\,,\, (1, \mathbf{X}_1)^\top \otimes \ldots \otimes (1, \mathbf{X}_n)^\top \rangle \;=:\; q_m(\mathbf{X}_1, \ldots, \mathbf{X}_n)\,. \tag{4.2}$$

See the end of this section for details on tensor notation. The function q_m is multilinear. As surrogates for the "augmented" variables $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$, we choose a collection of Gaussian vectors

$$\Xi \;:=\; (\xi_1, \ldots, \xi_n) \qquad \text{where} \quad \xi_1 \perp\!\!\!\perp \ldots \perp\!\!\!\perp \xi_n \quad \text{and} \quad \xi_i \sim \mathcal{N}(\mathbb{E}[\mathbf{X}_i], \mathrm{Var}[\mathbf{X}_i])\,.$$

Choosing the function f in (1.1) as either p_m or q_m then yields the approximations

$$p_m(X) \;\approx\; p_m(Z) \qquad \text{or} \qquad p_m(X) \;\approx\; \mathbb{E}[p_m(X)] + q_m(\Xi)\,.$$

Although q_m multiplies n coefficients, the tensor T_m is such that no resulting power exceeds m. The variable $q_m(\Xi)$ is thus a degree-m polynomial of Gaussians. The bound is given in terms of the moment terms

$$\sigma \coloneqq \sqrt{\operatorname{Var} q_m(\mathbf{X})} \qquad \text{and} \qquad M_{\nu;i} \coloneqq \|\partial_i q_m(\mathbf{W}_i)^\top \mathbf{X}_i\|_{L_\nu} , \tag{4.3}$$

where $\mathbf{W}_i \coloneqq (\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{0}, \xi_{i+1}, \ldots, \xi_n) \in \mathbb{R}^{nD}$ and $\|\bullet\|_{L_\nu} = (\mathbb{E}|\bullet|^\nu)^{1/\nu}$.

Tensor notation. By a tensor T of size $d_1 \times \ldots \times d_k$, we mean an element of $\mathbb{R}^{d_1 \times \ldots \times d_k}$. The tensor product of k vectors $x_i \in \mathbb{R}^{d_i}$ is the tensor $\otimes_i x_i$ with entries

$$\otimes_{i \le k} x_i(j_1, \ldots, j_k) \;=\; x_1(j_1) \cdots x_k(j_k) \quad \text{for } j_1 \le d_1, \ldots, j_k \le d_k , \tag{4.4}$$

and we write $x_1^{\otimes r} \coloneqq \otimes_{i \le r} x_1$ for short. The scalar product of two tensors of equal size is

$$\langle S, T \rangle \;\coloneqq\; \sum\nolimits_{i_1 \le d_1, \ldots, i_k \le d_k} S(i_1, \ldots, i_k) T(i_1, \ldots, i_k) \qquad \text{for } S, T \in \mathbb{R}^{d_1 \times \ldots \times d_k} ,$$

or equivalently the Euclidean scalar product in $\mathbb{R}^{d_1 \times \ldots \times d_k}$. Any polynomial p_m of degree $\le m$ can be represented as

$$p_m(x_1, \ldots, x_n) \;=\; T_0 + \sum\nolimits_{r_1 + \ldots + r_n \le m} \langle T_{r_1, \ldots, r_n} , x_1^{\otimes r_1} \otimes \ldots \otimes x_n^{\otimes r_n} \rangle ,$$

where T_0 is a scalar and each $T_{r_1, \ldots, r_n}$ is a tensor of size $d^{r_1} \times \ldots \times d^{r_n}$. Having defined the augmented random tensor $\mathbf{X}_i$ in (4.1), we can write p_m as

$$p_m(X) \;=\; \mathbb{E}[p_m(X)] \;+\; \langle T_m , (1, \mathbf{X}_1)^\top \otimes \ldots \otimes (1, \mathbf{X}_n)^\top \rangle \tag{4.5}$$

for a tensor T_m. T_m is determined completely by T_0 and $(T_{r_1, \ldots, r_n})_{r_1, \ldots, r_n \le m}$. Since p_m has degree m, T_m is such that for $m' > m$, all m'-fold products of $\mathbf{X}_1, \ldots, \mathbf{X}_n$ correspond to zero coefficients in T_m.

Throughout the thesis, we also make use of the tensor vectorisation operator. For a tensor T of size $d_1 \times \ldots \times d_k$, we define the corresponding flattened vector as

$$\operatorname{vec}(T) \;\coloneqq\; \big(T(1, \ldots, 1, 1) , T(1, \ldots, 1, 2) , \ldots , T(d_1, \ldots, d_{k-1}, d_k) \big)^\top \in \mathbb{R}^{d_1 \cdots d_k} .$$

4.2 Upper bounds

We present two results in this section. Theorem 4.1 gives an upper bound for the Gaussian universality approximation of the multilinear polynomial q_m, and Theorem 4.2 extends it to functions that are well-approximated by q_m. The approximation of $p_m(X) = q_m(\mathbf{X})$ directly by $p_m(Z)$, i.e. the universality result for the original, non-multilinear polynomial p_m, can be obtained as a special case of these results under mild conditions; we defer the formal result to Section 5.1.

Theorem 4.1 (Gaussian universality for polynomials)**.** *Fix $\nu \in (2, 3]$. Then there exists some absolute constant $C > 0$ such that*

$$\left|\mathbb{P}(\sigma^{-1} q_m(\mathbf{X}) \le t) - \mathbb{P}(\sigma^{-1} q_m(\Xi) \le t)\right| \;\le\; Cm \left(\frac{\sum_{i=1}^n M_{\nu;i}^\nu}{(1+t^2)^{\nu/2}\,\sigma^\nu}\right)^{\frac{1}{\nu m+1}}.$$

for every $n, m, d \in \mathbb{N}$, every $t \in \mathbb{R}$, and every $\sigma > 0$. Moreover, we have $\mathbb{E}[q_m(\mathbf{X})] = \mathbb{E}[q_m(\Xi)] = 0$ and $\mathrm{Var}[q_m(\mathbf{X})] = \mathrm{Var}[q_m(\Xi)]$.

Since the degree-two U-statistic in Chapter 3 can be approximated as a bilinear form of K-dimensional random vectors under Assumption 3.2, the universality result of Theorem 3.1 is a special case of Theorem 4.1. A few remarks on Theorem 4.1:

(i) *Proof technique.* Theorem 4.1 is proved by the Lindeberg method in Chapter 2. The main difference from the one-dimensional case is that we avoid applying the Cauchy-Schwarz inequality to any l_2 inner product of $\mathbb{R}^d$ vectors, in order to obtain a tighter dimension control. As we shall see in Chapter 5, this allows the bound to be well-controlled with martingale difference bounds that exploit the structure of the polynomial, which notably allows us to get mild to no dependence on the dimension d. We include the proof in Appendix C.2;

(ii) *Dimension dependence*. The upper bound depends on the dimension d only via the moment terms defined in (4.3). For many statistics of high-dimensional data, the moment ratio becomes independent of d (see Chapter 5). Intuitively, this is because the only effect of the input dimension $d(n)$ is to introduce dependence on n in the distribution of X_i's, which manifests in a Berry-Esseen type bound only through the first few moments;

(iii) *Asymptotic normality*. Since $q_m(\Xi)$ is a polynomial of Gaussians, results of Nualart and Peccati (2005) on the fourth moment phenomenon imply necessary and sufficient conditions for $q_m(\Xi)$ to be asymptotically Gaussian. See Section 4.4;

(iv) *Monotone transformations*. Theorem 4.1 also applies to strictly monotone functions of polynomials (since it controls a Kolmogorov distance, which is invariant under strictly monotone transforms). For example, an exponential of $q_m(\mathbf{X})$ can be approximated by an exponential of $q_m(\Xi)$. In other words, the result also applies to simple functions such as $x \mapsto \exp(x)$ that are not themselves approximatable by any degree-m polynomial;

(v) *Block dependence*. Let $(Y_{ij})_{i \le n, j \le k}$ be a block-dependent dataset of size $N = nk$, consisting of n blocks of k $\mathbb{R}^d$-valued data and where the data are independent across the n blocks but dependent within each block. By grouping each block of data into an $\mathbb{R}^{kd}$-dimensional vector and identifying each block as $X_i = (Y_{i1}, \ldots, Y_{ik})$, the upper bound of Theorem 4.1 still applies. However, to accommodate a growing block size k, one requires the dependency within each block to

be such that the moment ratio in Theorem 4.1 remains bounded. For example, consider an 1d empirical average $q_m(\mathbf{X}) = \frac{1}{\sqrt{n}\,k}\sum_{i=1}^{n}\sum_{j=1}^{k} Y_{ij}$, where the blocks are i.i.d. and each block $(Y_{i1},\dots,Y_{ik})$ is exchangeable with zero mean. Theorem 4.1 can be viewed as a result on the approximation of $q_m(\mathbf{X})$ by $q_m(\mathbf{Z}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i$, where $\xi_i \coloneqq \frac{1}{k}\sum_{j\le k} Z_{ij} \sim \mathcal{N}(0, \frac{1}{k}\mathrm{Var}[Z_{11}] + \frac{k-1}{k}\mathrm{Cov}[Z_{11}, Z_{12}])$. For $\nu = 3$, the moment ratio is some fractional power of

$$\frac{1}{\sqrt{n}}\,\frac{\|\frac{1}{\sqrt{k}}\sum_{j\le k} Y_{1j}\|_{L_3}^3}{\mathrm{Var}[\frac{1}{\sqrt{k}}\sum_{j\le k} Y_{1j}]^{3/2}} \;\le\; \frac{1}{\sqrt{n}}\,\frac{k^{3/2}\|Y_{11}\|_{L_3}^3}{(\mathrm{Var}[Y_{11}] + (k-1)\,\mathrm{Cov}[Y_{11}, Y_{12}])^{3/2}}\,.$$

If $\mathrm{Cov}[Y_{11}, Y_{12}] = 0$, e.g. because the dependence between Y_{11} and Y_{22} does not manifest through linear correlation, a sufficient condition for the moment ratio to be $o(1)$ is $k = o(n^{1/3})$.

We note that the dimension independence remark in (ii) (as well as in Chapter 3 and the subsequent Chapter 5) does not contradict the dimension constraint observed in (v): In all cases, dimensionality does not appear explicitly in the bound. While (v) illustrates a pathological case of how arbitrary dependence can lead to a dimensionality constraint, for the U-statistic considered in Chapter 3 with independent data, the moment ratio reduces to a (3^{rd} moment)/(2^{nd} moment) ratio on the U-statistic itself. This ratio is regarded as $O(1)$ for many practical U-statistics (Assumption 3.1). A similar assumption is made implicitly in order to interpret the bounds for the estimators in Chapter 5.

Let $L_2(X) = L_2(X_1,\dots,X_n)$ denote the space of square-integrable functions with respect to the probability measure of X. The polynomials of degree $\le m$ in $L_2(X)$ form a linear subspace of $L_2(X)$, and Theorem 4.1 provides universality approximations for functions in this subspace. A simple modification extends the result to functions that are close to the subspace in the L_2 norm:

Theorem 4.2 (Approximate polynomials)**.** *Fix $\nu \in (2,3]$. There exists some absolute constant $C > 0$ such that, for every $n, m, d \in \mathbb{N}$ and $\sigma > 0$,*

$$\sup_{t\in\mathbb{R}} \big|\mathbb{P}\big(\sigma^{-1}f(X) \le t\big) - \mathbb{P}\big(\sigma^{-1}q_m(\Xi) \le t\big)\big| \\ \le Cm\Big(\Big(\frac{\|f(X) - q_m(\mathbf{X})\|_{L_2}}{\sigma}\Big)^{\frac{2}{2m+1}} + \Big(\frac{\sum_{i=1}^{n} M_{\nu;i}^{\nu}}{\sigma^{\nu}}\Big)^{\frac{1}{\nu m+1}}\Big)$$

for every measurable function $f : \mathbb{R}^{nd} \to \mathbb{R}$.

Since $\mathbb{E}[q_m(\mathbf{X})] = 0$ by Theorem 4.1, the additional approximation error above is given in terms of a ratio of two L_2 norms,

$$\frac{\|f(X) - q_m(\mathbf{X})\|_{L_2}}{\sigma} \;=\; \frac{\|f(X) - q_m(\mathbf{X})\|_{L_2}}{\|q_m(\mathbf{X})\|}\,. \tag{4.6}$$

Theorem 4.2 plays a key role in determining which polynomial to use for approximating a general estimator. As a special case, the problem of determining the asymptotic of

a degree-two U-statistic in Chapter 3 is reduced to determining whether a degree-one Gaussian polynomial or a degree-two Gaussian polynomial dominates asymptotically. The ratio of variances arises as a determining factor precisely due to the approximation error (4.6). In Chapter 5, we will see how Theorem 4.2 can be applied to obtain a high-dimensional delta method, generalise known results on high-dimensional and infinite-order U-statistics, and extend results on fluctuations of subgraph counts.

4.3 Variance domination

One simple technique used extensively in our proofs and in particular, to prove Theorem 4.2, is the idea of variance domination: When computing the limit of a sum of two dependent, real-valued random variables $X' + Y'$, it suffices to ignore Y' in the limit provided that the variance of Y' is negligible compared to that of X'. The next result summarises the technique and is proved in Appendix B.4.

Proposition 4.3. *Let X' and Y' be two $\mathbb{R}$-valued, possibly dependent random variables with $\mathbb{E}[Y'] = 0$. Then for every $t \in \mathbb{R}$,*

$$\begin{aligned}\big|\mathbb{P}(X' + Y' \le t) - \mathbb{P}(X' \le t)\big| \\ \le \inf_{\epsilon > 0} \left(\max\left\{ \mathbb{P}(X' \in (t-\epsilon, t]) \,, \mathbb{P}(X' \in (t, t+\epsilon])\right\} + \frac{\mathrm{Var}[Y']}{\epsilon^2} \right) .\end{aligned}$$

If we further have $\sigma_{X'}^2 := \mathrm{Var}[X'] > 0$, then

$$\begin{aligned}\big|\mathbb{P}\big(\sigma_{X'}^{-1}(X' + Y') \le t\big) - \mathbb{P}\big(\sigma_{X'}^{-1} X' \le t\big)\big| \\ \le \inf_{\epsilon > 0} \left(\max\left\{ \mathbb{P}(\sigma_{X'}^{-1} X' \in (t-\epsilon, t]) \,, \mathbb{P}(\sigma_{X'}^{-1} X' \in (t, t+\epsilon])\right\} + \frac{\mathrm{Var}[Y']}{\mathrm{Var}[X']\,\epsilon^2} \right) .\end{aligned}$$

Remark 4.1. Several adaptations are useful: (i) To swap the roles of $X' + Y'$ and X', one may replace X' and Y' above by $X' + Y'$ and $-Y'$ respectively; (ii) To replace $\sigma_{X'}$ by another normalisation, e.g. $\sqrt{\mathrm{Var}[X' + Y']}$, one may rescale t and ϵ simultaneously.

Proposition 4.3 formalises the variance domination effect: If $\mathrm{Var}[Y']/\mathrm{Var}[X'] = o(1)$, by choosing $\epsilon = (\mathrm{Var}[Y']/\mathrm{Var}[X'])^{1/3}$, Proposition 4.3 implies that the c.d.f. difference $\big|\mathbb{P}(\sigma_{X'}^{-1}(X'+Y') \le t) - \mathbb{P}(\sigma_{X'}^{-1} X' \le t)\big| = o(1)$, provided that $\mathbb{P}(\sigma_{X'}^{-1} X' = t) = 0$. Meanwhile, to get a finite-sample control, one needs an anti-concentration bound on $\sigma_{X'}^{-1} X$. By the triangle inequality, it also suffices if $\sigma_{X'}^{-1} X$ is well-approximated in distribution by some random variable Z' with an anti-concentration bound. For polynomials of a random vector following a log-concave probability measure, a celebrated anti-concentration result is available due to Carbery and Wright (2001). This in particular applies to polynomials of Gaussian random vectors.

Fact 4.4 (Carbery-Wright inequality, Theorem 8 of Carbery and Wright (2001))**.** *Let $q_m(\mathbf{x})$ be a degree-m polynomial of $\mathbf{x} \in \mathbb{R}^d$ taking values in $\mathbb{R}$, and η be an $\mathbb{R}^d$-valued random vector following a log-concave probability measure. Then there exists a constant C independent of q_m, d, m or η such that, for every $\epsilon > 0$,*

$$\mathbb{P}\big(|q_m(\eta)| \leq \epsilon\big) \;\leq\; Cm\,\epsilon^{1/m}(\mathbb{E}[|q_m(\eta)|^2])^{-1/2m}\,.$$

Remark. We emphasise that under the Gaussian universality result of Theorem 4.1, we would only need to apply Fact 4.4 to Gaussian random vectors, which satisfies log-concavity immediately. While the proof of Theorem 4.1 also uses Fact 4.4, the argument is such that only anti-concentration bounds on a Gaussian polynomial is needed, which circumvents the need of checking log-concavity of the data distribution.

This immediately implies the following corollary of Proposition 4.3:

Corollary 4.5. *Let X', Y' and $\sigma_{X'} > 0$ be defined as in Proposition 4.3, and $q_m(\eta)$ be given as in Fact 4.4. Suppose $\mathrm{Var}[q_m(\eta)] = \mathrm{Var}\big[\sigma_{X'}^{-1}X'\big] = 1$. Then there is an absolute constant $C > 0$ such that for every $t \in \mathbb{R}$,*

$$\begin{aligned}&\sup_{t\in\mathbb{R}} \big|\mathbb{P}\big(X' + Y' \leq t\big) - \mathbb{P}\big(X' \leq t\big)\big| \\ &\qquad\leq\; Cm\Big(\frac{\mathrm{Var}[Y']}{\mathrm{Var}[X']}\Big)^{\frac{1}{2m+1}} + 2\sup_{t\in\mathbb{R}} \big|\mathbb{P}\big(\sigma_{X'}^{-1}X' \leq t\big) - \mathbb{P}\big(q_m(\eta) \leq t\big)\big|\,.\end{aligned}$$

Proof of Corollary 4.5. The result follows from combining Proposition 4.3 and Fact 4.4, noting that $\mathbb{E}[|q_m(\eta)|^2] \geq \mathrm{Var}[q_m(\eta)] = 1$, choosing $\epsilon = (\mathrm{Var}[Y']/\mathrm{Var}[X'])^{m/(2m+1)}$ and taking a supremum over $t \in \mathbb{R}$. □

Remark 4.2. The generality of Proposition 4.3 comes at a cost: It does not provide the tightest control on the c.d.f. difference in general, as the variance ratio term comes from Markov's inequality. When one has more information about the tail behaviour of $Y'/\sqrt{\mathrm{Var}[X']}$, the bound can usually be improved.

We conclude by remarking that, in the case $\mathbb{E}[f(X)] = 0$, Theorem 4.2 is essentially proved by identifying $X' = \sigma^{-1}q_m(\mathbf{X})$ and $Y' = \sigma^{-1}(f(X) - q_m(\mathbf{X}))$, and by applying the universality approximation bound of Theorem 4.1 to replace $q_m(\mathbf{X})$ by $q_m(\Xi)$. When $\mathbb{E}[f(X)] \neq 0$ and therefore $Y' \neq 0$, a minor adjustment is made to the bound in Proposition 4.3 to replace $\mathrm{Var}[Y']$ by $\|Y'\|_{L_2}$, which gives rise to the L_2 moment ratio in Theorem 4.2.

4.4 A necessary and sufficient condition for Gaussianity

To understand whether $p_m(X)$ may be asymptotically Gaussian, Theorem 4.1 reduces the problem to studying the asymptotic normality of $q_m(\Xi)$, a degree-m polynomial of nd i.i.d. univariate standard normal variables. This polynomial lives in the span of products of Hermite polynomials with total degree $\leq m$, and is therefore an m-th order Wiener chaos; see Nourdin (2013) for an introduction. If m is fixed, the fourth moment theorem by Nualart and Peccati (2005) applies and shows that $q_m(\Xi)$ is asymptotically Gaussian as $n \to \infty$ if and only if its excess kurtosis, defined as

$$\text{Kurt}[q_m(\Xi)] \;:=\; \mathbb{E}[(q_m(\Xi) - \mathbb{E}[q_m(\Xi)])^4] \,/\, \text{Var}[q_m(\Xi)]^2 - 3 \;=\; \sigma^{-4}\,\mathbb{E}[(q_m(\Xi))^4] - 3\,,$$

is asymptotically zero. Denote the total variation distance by d_{TV}. By directly using a finite-sample bound developed by Nourdin and Peccati (2015), the next result says that this is sufficient even as m grows, and is necessary under a uniform integrability condition.

Proposition 4.6 (Fourth moment phenomenon)**.** *Let η be a univariate standard normal variable independent of all other quantities. For every $n, m, d \in \mathbb{N}$ and $\sigma > 0$, we have*

$$d_{\text{TV}}(\sigma^{-1}q_m(\Xi)\,,\,\eta) \;\leq \left(\frac{4m-4}{3m}\left|\text{Kurt}[q_m(\Xi)]\right|\right)^{1/2}.$$

Under the high-dimensional asymptotic regime, as $n \to \infty$,

(i) $\text{Kurt}[q_m(\Xi)] \to 0$ *is a sufficient condition for* $\sigma^{-1}q_m(\Xi) \xrightarrow{d} \eta$*;*

(ii) if $\sigma^{-4}q_m(\Xi)^4$ *is uniformly integrable, then* $\text{Kurt}[q_m(\Xi)] \to 0$ *is also necessary.*

Proof of Proposition 4.6. The finite-sample bound is a restatement of Theorem 1.1 of Nourdin and Peccati (2015) for $q_m(\Xi)$, and directly implies (i). (ii) is proved by noting that, when $\sigma^{-1}q_m(\Xi) \xrightarrow{d} \eta$, by continuous mapping theorem, $\sigma^{-4}q_m(\Xi)^4 \xrightarrow{d} \eta^4$. Uniform integrability then implies the desired moment convergence (see Theorem 25.12 of Billingsley (1995)). □

Remark 4.3. Since the bound from Nourdin and Peccati (2015) works for a general sequence of Wiener chaos, Proposition 4.6 also holds if the Gaussian polynomial $\sigma^{-1}q_m(\Xi)$ is replaced by $\text{Var}[p_m(Z)]^{-1/2}(p_m(Z) - \mathbb{E}[p_m(Z)])$.

A consequence of Theorem 4.1 and Proposition 4.6 is that, even if $p_m(X)$ is not asymptotically Gaussian when d and m are fixed, it can become Gaussian in the high-dimensional regime. Similar phenomenon has already been observed for U-statistics (Janson and Nowicki, 1991; Bhattacharya et al., 2022). See Section 3.1 and Section 3.2.3 for a discussion on degree-two U-statistics of high-dimensional vectors.

4.5 Lower bound

A question not addressed by Theorem 4.1 is whether the bound is tight. Clearly, it may be suboptimal if additional information about f_n or the law of X is available. For example, choose X_i's as centred i.i.d. random elements of $\mathbb{R}$ with unit variance, $p_m(X)$ as the empirical average $\sum_{i=1}^n X_i/n$, and $\nu = 3$. The Berry-Esseen theorem then applies and guarantees a rate of $O(n^{-1/2})$, whereas Theorem 4.1 yields

$$\left|\mathbb{P}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i \le t\Big) - \mathbb{P}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i \le t\Big)\right| \;\le\; C\left(\frac{\|X_1\|_{L_3}}{\sqrt{n}\,(1+t^2)^{3/2}}\right)^{\frac{1}{4}} \;=\; O(n^{-1/8})\,.$$

If the only structure we know of our estimator $f(X)$ is that f is polynomial of degree $\le m$, however, the bound is essentially tight:

Theorem 4.7 (Lower bound)**.** *Fix $\nu \in (2,3]$, and assume that m is even with $m = o(\log n)$. Then there exist a sequence of probability measures $\mu^{(n)}_{\nu,\sigma_0}$, a sequence of polynomials $p^*_m = p^*_{m(n)}$, and absolute constants $c, C > 0$ and $N \in \mathbb{N}$, such that*

$$\begin{aligned} cn^{-\frac{\nu-2}{2\nu m}} \;&\le\; \sup_{t\in\mathbb{R}}\left|\mathbb{P}(p^*_m(X)\le t) - \mathbb{P}(p^*_m(Z)\le t)\right| \;\le\; Cmn^{-\frac{\nu-2}{2\nu m+2}}\,,\\ cn^{-\frac{\nu-2}{2\nu m}} \;&\le\; \sup_{t\in\mathbb{R}}\left|\mathbb{P}\big(q^*_m(\mathbf{X})\le t\big) - \mathbb{P}\big(q^*_m(\Xi)\le t\big)\right| \;\le\; Cmn^{-\frac{\nu-2}{2\nu m+2}} \end{aligned}$$

for i.i.d. variables $X_1,\dots,X_n \sim \mu^{(n)}_{\nu,\sigma_0}$ and all $n \ge N$.

Recall that for a generic function $f : \mathbb{R}^{nd} \to \mathbb{R}$, Theorem 4.2 establishes a universality result when f is well-approximated by its m-th order Taylor expansion. Theorem 4.7 shows that this Taylor approximation, which becomes more accurate as m grows, trades off against the Gaussian universality approximation,which deteriorates with m. In other words, when no additional information about f or the law of X is available, the error bound obtained from Theorem 4.1 is near-optimal, and becomes tighter as m grows.

The main technical difficulty of Theorem 4.7 is the lower bound, which cannot be established by the techniques on upper bounds from Chapter 2. We devote the rest of the section to a discussion on the proof techniques and the exact constructions of $\mu^{(n)}_{\nu,\sigma_0}$ and $p^*_m = p^*_{m(n)}$, and include the full proofs in Appendix C.3.

The key ingredient of the proof of Theorem 3 is a suitable sequence of heavy-tailed probability measures $\mu^{(n)}_{\nu,\sigma_0}$ and statistic p^*_m. Before stating the exact constructions, we first motivate them by discussing our proof technique.

Adapting an asymmetry argument from Senatov (1998). The overall idea is to construct a mixture of an average of a heavy-tailed random variable, which is poorly approximated via Gaussian universality, and a degree-m V-statistic of Gaussians, which has good approximation by the same V-statistic of an i.i.d. copy of Gaussians at small m but poor approximation at large m. This will be made precise in (4.7). The technical

steps to obtain a tight lower bound, however, are non-trivial. Our strategy is inspired by Example 9.1.3 of Senatov (1998): They construct a sequence of probability measures on $\mathbb{R}^d$ to demonstrate how a multivariate normal approximation bound on a chosen sequence of Euclidean balls may depend on eigenvalues of the covariance matrix. Two key mathematical ingredients of their work are

(i) a heavy-tailed univariate variable, whose i.i.d. average is poorly approximated by a Gaussian, and

(ii) an asymmetry argument by considering Euclidean balls not centred at the origin while studying averages of mean-zero random vectors.

To adapt the construction in Senatov (1998) for our problem, we make one key observation: The probability of a mean-zero average lying in a radius-r Euclidean ball centred at the origin is exactly the probability of a degree-2 V-statistic taking values in $[0, r^2)$. In other words, (i) gives the heavy-tailed variable we need, and (ii) is almost our V-statistic except that the Euclidean balls in Senatov (1998) are not centred at the origin. Unfortunately, while the asymmetry in (ii) is key to the proof by Senatov (1998), it breaks the connection to V-statistics, and a naive application of Senatov's results gives very loose bounds. Instead, we create a different asymmetry by asking $p^*_m(X)$ to be a mixture of an odd-degree V-statistic and an even-degree V-statistic, which allows Senatov's argument to be adapted. An additional distinction of our proof from Senatov (1998) is that, to accommodate a growing degree m, we need to derive finer moment controls.

We are ready to state the constructions. For $m = m(n)$ even, consider the polynomial

$$p^*_m(x_1, \ldots, x_n) \coloneqq \frac{1}{\sqrt{n}} \sum_{i=1}^n x_{i1} + \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n x_{ij2} \right)^m \quad \text{for } x_i = (x_{i1}, x_{i2}) \in \mathbb{R}^2 \, .$$

The distribution $\mu^{(n)}_{\nu,\sigma_0}$ is the law of a bivariate random vector with a heavy-tailed coordinate and a Gaussian coordinate. We first construct the heavy-tailed variable.

A heavy-tailed univariate distribution $\tilde{\mu}_{\nu,\sigma}$. For $\nu > 2$ and $\sigma \in (0, 1]$, we define

$$p \; = \sigma^{2\nu/(\nu-2)} \in (0, 1] \qquad \text{and} \qquad x_0 \; = \sigma/\sqrt{6p} \; = \; x \; = \; \frac{1}{\sqrt{6}} \sigma^{-\frac{2}{\nu-2}} \, ,$$

and let U_1 be a discrete random variable taking values in $\{-x_0, 0, 2x_0\}$ with

$$\mathbb{P}(U_1 = -x_0) \; = \; 2p \, , \qquad \mathbb{P}(U_1 = 0) \; = \; 1 - 3p \quad \text{and} \quad \mathbb{P}(U_1 = 2x_0) \; = \; p \, .$$

Let $Z_1 \sim \mathcal{N}(0, \sigma^2)$ and define a smoothed version of U as

$$V_1 \; \coloneqq \; 2^{-1/2} U_1 + 2^{-1/2} Z_1 \, .$$

Note that the only role Z_1 plays in the proof is for V_1 to have a continuous density function, which allows us to apply results relating distribution functions to characteristic functions. Write $\tilde{\mu}_{\nu,\sigma}$ as the law of V_1. When $\nu = 3$, this is the heavy-tailed distribution

constructed in Example 9.1.3 of Senatov (1998). Roughly speaking, the construction is such that if we set $\sigma = \sigma_n \to 0$ as $n \to \infty$, then $\mathrm{Var}[V_1] = \sigma_n^2 \to 0$, but the ν-th central moment of V_1 remains $\Theta(1)$.

Let $V_2, \ldots, V_n \overset{i.i.d.}{\sim} \tilde{\mu}_{\nu,\sigma}$ and draw an independent $Z_1' \sim \mathcal{N}(0, \sigma^2)$. We now list three results on $\tilde{\mu}_{\nu,\sigma}$, which extend the results by Senatov (1998) to a general ν and admit similar proofs. Since the original proofs are highly condensed, we include proofs for all lemmas with more detailed steps and intuitions in Appendix C.8.2. Lemma 4.8 below controls moments of V_1. In particular, $\mathbb{E}|V_1|^\nu = O(1)$ and the upper bound on $\mathbb{E}|V_1|^\omega$ diverges for $\omega > \nu$. This upper bound will allow us to handle moments of polynomials of V_i's of growing degrees.

Lemma 4.8. $\mathbb{E}\, V_1 = 0$ *and* $\mathrm{Var}\, V_1 = \sigma^2$. *Moreover, there exist absolute constants* $c_1, c_2 > 0$ *such that for all* $\omega \geq 1$, $\mathbb{E}|V_1|^\omega \leq c_1^\omega \sigma^{-\frac{2(\omega-\nu)}{\nu-2}} + c_2^\omega \sigma^\omega \omega^{\omega/2}$.

Lemma 4.9 below provides a finer control on the normal approximation error of an empirical average, by performing a higher-order Taylor expansion in the space of characteristic functions. This higher-order Taylor term is then inverted back to the space of distribution functions and captured by F_q below.

Lemma 4.9. *Write* $A := \frac{1}{2^{7/2}3^{3/2}\pi^{1/2}}$ *and* $F_q(x) := \frac{A}{n^{1/2}\sigma^{\nu/(\nu-2)}}\big(1 - \frac{x^2}{\sigma^2}\big)e^{-x^2/(2\sigma^2)}$. *Then*

$$\sup_{x\in\mathbb{R}} \left|\mathbb{P}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n V_i < x\Big) - \mathbb{P}(Z_1' < x) - F_q(x)\right| \;\leq\; \frac{2}{n\sigma^{2\nu/(\nu-2)}}\,.$$

Remark. F_q is related to the higher order edgeworth expansion of the distribution of $\frac{1}{\sqrt{n}}\sum_{i\leq n} V_i$, except that the result above involves an approximation of the cumulative density function whereas edgeworth expansion typically concerns the probability density function.

Lemma 4.10 below provides a finer upper bound on the normal approximation error of an average by exploiting the distribution of V_i's we constructed.

Lemma 4.10. *Suppose there exists some constant* $M \geq 10$ *such that* $\sigma^{\nu/(\nu-2)} \geq Mn^{-1/2}$ *and* $n \geq 6M^2$. *Then there is some constant* $C_M > 0$ *that depends only on* M *such that, for all* $x \in \mathbb{R}$, *we have*

$$\left|\mathbb{P}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n V_i < x\Big) - \mathbb{P}(Z_1' < x)\right| \;\leq\; C_M\left(\frac{\max\{1,\sigma^3\}}{n^{1/2}\sigma^{\nu/(\nu-2)}}e^{-\frac{x^2}{16\sigma^2}} + \frac{1}{n^{3/2}x^4\sigma^{(8-\nu)/(\nu-2)}}\right).$$

The bivariate distribution $\mu_{\nu,\sigma_0}^{(n)}$. Now for a fixed $\nu \in (2, 3]$ and some absolute constant $\sigma_0 > 0$, set a standard deviation parameter

$$\sigma_n \;:=\; \min\left\{\sigma_0\, n^{-\frac{\nu-2}{2\nu}},\, 1\right\} \;\in\; (0, 1]\,.$$

With a slight abuse of notation, we consider

$$V_1, \dots, V_n \overset{i.i.d.}{\sim} \tilde{\mu}_{\nu,\sigma_n} \qquad \text{and} \qquad Y_1, \dots, Y_n \overset{i.i.d.}{\sim} \mathcal{N}(0,1) \ ,$$

and define $\mu^{(n)}_{\nu,\sigma_0}$ as the law of (V_1, Y_1).

In summary, the statistic we consider is a mixture of a heavy-tailed average and an even-degree V-statistic of Gaussian random variables:

$$p^*_m(X) \ := \frac{1}{\sqrt{n}} \sum_{i=1}^n V_i + \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i \right)^m \ = \ v^*_1(V) + v^*_m(Y) \ , \tag{4.7}$$

where we denoted $V := (V_i)_{i \le n}$, $Y := (Y_i)_{i \le n}$ and the rescaled degree-m' V-statistic as

$$v^*_{m'}(x'_1, \dots, x'_n) \ := \ \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n x'_i \right)^{m'} \qquad \text{for } x'_1, \dots, x'_n \in \mathbb{R} \ .$$

4.5.1. Lower bound construction for degree-two U-statistics and V-statistics

Since p^*_m in (4.7) is a rather specific polynomial, a natural question is its applicability to more natural classes of statistics such as U-statistics and V-statistics. It turns out that in the case $m = 2$, we can adapt p^*_m and Theorem 4.7 to obtain similar lower bound results for degree-two U-statistics and degree-two V-statistics. The U-statistics result has been presented as Theorem 3.8 in Section 3.3 and the V-statistics result is included as Theorem B.1 in the appendix.

To state the construction, notice that in the case $m = 2$, p^*_m is a polynomial $p^*_2 : (\mathbb{R}^2)^n \to \mathbb{R}$ given by

$$p^*_2(w_1, \dots, w_n) \ := \frac{1}{\sqrt{n}} \sum_{i=1}^n w_{i1} + \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n w_{i2} \right)^2 \qquad \text{for } w_i = (w_{i1}, w_{i2}) \in \mathbb{R}^2 \ .$$

We also consider the collection of i.i.d. $\mathbb{R}^2$ random vectors $W := (W_i)_{i \le n}$ with $W_i \sim \mu^{(n)}_{\nu,\sigma_0}$. To adapt $p^*_2(W)$, we first note that $p^*_2(W)$ can be rewritten as a V-statistic:

$$p^*_2(w_1, \dots, w_n) \ = \frac{1}{n} \sum_{i,j=1}^n \left(\frac{w_{i1}}{2\sqrt{n}} + \frac{w_{j1}}{2\sqrt{n}} + w_{i2}\, w_{j2} \right) \ = \ n\, \tilde{v}_n(w_1, \dots, w_n) \ ,$$

where we have defined, for $w_1, \dots, w_n \in \mathbb{R}^2$ and $a_1, a_2, b_1, b_2 \in \mathbb{R}$,

$$\begin{aligned} \tilde{v}_n(w_1, \dots, w_n) \ &:= \frac{1}{n^2} \sum_{i,j=1}^n \tilde{k}_v(w_i, w_j) \ , \\ \tilde{k}_v((a_1, a_2), (b_1, b_2)) \ &:= \frac{a_1}{2\sqrt{n}} + \frac{b_1}{2\sqrt{n}} + a_2 b_2 \ . \end{aligned}$$

Moreover, since we have no restrictions on how $d(n)$ depends on n, there are non-unique choices of a function $\phi_{d(n)} : \mathbb{R}^{d(n)} \to \mathbb{R}$ and a probability measure $\mu_{d(n)}$ on $\mathbb{R}^{d(n)}$ such that

$$X_1 \ \sim \mu_{d(n)} \qquad \Leftrightarrow \qquad \phi_{d(n)}(X_1) \ \overset{d}{=} W_{1;\sigma_n} \ .$$

Our construction of the V-statistic is thus given by taking $X_i \overset{\text{i.i.d.}}{\sim} \mu_{d(n)}$ and $k_v(x_1, x_2) :=$

$\tilde{k}_v(\phi_{d(n)}(x_1), \phi_{d(n)}(x_2))$, which gives

$$v_n(X) \;=\; \frac{1}{n^2}\sum\nolimits_{1\le i,j\le n} k_v(X_i, X_j) \;\stackrel{d}{=}\; \tilde{v}_n(W) \;=\; \frac{1}{n}\, p_2^*(W)\;,$$

where $\stackrel{d}{=}$ denotes equality in distribution. This makes Theorem 4.7 immediately applicable, which yields Theorem B.1. For the U-statistics construction, we observe that

$$\begin{aligned}
&p_2^*(w_1, \ldots, w_n) \\
&= \frac{1}{\sqrt{n(n-1)}}\sum\nolimits_{i\neq j}^{n}\left(\frac{w_{i1}}{2\sqrt{n-1}} + \frac{w_{i2}}{2\sqrt{n-1}} + \frac{\sqrt{n-1}}{\sqrt{n}}\, w_{i2}w_{j2}\right) + \frac{1}{n}\sum\nolimits_{i=1}^{n} w_{i2}^2 \\
&= \sqrt{n(n-1)}\,\tilde{u}_n(w_1, \ldots, w_n) + R_n(w_1, \ldots, w_n)\;,
\end{aligned}$$

where we have defined, for $w_1, \ldots, w_n \in \mathbb{R}^2$ and $a_1, a_2, b_1, b_2 \in \mathbb{R}$,

$$\tilde{u}_n(w_1, \ldots, w_n) \;\coloneqq\; \frac{1}{n(n-1)}\sum\nolimits_{i\neq j}\tilde{k}_u(w_i, w_j)\;, \qquad R_n(w_1, \ldots, w_n) \;\coloneqq\; \frac{1}{n}\sum\nolimits_{i=1}^{n} w_{i2}^2\;,$$

$$\tilde{k}_u((a_1, a_2), (b_1, b_2)) \;\coloneqq\; \frac{a_1}{2\sqrt{n-1}} + \frac{b_1}{2\sqrt{n-1}} + \frac{\sqrt{n-1}}{\sqrt{n}}\, a_2 b_2\;.$$

Therefore for the U-statistic, we take $k_u(x_1, x_2) \coloneqq \tilde{k}_u(\phi_{d(n)}(x_1), \phi_{d(n)}(x_2))$, which gives

$$u_n(X) \;=\; \frac{1}{n(n-1)}\sum\nolimits_{1\le i\neq j\le n} k_u(X_i, X_j) \;\stackrel{d}{=}\; \tilde{u}_n(W) \;=\; p_2^*(W) - R_n(W)\;.$$

To obtain Theorem 3.8 from Theorem 4.7, the only technical hurdle is to show that $R_n(W)$ has a negligible effect other than centering the chi-squared distribution. This is achieved by applying the variance domination results in Section 4.3 and exploiting the anti-concentration of the chi-squared distribution.

Chapter 5

Degree-m polynomials of high-dimensional data

This chapter focuses on generalising existing approximation results as applications of our universality theorems in Chapter 4, and we assume the notation used in Chapter 4. For simplicity, we mostly consider i.i.d. data and symmetric functions. As our general results also hold in the non-i.i.d. and asymmetric case, the interesting phenomena we observe can be readily extended, and Section 5.4 provides one such example. Notably, since any degree-m, symmetric polynomial of n variables can be written as a weighted sum of U-statistics and V-statistics with degree $\leq m$, all our applications will be reduced to studying the asymptotic distributions of U-statistics and V-statistics. The proofs for all results in this section are included in Appendix C.

The rest of the chapter proves universality results for the following examples:

- Section 5.1 concerns a family of simple V-statistics. This example also illustrates how universality holds for a non-multilinear polynomial p_m;
- Section 5.2 concerns a high-dimensional delta method. The resulting limit distributions for functions of sample averages may be non-Gaussian and even non-consistent, depending on which polynomial component of the estimator dominates. See Proposition 5.5 for the result and Appendix C.1.1 for a simple example where such transition happens in high dimensions;
- Section 5.3 provides finite sample bounds that generalise a range of existing results on high-dimensional and infinite-order U-statistics (van Es and Helmers, 1988; Chen and Qin, 2010; Harchaoui et al., 2020; Wang et al., 2015; Yan and Zhang, 2022; Gao and Shao, 2023; Bhattacharya et al., 2022). The degree of the U-statistic may grow faster than $\log n$, provided the degree M of the approximating polynomial of Gaussians satisfies $M \log M = o(\log n)$;
- Section 5.4 provides results on fluctuations of subgraph densities at different orders, which extend those in Hladký et al. (2021); Bhattacharya et al. (2023); Kaur and Röllin (2021) by characterising a full range of vertex-level and edge-level fluctuations with finite-sample bounds.

5.1 Simple V-statistics

In Sections 4.1 and 4.2, we have motivated the approximation of $p_m(X) - \mathbb{E}[p_m(X)]$ by $q_m(\Xi)$, a polynomial of the augmented variables, by noting how q_m adapts well to the Lindeberg method. On the other hand, in the case of $p_m(X) = g(\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i)$, the continuous mapping theorem suggests an approximation by $p_m(Z)$, a polynomial of the original variables, at least when g and d are fixed in n. A natural question is whether the two approximations, i.e. the two different notions of Gaussian universality, coincide. Note that in the high-dimensional regime, it is non-trivial to show that one may replace X_i's by Z_i's in p_m: Gaunt (2020); Gaunt and Sutcliffe (2023) have used Stein's method and a Lipschitz argument to control the approximation error by $p_m(Z)$ in smooth function bounds, but those bounds are only well-controlled when dimension d grows sublinearly in n and are not well-adapted to the high-dimensional setting.

In this section, we confirm that the two notions of universality do coincide in the high-dimensional regime, under a mild moment condition and for simple V-statistics. We focus on the case where $X_1, \dots, X_n$ are i.i.d. zero-mean and consider, for $x_1, \dots, x_n \in \mathbb{R}^d$, the statistic

$$v_m(x_1, \dots, x_n) \;:=\; \frac{1}{n^m} \sum\nolimits_{i_1,\dots,i_m \in [n]} \langle S,\, x_{i_1} \otimes \dots \otimes x_{i_m} \rangle \;, \tag{5.1}$$

where $S \in \mathbb{R}^{d^m}$ is a deterministic, symmetric tensor. Note that this is exactly the dominating quantity under an m-th order delta method, and when S is not required to be symmetric, any $p_m(X)$ of the form (4.5) can be written as a weighted sum of such V-statistics. We also note that the centering of X_i's implies that $v_m(X)$ is a degenerate V-statistic, which simplifies our presentation; for distributional approximations of similar sums of uncentred variables, we refer interested readers to Temčinas et al. (2024) and the references therewithin.

The moment condition is cumbersome to state, but can be motivated as follows. We first notice that for the distributions of $p_m(Z)$ and $q_m(\Xi)$ to agree, we require the contribution of terms like $Z_i^{\otimes m}$ (found in $p_m(Z)$) and $\mathcal{N}(\mathbb{E}[X_i^{\otimes m}], \mathrm{Var}[X_i^{\otimes m}])$ (found in $q_m(\Xi)$) to be negligible. To make the condition well-adapted to a setting where d and the law of X_i's may vary in n, we need careful controls for all possible m-fold products of Z_i's and of X_i's with at least one repeated element. We coin this moment condition δ-regularity:

Definition 5.1 (δ-regularity)**.** v_m is δ-regular with respect to X if, for some $\delta \in [0, 1)$, there exists an absolute constant $C > 0$ such that for all $j \in [m-1]$,

$$\max_{q_1+\dots+q_j=m\,,\, q_l \in \mathbb{N}} \max\left\{ \mathrm{Var}\left[\left\langle S\,,\, \bigotimes_{l=1}^{j} X_l^{\otimes q_l} \right\rangle \right],\, \mathrm{Var}\left[\left\langle S\,,\, \bigotimes_{l=1}^{j} Z_l^{\otimes q_l} \right\rangle \right] \right\} \tag{5.2}$$
$$\leq C n^{m+\delta} m^{-j}\, \mathrm{Var}[v_m(X)] \;,$$

$$\max_{q_1+\ldots+q_j=m\,,\,q_l\in\mathbb{N}} \max\left\{\left(\mathbb{E}\left[\left\langle S\,,\,\bigotimes_{l=1}^{j} X_l^{\otimes q_l}\right\rangle\right]\right)^2,\left(\mathbb{E}\left[\left\langle S\,,\,\bigotimes_{l=1}^{j} Z_l^{\otimes q_l}\right\rangle\right]\right)^2\right\} \leq C\,n^{m+\delta}\,\mathrm{Var}[v_m(X)]\,. \tag{5.3}$$

Remark 5.1. The n^m factor on the RHS balances the variance of the un-normalised V-statistic $v_m(X)$. The m^{-j} factor accounts for both the $\binom{m-1}{j-1} = O(m^{j-1})$ multiplicity of the terms in $\mathrm{Var}[v_m(X)]$, and the fact that the moment of a univariate standard Gaussian η grows as $\mathbb{E}[\eta^m] = O(m^{m/2})$. In Lemma C.8 in the appendix, we provide a verification of δ-regularity for a degree-m V-statistic of univariate Gaussians as part of the proof of Theorem 4.7. Sufficient conditions for verifying δ-regularity are deferred to Remark 5.4. There, we show that when m and d are fixed, up to a spectral condition, δ-regularity always holds. It suggests that the requirement of δ-regularity is specific to our high-dimensional setup, where both m and d may grow in n.

We are ready to state the main result of this subsection, which confirms that under δ-regularity for the V-statistic v_m in (5.1), the approximations by $q_m(\Xi)$ and $p_m(Z) - \mathbb{E}[p_m(X)]$ do agree. It also immediately implies that the universality result of Theorem 4.1 holds also for the non-multilinear polynomial $p_m \equiv v_m$. Here and below, we let q_m^v be the multilinear representation of v_m defined in (4.2).

Proposition 5.2. *Fix $\nu \in (2,3]$ and let $(X_i)_{i\leq n}$ be i.i.d. zero-mean. If v_m is δ-regular with respect to X for some $\delta \in [0,1)$, there exist some absolute constants $C, C' > 0$ such that*

$$\begin{aligned}
&\sup_{t\in\mathbb{R}} \left|\mathbb{P}\big(\sigma^{-1}(q_m^v(\Xi) + \mathbb{E}[v_m(X)]) \leq t\big) - \mathbb{P}\big(\sigma^{-1}v_m(Z) \leq t\big)\right| \;\leq\; Cmn^{-\frac{1-\delta}{2m+1}}\,,\\
&\sup_{t\in\mathbb{R}} \left|\mathbb{P}(\sigma^{-1}v_m(X) \leq t) - \mathbb{P}(\sigma^{-1}v_m(Z) \leq t)\right| \;\leq\; Cmn^{-\frac{1-\delta}{2m+1}} + C'm\Delta_\delta\,,\\
&\sup_{t\in\mathbb{R}} \left|\mathbb{P}(\sigma^{-1}(v_m(X) - \mathbb{E}[v_m(X)]) \leq t) - \mathbb{P}(\sigma^{-1}(v_m(Z) - \mathbb{E}[v_m(Z)]) \leq t)\right|\\
&\qquad\qquad \leq\; Cmn^{-\frac{1-\delta}{2m+1}} + C'm\Delta_\delta\,,
\end{aligned}$$

where the error term is defined as

$$\Delta_\delta \;:=\; \left(\frac{\sum_{i=1}^n M_{\nu;i}^\nu}{\sigma^\nu}\right)^{\frac{1}{\nu m+1}}\,.$$

Remark 5.2. (i) Proposition 5.2 is proved by an adaptation of variance domination in Theorem 4.2 together with the variance ratio bounds in Lemma 5.3. (ii) While $\mathbb{E}[v_m(X)]$ does not necessarily equal $\mathbb{E}[v_m(Z)]$, the third bound of Proposition 5.2 implies that the difference is asymptotically negligible. (iii) An explicit bound on Δ_δ is given in Lemma 5.4.

The proof of Proposition 5.2 is included in Appendix C.6, and an illustration of Gaussian universality for a toy V-statistic in Appendix C.1.1. The proof makes use of the next

two lemmas, which are also useful for subsequent applications. Their proofs are included in Appendix C.6. The first lemma shows that δ-regularity can be used to control two variance ratios, which are similar to the moment ratio obtained from variance domination (Theorem 4.2):

Lemma 5.3. *Suppose $(X_i)_{i\le n}$ are i.i.d. zero-mean. Assume that v_m is δ-regular with respect to X for some $\delta \in [0,1)$. Also assume a coupling between ξ_i and Z_i such that $\xi_{i1} = Z_i$ almost surely. Then there exists some absolute constants $C, C_* > 0$ such that*

$$\frac{\|q_m^v(\Xi) + \mathbb{E}[v_m(X)] - v_m(Z)\|_{L_2}^2}{\mathrm{Var}[q_m^v(\Xi)]} \;\le\; \frac{C^m}{n^{1-\delta}}$$

and $\big(1 - (C_*)^m n^{-(1-\delta)/2}\big)^2 \le \frac{\mathrm{Var}[v_m(Z)]}{\mathrm{Var}[v_m(X)]} \le \big(1 + (C_*)^m n^{-(1-\delta)/2}\big)^2$.

The next lemma applies Theorem 4.1 to show Gaussian universality with respect to the augmented variables $\mathbf{X}$, and use the simple structure of v_m to bound the moment ratio explicitly. The error bounds are given in terms of both $M_{\nu;i}$ defined in Theorem 4.1 for q_m^v and

$$\alpha_\nu(S,k) \;:=\; \Big\| \sum\nolimits_{\substack{p_1+\ldots+p_k=m \\ p_1,\ldots,p_k\ge 1}} \Big\langle S\,,\, \overline{X_1^{\otimes p_1}} \otimes \ldots \otimes \overline{X_k^{\otimes p_k}} \Big\rangle \Big\|_{L_\nu} \qquad \text{for } \nu \in [2,3] \text{ and } k \in [m]\,.$$

We also provide upper and lower bounds on $\mathrm{Var}[v_m(X)]$ in terms of $\alpha_2(S,k)$.

Lemma 5.4. *Fix $\nu \in (2,3]$ and let $(X_i)_{i\le n}$ be i.i.d. zero-mean. Then there exists some absolute constant $C > 0$ such that, for every $n, m, d \in \mathbb{N}$, $t \in \mathbb{R}$ and $\sigma = \mathrm{Var}[v_m(X)] > 0$,*

$$\sup_{t\in\mathbb{R}} \big|\mathbb{P}(\sigma^{-1} q_m^v(\mathbf{X}) \le t) - \mathbb{P}(\sigma^{-1} q_m^v(\Xi) \le t)\big| \;\le\; Cm\Big(\frac{\sum_{i=1}^n M_{\nu;i}^\nu}{\sigma^\nu}\Big)^{\frac{1}{\nu m+1}} \;=\; Cm\Delta_\delta\,.$$

Meanwhile, if $n \ge 2m^2$, there exist absolute constants $C', C_1, C_2 > 0$ such that

$$\Delta_\delta \;\le\; C'\, n^{-\frac{\nu-2}{2\nu m+2}}\, \beta_{m,\nu}^{\frac{\nu}{2\nu m+2}}\,, \qquad \beta_{m,\nu} \;:=\; \frac{\sum_{k=1}^m \binom{n}{k} (\alpha_\nu(S,k))^2}{\sum_{k=1}^m \binom{n}{k} (\alpha_2(S,k))^2}\,,$$

$$\textit{and} \quad \frac{(C_1)^m}{n^{2m}} \sum\nolimits_{k=1}^m \binom{n}{k} (\alpha_2(S,k))^2 \;\le\; \mathrm{Var}[v_m(X)] \;\le\; \frac{(C_2)^m}{n^{2m}} \sum\nolimits_{k=1}^m \binom{n}{k} (\alpha_2(S,k))^2\,.$$

Remark 5.3. The moment bound in Lemma 5.4 involves a ratio of the ν-th moment versus the second moment of the same quantity. In fact, the ratio is a strict generalisation of the moment ratio from the classical Berry-Esseen bound for sample averages ($m = 1$), and similar to the Berry-Esseen ratio, we expect it to be $O(1)$ for distributions that are not too heavy-tailed. A crude upper bound on this ratio is deferred to Remark 5.5.

Remark 5.4 (Sufficient condition for verifying δ-regularity)**.** A sufficient condition for δ-regularity can be obtained from a moment condition and a spectral condition. Recall that $\mathrm{vec}(T)$ is the vectorisation of a tensor T (Section 4.1), and let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the maximum and minimum eigenvalues of a real symmetric matrix A. Suppose

there exist $\omega_{\max}^{(1)}, \omega_{\max}^{(2)} > 0$ such that for all $j \leq m-1$ and any $q_1, \ldots, q_j \in \mathbb{N}$ that sum up to m,

$$\lambda_{\max}\Big(\mathrm{Var}\Big[\mathrm{vec}\Big(\otimes_{l=1}^{j} X_l^{\otimes q_l}\Big)\Big]\Big) \leq \omega_{\max}^{(1)} ,$$
$$\lambda_{\max}\Big(\mathbb{E}\Big[\mathrm{vec}\Big(\otimes_{l=1}^{j} X_l^{\otimes q_l}\Big)\mathrm{vec}\Big(\otimes_{l=1}^{j} X_l^{\otimes q_l}\Big)^\top\Big]\Big) \leq \omega_{\max}^{(2)} .$$

and suppose $\lambda_{\min}(\Sigma_2) > 0$, where

$$\Sigma_2 := \mathrm{Var}[\mathrm{vec}(Q_X)] , \qquad Q_X := \sum_{\substack{p_1+\ldots+p_k=m \\ p_1,\ldots,p_k \geq 1}} \overline{X_1^{\otimes p_1}} \otimes \ldots \otimes \overline{X_k^{\otimes p_k}} .$$

WLOG take $\|\mathrm{vec}(S)\| = 1$. Then we have

$$(5.2) \leq \omega_{\max}^{(1)} , \qquad (5.3) \leq \omega_{\max}^{(2)} , \qquad \mathrm{Var}[v_m(X)] \geq C^m n^{-m} \lambda_{\min}(\Sigma_2)$$

for some absolute constant $C > 0$ provided that $n \geq 2m^2$; the third bound is obtained from Lemma 5.4. Since $m = o(\log(n))$, $C^m = O(n^\epsilon)$ for any $\epsilon > 0$, so to show δ-regularity, it suffices to check that for some constant $C > 0$, some $\delta \in [0, 1)$ and some $\epsilon > 0$,

$$\max\{\omega_{\max}^{(1)}, \omega_{\max}^{(2)}\} \leq C\, n^{\delta-\epsilon}\, \lambda_{\min}(\Sigma_2) .$$

Notice that $\lambda_{\min}(\Sigma_2)$ is the minimum eigenvalue of the covariance matrix of X_1 when $m = 1$. Moreover, when m and d are fixed, up to a positive minimum eigenvalue condition, δ-regularity always holds. We emphasise that the above only gives crude bounds on the quantities in Definition 5.1. Instead of taking the minimum eigenvalue of Σ_2, it can often be easier to compute $\mathrm{Var}[v_m(X)]$ directly: One approach is to decompose a V-statistic as a sum of U-statistics (see Lemma C.12 in the appendix) followed by using known moment formulas for U-statistics (see Appendix C.5). Such an approach is used in Lemma C.8 in the appendix as part of the proof of Theorem 4.7.

Remark 5.5 (Upper bound on the moment ratio in Lemma 5.4)**.** We again use a spectral condition. Consider Σ_2 and Q_X defined in Remark 5.4, and the analogue of Σ_2 but with a fourth moment:

$$\Sigma_4 := \mathbb{E}\Big[\,\mathrm{vec}\big(Q_X^{\otimes 2}\big)\,\mathrm{vec}\big(Q_X^{\otimes 2}\big)^\top\,\Big] .$$

Define $\lambda_{\max}$ and $\lambda_{\min}$ as in Remark 5.4 and suppose $\lambda_{\min}(\Sigma_2) > 0$. Then by noting $\nu < 4$, an upper bound on the moment ratio in Lemma 5.4 is given as

$$\beta_{m,\nu} \leq \frac{\lambda_{\max}(\Sigma_4)^{1/2}}{\lambda_{\min}(\Sigma_2)} .$$

5.2 A high-dimensional delta method

We consider the classical delta method (Rao et al., 1973; Bishop et al., 1974) and show how a high-dimensional version may be derived from our universality result. Assume that $(X_i)_{i\le n}$ are i.i.d. and, for some smooth function $g : \mathbb{R}^d \to \mathbb{R}$, define

$$\hat{g}(X) \;\coloneqq\; g\Big(\frac{1}{n}\sum_{i=1}^n X_i\Big) \;=\; g\Big(\frac{1}{n}\sum_{i=1}^n \bar{X}_i + \mathbb{E}[X_1]\Big)\,, \qquad \text{where } \bar{X}_i = X_i - \mathbb{E}[X_1]\,.$$

$\hat{g}(X)$ is called a plug-in estimator of $g(\mathbb{E}X_1)$. A guiding example is $g_{\text{toy}}(x) \coloneqq x^\top x$, which yields a simple V-statistic

$$\hat{g}_{\text{toy}}(X) \;\coloneqq\; g_{\text{toy}}\Big(\frac{1}{n}\sum_{i=1}^n X_i\Big) \;=\; \frac{1}{n^2}\sum_{i,j=1}^n X_i^\top X_j\,.$$

To see how the distribution of the plug-in estimator behaves through variance domination, consider a Taylor expansion

$$\hat{g}_{\text{toy}}(X) - g_{\text{toy}}(\mathbb{E}[X_1]) \;=\; \frac{2}{n}\sum_{i=1}^n \bar{X}_i^\top \mathbb{E}[X_1] \;+\; \frac{1}{n^2}\sum_{i,j=1}^n \bar{X}_i^\top \bar{X}_j\,.$$

The variances of the first and second-order expansions are respectively on the order

$$\Theta\Big(\frac{\mathbb{E}[X_1]^\top \text{Var}[X_1]\mathbb{E}[X_1]}{n}\Big) \qquad \text{and} \qquad \Theta\Big(\frac{\text{Var}[\bar{X}_1^\top \bar{X}_2]}{n^2} + \frac{\text{Var}[\bar{X}_1^\top \bar{X}_1]}{n^3}\Big)\,. \tag{5.4}$$

Classically when d is fixed, normality of the plug-in estimator follows from the first-order delta method whenever $\partial g(\mathbb{E}[X_1]) \neq 0$. For $\hat{g}_{\text{toy}}(X)$, this condition is equivalent to $\mathbb{E}[X_1] \neq 0$, and exactly corresponds to when the first-order Taylor expansion dominates in variance. In the high-dimensional regime, however, the moment terms above can vary with n through their dependence on d: Take $X_i \sim \mathcal{N}(\mu, I_d)$ with $\|\mu\|_2 = 1$, in which case (5.4) becomes

$$\Theta\Big(\frac{1}{n}\Big) \qquad \text{and} \qquad \Theta\Big(\frac{d}{n^2} + \frac{d}{n^3}\Big)\,.$$

Even though $\mu = \mathbb{E}[X_1]$ and therefore $\partial g(\mathbb{E}[X_1])$ is non-zero, the first-order term is still dominated in variance by the second-order term when $d = \omega(\sqrt{n})$: The first-order delta method fails to hold, and the plug-in estimator is no longer asymptotically normal. We now formalise this in a result concerning an m-th order delta method.

Derivatives. A delta method relies on the derivatives of g, and some shorthands are introduced next. For $x, x_1, \ldots, x_m \in \mathbb{R}^d$ and $m \in \mathbb{N} \cup \{0\}$, we define

$$\begin{aligned} g_m^{(x)}(x_1, \ldots, x_m) \;&\coloneqq\; \frac{1}{m!}\big\langle \partial_m\, g(x)\,,\, (x_1 - \mathbb{E}X_1) \otimes \ldots \otimes (x_m - \mathbb{E}X_1)\big\rangle \\ \hat{g}_m^{(\mathbb{E}X_1)}(x_1, \ldots, x_n) \;&\coloneqq\; n^{-m} \sum_{i_1,\ldots,i_m \in [n]} g_m^{(\mathbb{E}X_1)}\big(x_{i_1}, \ldots, x_{i_m}\big)\,. \end{aligned}$$

which is a degree-m polynomial function. If g is $(m+1)$-times continuously differentiable everywhere, we can perform an m-th order Taylor expansion with the integral remainder. Let $S_n \coloneqq n^{-1}\sum_{i\le n} \bar{X}_i$ and draw $\Theta \sim \text{Uniform}[0,1]$ independently of X.

Then almost surely

$$\hat{g}(X) \;=\; g(\mathbb{E}X_1) + \sum_{j=1}^{m} \hat{g}_j^{(\mathbb{E}X_1)}(X) \;+\; (m+1)\mathbb{E}\big[(1-\Theta)^{m+1}\, \hat{g}_{m+1}^{(\mathbb{E}X_1+\Theta S_n)}(X)\,\big|\, X\big]\;.$$

Note that when $\hat{g}_{m+1}^{(\mathbb{E}X_1+\Theta S_n)}$ is independent of Θ, the $(m+1)$ factor is cancelled out by $\mathbb{E}[(1-\Theta)^{m+1}] = (m+1)^{-1}$. The penultimate approximation will be identified by comparing which term of the Taylor expansion dominates in variance.

For simplicity of notation, we will assume that $\hat{g}_m^{(\mathbb{E}X_1)}$ is δ-regular with respect to $(\bar{X}_i)_{i\le n}$ (see Definition 5.1 and Remark 5.4 for a sufficient condition), so that we can apply universality directly to X and not just the augmented variables. In the case of g_{toy} with $\bar{X}_i \sim \mathcal{N}(0, I_d)$, this corresponds to requiring the variance term involving $\bar{X}_1^\top \bar{X}_1$ to be negligible, which always holds in view of (5.4).

Moments. The result involves several moment terms. For each $m \in \mathbb{N}\cup\{0\}$, define

$$\mu_m \;:=\; \mathbb{E}\,\hat{g}_m^{(\mathbb{E}X_1)}(X)\;.$$

Note that $\mu_0 = g(\mathbb{E}X_1)$, which is the target of estimation under the plug-in principle. We also define, for $\nu \in (2,3]$ and $k \in [m]$, the moment terms

$$\epsilon_m \;:=\; \frac{\mathrm{Var}\Big[\sum_{l=1}^{m-1} \hat{g}_l^{(\mathbb{E}X_1)}(X)\Big]}{\mathrm{Var}\Big[\hat{g}_m^{(\mathbb{E}X_1)}(X)\Big]} + \frac{(m+1)^2 \Big\|\mathbb{E}\Big[(1-\Theta)^{m+1}\, \hat{g}_{m+1}^{\left(\mathbb{E}X_1+\Theta n^{-1}\sum_{i\le n}\bar{X}_i\right)}(X)\Big|X\Big]\Big\|_{L_2}^2}{\mathrm{Var}\Big[\hat{g}_m^{(\mathbb{E}X_1)}(X)\Big]}\,,$$

$$\alpha_{m,\nu}(k) := \Big\| \sum_{\substack{p_1+\ldots+p_k=m\\ p_1,\ldots,p_k\ge 1}} \Big\langle \frac{\partial_m g(\mathbb{E}X_1)}{m!}\,,\, \overline{(X_1-\mathbb{E}X_1)^{\otimes p_1}} \otimes \ldots \otimes \overline{(X_k-\mathbb{E}X_1)^{\otimes p_k}} \Big\rangle \Big\|_{L_\nu}\,,$$

$$\text{and } \beta_{m,\nu} \;:=\; \frac{\sum_{k=1}^m \binom{n}{k}(\alpha_{m,\nu}(k))^2}{\sum_{k=1}^m \binom{n}{k}(\alpha_{m,2}(k))^2}\;.$$

ϵ_m is a variance ratio that is small when the m-th order term dominates other terms of Taylor expansion in variance (Theorem 4.2; see Remark 5.7 for a further upper bound). $\beta_{m,\nu}$ is a V-statistic analogue of the classical Berry-Esseen ratio from Lemma 5.4 (see Remark 5.5 for a further upper bound). The next result gives a bound for approximating $\hat{g}(X)$ by $\hat{g}_m^{(\mathbb{E}X_1)}(Z)$.

Proposition 5.5. *Assume that g is $(m+1)$-times continuously differentiable and that $\hat{g}_m^{(\mathbb{E}X_1)}$ is δ-regular (Definition 5.1) with respect to $(X_i - \mathbb{E}X_1)_{i\le n}$ for some $\delta \in (0,1)$. Then, there is some absolute constant $C>0$ such that for all $n, m, d \in \mathbb{N}$ with $n \ge 2m^2$ and any $\nu \in (2,3]$, we have*

$$\sup_{t\in\mathbb{R}} \Big|\mathbb{P}\big(\hat{g}(X) - \sum_{l=0}^m \mu_l \le t\big) - \mathbb{P}\big(\,\hat{g}_m^{(\mathbb{E}X_1)}(Z) - \mathbb{E}\big[\hat{g}_m^{(\mathbb{E}X_1)}(Z)\big] \le t\,\big)\Big| \;\le\; \Delta_{\hat{g},m}\,,$$
$$\text{where}\quad \Delta_{\hat{g},m} \;:=\; Cm\Big(\epsilon_m^{\frac{1}{2m+1}} + n^{-\frac{1-\delta}{2m+1}} + n^{-\frac{\nu-2}{2\nu m+2}}\,\beta_{m,\nu}^{\frac{\nu}{2\nu m+2}}\Big)\;.$$

Write Φ as the c.d.f. of $\mathcal{N}(0,1)$ and suppose $\mathrm{Var}\big[\hat{g}_m^{(\mathbb{E}X_1)}(X)\big], \mathrm{Var}\big[\hat{g}_m^{(\mathbb{E}X_1)}(Z)\big] > 0$. *Then*

there are some absolute constants $C', C'' > 0$ *such that*

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\Big(\mathrm{Var}\big[\hat{g}_m^{(\mathbb{E}X_1)}(X)\big]^{-1/2} \big(\hat{g}(X) - \textstyle\sum_{l=0}^{m} \mu_l\big) \le t \Big) - \Phi(t) \right| \\ \le \Delta_{\hat{g},m} + C' \frac{(C'')^m\, n^{-(1-\delta)/2}}{2 - (C'')^m\, n^{-(1-\delta)/2}} + \left(\frac{4m-4}{3m} \left| \mathrm{Kurt}\big[\hat{g}_m^{(\mathbb{E}X_1)}(Z)\big] \right| \right)^{1/2}.$$

Remark 5.6. Proposition 5.5 considers only the case when one of the derivatives dominates, but can be readily extended to the case when several terms dominate at the same time: This is done by applying variance domination (Theorem 4.2) to the appropriate dominating quantity.

Remark 5.7 (Upper bound on ϵ_m)**.** Using the triangle inequality and the Jensen's inequality, we can obtain a further upper bound on the first term of ϵ_m as

$$(m-1) \sum_{l=1}^{m-1} \frac{\mathrm{Var}\left[\hat{g}_l^{(\mathbb{E}X_1)}(X)\right]}{\mathrm{Var}\left[\hat{g}_m^{(\mathbb{E}X_1)}(X)\right]} \,.$$

Each summand involves variances of V-statistics, whose upper and lower bounds are given in Lemma 5.4. As discussed in Remarks 5.4 and 5.5, these variance ratios can either be controlled by a further spectral bound or an explicit computation. An analogous bound applies to the second term, provided that the $(m+1)$-th derivative satisfies some stability condition: One example of such a condition is the existence of some $\nu > 2$ such that

$$\left\| \hat{g}_{m+1}^{(\mathbb{E}X_1+\Theta S_n)}(X) - \hat{g}_{m+1}^{(\mathbb{E}X_1)}(X) \right\|_{L_\nu} \;=\; o(1) \,.$$

When g is infinitely differentiable, Proposition 5.5 implies that the asymptotic distribution of $\hat{g}(X)$ is determined by its m-th order Taylor expansion for the smallest m such that $\epsilon_m \to 0$, i.e. the m-th order term that dominates in variance. As such, Proposition 5.5 strictly generalises the m-th order delta method by replacing the requirement that all lower derivatives are zero with the condition that $\epsilon_m = o(1)$, whereas the finite-sample bound introduces a second condition $\beta_{m,\nu} = o(n^{(\nu-2)/2\nu})$ analogous to a Berry-Esseen moment ratio. In particular, under these two conditions, the bound does not have any additional dependence on d. In the high-dimensional regime, Proposition 5.5 generalises the observations in g_{toy} to show two aspects of delta method that depart from classical behaviours:

- *Non-Gaussianity despite a non-zero first derivative.* The approximation by a degree-m polynomial of Gaussians can dominate the first-order Gaussian term, when the first derivative is negligible compared to some m-th derivative. If the excess kurtosis of this degree-m Gaussian polynomial does not vanish, Proposition 4.6 implies that $\hat{g}(X)$ is not asymptotically normal under a further uniform integrability condition. In particular, this may hold despite a non-zero first derivative term.

- *Non-consistency (under the plug-in principle).* By applying a further Markov's inequality on $\mathbb{P}\big(\hat{g}_m^{(\mathbb{E}X_1)}(Z) - \mathbb{E}\big[\hat{g}_m^{(\mathbb{E}X_1)}(Z)\big] \le t\big)$, we see that if $\Delta_{\hat{g},m} = o(1)$ and $\mathrm{Var}\big[\hat{g}_m^{(\mathbb{E}X_1)}(Z)\big] = o(1)$, then

$$\hat{g}(X) - \textstyle\sum_{l=0}^{m} \mu_l \xrightarrow{\mathbb{P}} 0 \,.$$

  Recall that under the plug-in principle, $\hat{g}(X)$ is intended as an estimator for $g(\mathbb{E}X_1) = \mu_0$. Since, by linearity,

$$\mu_1 \;=\; \mathbb{E}\big[\langle \partial g(\mathbb{E}X_1), X_1 - \mathbb{E}X_1\rangle\big] \;=\; 0 \,,$$

  $\hat{g}(X)$ is indeed consistent when Proposition 5.5 holds with $m = 1$ (e.g. in the classical case with non-zero first derivative). When $m > 1$, however, $\hat{g}(X)$ may be non-consistent, and the bias is exactly given by the limit of $\sum_{l=2}^{m} \mu_l$.

Remark 5.8 (Gaussianity under a zero first derivative)**.** Proposition 4.6 also implies that where m and d are allowed to grow, a degree-m polynomial of d-dimensional Gaussians may itself be asymptotically Gaussian: This can be seen from $\hat{g}_{\mathrm{toy}}(X)$ with $X_i \sim \mathcal{N}(\mu, I_d)$, whose second-order term satisfies

$$\frac{1}{n^2}\sum_{i,j=1}^{n} \bar{X}_i^\top \bar{X}_j \;\overset{d}{=}\; \frac{d}{n} \times \frac{1}{d}\sum_{j=1}^{d} \eta_j^2 \qquad \text{with } \eta_j \overset{i.i.d.}{\sim} \mathcal{N}(0,1) \,.$$

This is an average of i.i.d. variables, and becomes asymptotically normal when d and n both grow. A consequence is that in the high-dimensional regime, Gaussianity may also occur even when the first derivative is zero. A phase transtion from the Gaussian limit to the chi-squared limit can happen, if the second-order term dominates as d becomes large.

Remark 5.9. Theorem 4.1 and Theorem 4.2 in fact apply to a more general version of delta method on $g_n(T_n)$, where we consider a sequence of random variables $\{T_n\}_{n=1}^{\infty}$, with each T_n determined by $X_1, \ldots, X_n$, and a sequence of well-behaved functions $\{g_n\}_{n=1}^{\infty}$. The only requirement is a suitable Taylor approximation of the estimator, which yields the polynomial structure. Degree-m U-statistics, discussed next, give such an example.

5.3 Effect of large dimensions on degree-m U-statistics

Let $Y_1, \ldots, Y_n$ be i.i.d. random variables taking values in some general (not necessarily Euclidean), possibly n-dependent measurable space $\mathcal{E} \equiv \mathcal{E}(n)$. Given a symmetric function $u : \mathcal{E}^m \to \mathbb{R}$, consider the degree-$m$ U-statistic given by

$$u_m(Y) \;\coloneqq\; \frac{1}{n(n-1)\ldots(n-m+1)} \sum_{i_1,\ldots,i_m \in [n] \text{ distinct}} u(Y_{i_1}, \ldots, Y_{i_m}) \,. \tag{5.5}$$

Classically, under a technique called the Hájek projection, the asymptotic distribution of $u_m(Y)$ can be approximated by that of a degree-M polynomial of Gaussians, where

$$M \coloneqq \min\{j \in [m] : \sigma_j > 0\}, \qquad \text{and} \qquad \sigma_j^2 \coloneqq \operatorname{Var} \mathbb{E}[u(Y_1, \dots, Y_m) \,|\, Y_1, \dots, Y_j] .$$

See Filippova (1962), Rubin and Vitale (1980) and Chapter 5.4 of Serfling (1980) for references on the classical theory for U-statistics. In this section, we show how the high-dimensional asymptotics of $u_m(Y)$, where $\mathcal{E} = \mathbb{R}^d$ with m and d allowed to be large relative to n, can again be obtained with our general results.

We will invoke a functional decomposition assumption analogous to Assumption 3.2, which was used to obtain finite-sample bounds for degree-two U-statistics. For each $K \in \mathbb{N}$, we consider approximating $u(y_1, \dots, y_m)$ by a polynomial of m vectors in $\mathbb{R}^K$,

$$\sum_{k_1,\dots,k_m=1}^{K} \lambda^{(K)}_{k_1 \dots k_m} \phi^{(K)}_{k_1}(y_1) \times \dots \times \phi^{(K)}_{k_m}(y_m) ,$$

where for each K, the coordinates of the m vectors are defined via a triangular array of $\mathcal{E} \to \mathbb{R}$ functions $\{\phi^{(K)}_k\}_{k \le K, K \in \mathbb{N}}$ evaluated at $y_1, \dots, y_m$, and the polynomial weights are given by a triangular array of real values $\{\lambda^{(K)}_{k_1 \dots k_m}\}_{k_1,\dots,k_m \le K, K \in \mathbb{N}}$. We also denote the L_ν approximation error as

$$\varepsilon_{K;\nu} \coloneqq \Big\| \sum_{k_1,\dots,k_m=1}^{K} \lambda^{(K)}_{k_1 \dots k_m} \phi^{(K)}_{k_1}(Y_1) \times \dots \times \phi^{(K)}_{k_m}(Y_m) - u(Y_1, \dots, Y_m) \Big\|_{L_\nu} .$$

Assumption 5.1. There exists some $\nu \in (2, 3]$ such that, for every fixed n, m, d and $\mathcal{E}$, as $K \to \infty$, the error $\varepsilon_{K;\nu} \to 0$ for some choice of $\{\phi^{(K)}_k\}_{k \le K, K \in \mathbb{N}}$, $\{\lambda^{(K)}_{k_1 \dots k_m}\}_{k_1,\dots,k_m \le K, K \in \mathbb{N}}$.

Remark 5.10. (i) As discussed in Assumption 3.2, Assumption 5.1 is very mild. In the appendix, we show how it holds with $\nu = 2$ under mild assumptions on the L_2 space equipped with the m-fold product measure of Y_1 (Lemma C.1), and how it holds for $\mathcal{E} = \mathbb{R}^d$ and any u well-approximated by a Taylor expansion (Lemma C.2). (ii) Unlike asymptotic U-statistics results that use the L_2 decomposition from Hilbert-Schmidt operator theory (see e.g. Chapter 5.4 of Serfling (1980)), Assumption 5.1 does not require orthogonality or boundedness conditions on $\phi^{(K)}_k$ or $\lambda^{(K)}_{k_1 \dots k_m}$ and also allow them to vary with K or n. In particular, they are generally not unique and can be chosen at convenience for verification of the assumption, e.g. by a suitable Taylor expansion.

Penultimate approximation. To identify the dominating component, we first recall that by Hoeffding's decomposition theorem (see e.g. Theorem 1.2.1 of Denker (1985)),

$$u_m(y_1, \dots, y_n) = \mathbb{E}[u(Y_1, \dots, Y_m)] + \sum_{j=1}^{m} \binom{m}{j} U^{\mathrm{H}}_j(y_1, \dots, y_n) , \tag{5.6}$$

where each Hoeffding's decomposition $U^{\mathrm{H}}_j(Y)$ is a degenerate degree-j U-statistic defined by

$$U^{\mathrm{H}}_j(y_1, \dots, y_n) \coloneqq \frac{1}{n(n-1)\dots(n-j+1)} \sum_{i_1,\dots,i_j \in [n] \text{ distinct}} u^{\mathrm{H}}_j(y_{i_1}, \dots, y_{i_j}) ,$$

$$u_j^{\mathrm{H}}(y_1,\dots,y_j) \;:=\; \sum_{r=0}^{j}(-1)^{j-r}\sum_{1\le l_1<\dots<l_r\le j}\mathbb{E}\big[u(y_{l_1},\dots,y_{l_r},Y_1,\dots,Y_{m-r})\big]\;.$$

Under Assumption 5.1, each Hoeffding's decomposition can be approximated by a polynomial of $\mathbb{R}^K$ vectors, to which our universality results can be applied. Let $\Xi^{(K)} := \{\xi_1^{(K)},\dots,\xi_n^{(K)}\}$ be a collection of i.i.d. zero-mean Gaussian random vectors in $\mathbb{R}^{mK}$ with the same variance as $(\phi_1^{(K)}(Y_1),\dots,\phi_K^{(K)}(Y_1))^\top$. For $j\in[m]$ and $K\in\mathbb{N}$, the penultimate approximation for $U_j^{\mathrm{H}}(Y)$ is therefore given as a polynomial of $\mathbb{R}^K$ Gaussians,

$$\begin{aligned} U_j^{(K)}(\Xi^{(K)}) \;&:=\; \frac{1}{n(n-1)\dots(n-j+1)}\sum_{i_1,\dots,i_j\in[n]\text{ distinct}}\tilde{u}_j^{(K)}(\xi_{i_1}^{(K)},\dots,\xi_{i_j}^{(K)})\;, \qquad (5.7)\\ u_j^{(K)}(v_1,\dots,v_j) \;&:=\; \sum_{k_1,\dots,k_m=1}^{K}\lambda_{k_1\dots k_m}^{(K)}v_{1k_1}\cdots v_{jk_j}\,\mathbb{E}\big[\phi_{k_{j+1}}^{(K)}(Y_1)\big]\cdots\mathbb{E}\big[\phi_{k_m}^{(K)}(Y_1)\big]\;. \end{aligned}$$

Moments. By variance domination, the asymptotic distribution of $u_m(Y)$ is determined by the relative size of the rescaled variances, given for $M\in[m]$ as

$$\rho_{m,n;M} \;:=\; \frac{\sum_{r\in[m]\setminus\{M\}}\sigma_{m,n;r}^2}{\sigma_{m,n;M}^2}\;,$$
$$\text{where}\quad \sigma_{m,n;j}^2 \;:=\; \binom{m}{j}^2\binom{n}{j}^{-1}\mathrm{Var}\,\mathbb{E}[u(Y_1,\dots,Y_m)\,|\,Y_1,\dots,Y_j] \qquad \text{for } j\in[m]\;.$$

$\sigma_{m,n;j}^2$ describes the contribution of U_j^{H} to the overall variance of $u_m(Y)$ (Lemma C.11 in the appendix). The bound also depends on a Berry-Esseen moment ratio, given for $\nu\in(2,3]$ and $M\in[m]$ as

$$\tilde{\beta}_{M,\nu} \;:=\; \frac{\|u_M^{\mathrm{H}}(Y_1,\dots,Y_M)\|_{L_\nu}}{\|u_M^{\mathrm{H}}(Y_1,\dots,Y_M)\|_{L_2}}\;.$$

Proposition 5.6. *Suppose Assumption 5.1 holds for some $\nu\in(2,3]$. Then there is some absolute constant $C>0$ such that for every $n,m,d\in\mathbb{N}$ and $M\in[m]$, we have*

$$\begin{aligned} &\sup_{t\in\mathbb{R}}\Big|\mathbb{P}\big(u_m(Y)-\mathbb{E}[u(Y_1,\dots,Y_m)]\le t\big)-\lim_{K\to\infty}\mathbb{P}\Big(\binom{m}{M}U_M^{(K)}(\Xi^{(K)})\le t\Big)\Big|\\ &\qquad\le\; Cm\Big(n^{-\frac{\nu-2}{2(\nu M+1)}}\tilde{\beta}_{M,\nu}^{\frac{\nu}{\nu M+1}}+\rho_{m,n;M}^{\frac{1}{2M+1}}\Big)\;. \end{aligned}$$

Moreover, writing Φ as the c.d.f. of $\mathcal{N}(0,1)$, we have that

$$\begin{aligned} &\sup_{t\in\mathbb{R}}\Big|\mathbb{P}\big(u_m(Y)-\mathbb{E}[u(Y_1,\dots,Y_m)]\le t\big)-\lim_{K\to\infty}\Phi\Big(\binom{m}{M}^{-1}\mathrm{Var}\big[U_M^{(K)}(\Xi^{(K)})\big]^{-1/2}t\Big)\Big|\\ &\quad\le Cm\Big(n^{-\frac{\nu-2}{2(\nu M+1)}}\beta_{M,\nu}^{\frac{\nu}{\nu M+1}}+\rho_{m,n;M}^{\frac{1}{2M+1}}\Big)+\lim_{K\to\infty}\Big(\frac{4m-4}{3m}\Big|\mathrm{Kurt}\big[U_M^{(K)}(\Xi^{(K)})\big]\Big|\Big)^{1/2}\;. \end{aligned}$$

We emphasise that Proposition 5.6 allows the degree of the U-statistic m to be much larger than $\log n$: In the above bound, when $\tilde{\beta}_{M,\nu}=O(1)$ and $\rho_{m,n;M}=o(1)$, we only require the degree of the polynomial of Gaussians to satisfy $M\log M=o(\log n)$, with no explicit requirement on m. As with Proposition 5.5, under these conditions, the bound also does not have any explicit dependence on d. This allows Proposition 5.6 to recover many known results on high-dimensional U-statistics. Similar to Proposition 5.5,

there are two cases under which a normal approximation is allowed, but with different variances:

- When $M = 1$, the distributional limit $\lim_{K\to\infty} U_1^{(K)}(\Xi^{(K)})$, when exists, is Gaussian. In this case, Proposition 5.6 provides a finite-sample bound for the Gaussian approximation of $u_m(Y)$, a potentially infinite-order U-statistic (IOUS). Notably if $\tilde{\beta}_{M,\nu}$ is bounded, the only dependence on m in the bound is through $\rho_{m,n;M}$; when d is fixed and $\mathrm{Var}\,\mathbb{E}[u(Y_1,\dots,Y_m)\,|\,Y_1,\dots,Y_j] \in (0,\infty)$ is fixed for all $j \in [m]$, we can compute

$$\rho_{m,n;1} \le \frac{n}{m^2}\sum_{r=2}^{m}\left(\frac{m^2e^2}{n}\right)^r = O\left(\frac{m^2}{n}\right) \qquad \text{if } m^2 < \frac{n}{2e^2}\,.$$

  Therefore when d is fixed and $M = 1$, Gaussianity holds if $m = o(n^{-1/2})$. This agrees with known optimal growth condition on m for Gaussianity of 1d IOUS (van Es and Helmers, 1988). On the other hand, when d and the variances are allowed to change, $\rho_{m,n;1}$ has more complex dependence on n; this can lead to a more stringent or relaxed condition on the size of m.

- When $M > 1$, $U_M^{(K)}(\Xi^{(K)})$ is a degree-M polynomial of Gaussians. By the fourth moment phenomenon (Proposition 4.6), the asymptotic limit can still be Gaussian in high dimensions, for which Proposition 5.5 provides a finite-sample bound. This is a behaviour specific to the high-dimensional regime, where the Gaussian polynomial approximation is allowed to vary in n.

The above two cases recover many existing works on the normal approximation of a high-dimensional degree-two U-statistic: Some of those results come without a fourth-moment condition and require assumptions similar to those for (i) (Chen and Qin, 2010; Harchaoui et al., 2020; Wang et al., 2015; Yan and Zhang, 2022), whereas others consider a fourth moment condition and fall under (ii) (Gao and Shao, 2023; Bhattacharya et al., 2022). A practical consequence of (ii) is that a degenerate U-statistic – often found in hypothesis testing and which requires numerical approximation due to its classical non-Gaussianity Leucht and Neumann (2013) – may be Gaussian in high dimensions. In view of the necessity statement in Proposition 4.6, one may argue that the two cases cover most, if not all of the situations of Gaussianity.

Meanwhile, as with Proposition 5.5, Proposition 5.6 highlights when $u_m(Y)$ may be asymptotically non-Gaussian. Here, non-Gaussianity happens not according to the degeneracy of the U-statistic, but depending on the relative sizes of the rescaled variances $\sigma^2_{m,n;j}$, each corresponding to the variance of the degree-j Hoeffding decomposition. If the ratio of the variances change as d grows, the limiting distribution of $u_m(Y)$ can transition from one low-degree polynomial of Gaussians to a higher one, or vice versa. This

change-of-asymptotic-limit effect generalises the observation in Chapter 3 for degree-two U-statistics. In Section 3.4, we have also seen empirical evidences of the transition for a U-statistic used in high-dimensional kernel-based distribution tests: There, the ratio of variances reduces to a comparison between problem-specific hyperparameter choices, which determine which asymptotic limit dominates.

Remark 5.11. The same argument in this section can be extended to a U-statistic of non-identically distributed data, although a more elaborate decomposition than (5.6) is required to accommodate for the fact that $\mathbb{E}[u(y_1, \dots, y_r, Y_1, \dots, Y_{m-r})]$ may not be equal to $\mathbb{E}[u(y_1, \dots, y_r, Y_2, \dots, Y_{m-r+1})]$. Section 5.4 illustrates such an example.

5.4 Finite-sample bounds for subgraph count statistics

We apply our results to characterise the possible asymptotic distributions of subgraph counts. We shall see that variance domination (Theorem 4.2) recovers and extends the geometric conditions considered by Hladký et al. (2021) and Bhattacharya et al. (2023) for obtaining different limits, and our results also apply to the edge-level fluctuations considered by Kaur and Röllin (2021).

Graph model. Given a symmetric, measurable function $w : [0,1]^2 \to [0,1]$, we generate a graph G_n with vertex set $[n]$ by drawing independent (not necessarily identically distributed) random variables $(U_i)_{i \le n}$ and $(V_{ij})_{1 \le i < j \le n}$, and joining an edge $i \sim j$ according to

$$Y_{ij} \;:=\; \mathbb{I}_{\{V_{ij} \le w(U_i, U_j)\}} \,.$$

When $U_i, V_{ij} \overset{\text{i.i.d.}}{\sim} \text{Uniform}[0,1]$, this generates an exchangeable graph from w (Diaconis and Janson, 2008).

Subgraph count statistics. Denote $K(S)$ as the complete graph on the vertex set $S \subseteq [n]$, $K_n := K(\{1, \dots, n\})$ and $E(H')$ as the edge set of a graph H'. Given a non-empty simple graph H with $m = m(n)$ vertices and $k = k(n)$ edges, we are interested in the number of subgraphs in G_n that are isomorphic to H, described by

$$\kappa(Y) \;:=\; \sum\nolimits_{1 \le i_1 < \dots < i_m \le n} \; \sum\nolimits_{H' \subseteq \mathcal{G}_H(\{i_1, \dots, i_m\})} \; \prod\nolimits_{(i_s, i_t) \in E(H')} Y_{i_s i_t} \,,$$

where $\mathcal{G}_H(S)$ denotes the collection of all subgraphs of $K(S)$ that are isomorphic to H.

Consider the conditionally centred indicator variable $\bar{Y}_{ij} := Y_{ij} - w(U_i, U_j)$. To derive the asymptotic limit of $\kappa(Y)$, we first split $\kappa(Y)$ into a sum over vertices and a sum over edges:

$$\kappa(Y) \;=\; \sum\nolimits_{1 \le i_1 < \dots < i_m \le n} \; \sum\nolimits_{H' \subseteq \mathcal{G}_H(\{i_1, \dots, i_m\})} \; \prod\nolimits_{(i_s, i_t) \in E(H')} w(U_{i_s}, U_{i_t})$$

$$+ \sum_{\substack{\{(i_1,j_1),\dots,(i_k,j_k)\}\subseteq E(K_n) \\ \text{is a set of distinct edges}}} \delta_H\big(\{(i_s,j_s)\}_{s\in[k]}\big) \prod_{s=1}^{k} \bar{Y}_{i_s j_s} \;=:\; \kappa_1(U) + \kappa_2(\bar{Y}) \;,$$

where $\delta_H(I)$ is the indicator function on whether the graph formed by the edge set I is isomorphic to H. $\kappa_1(U)$ represents vertex-level fluctuations, and $\kappa_2(\bar{Y})$ represents edge-level fluctuations. We first study the two fluctuations separately, and show how they affect the distribution on $\kappa(Y)$ by variance domination.

Vertex-level fluctuations. Let $U := (U_i)_{i\le n}$ be i.i.d. for notational simplicity. $\kappa_1(U)$ is now the degree-m U-statistic considered in Proposition 5.6, with the kernel

$$u_1(x_1,\dots,x_m) \;:=\; \binom{n}{m} \sum_{H'\subseteq \mathcal{G}_H(\{1,\dots,m\})} \prod_{(i_s,i_t)\in E(H')} w(x_{i_s},x_{i_t}) \;.$$

Applying Proposition 5.6 directly gives the following corollary, which approximates $\kappa_1(U)$ by its M-th Hoeffding decomposition. The notation $U_M^{(K)}(\Xi^{(K)})$, $\tilde{\beta}_{M,\nu}$, $\sigma^2_{m,n;M}$ and $\rho_{m,n;M}$ are defined in terms of the U-statistic $\kappa_1(U)$.

Corollary 5.7. *Suppose Assumption 5.1 holds for the kernel function u_1 with some $\nu \in (2,3]$. Then there is some absolute constant $C > 0$ s.t. for every $n,m,d \in \mathbb{N}$ and $M \in [m]$,*

$$\sup_{t\in\mathbb{R}} \Big|\mathbb{P}\big(\kappa_1(U) - \mathbb{E}[\kappa_1(U)] \le t\big) - \lim_{K\to\infty} \mathbb{P}\Big(\binom{m}{M} U_M^{(K)}(\Xi^{(K)}) \le t\Big)\Big| \\ \le\; Cm\Big(n^{-\frac{\nu-2}{2(\nu M+1)}} \tilde{\beta}_{M,\nu}^{\frac{\nu}{\nu M+1}} + \rho_{m,n;M}^{\frac{1}{2M+1}}\Big) \;.$$

The variance ratio $\rho_{m,n;M} = \frac{\sum_{r\in[m]\setminus M} \sigma^2_{m,n;r}}{\sigma^2_{m,n;M}}$ can be computed in terms of

$$\sigma^2_{m,n;r} \;= \binom{m}{r}\binom{n}{m}\binom{n-r}{m-r} \operatorname{Var} \mathbb{E}\left[\sum_{H'\subseteq \mathcal{G}_H(\{1,\dots,m\})} \prod_{(i_s,i_t)\in E(H')} w(U_{i_s},U_{i_t}) \,\middle|\, U_1,\dots,U_r \right] .$$

Remark 5.12. (i) A common tool for proving limit theorems for subgraph counts (Kaur and Röllin, 2021; Bhattacharya et al., 2023) is the orthogonal decomposition of a generalised U-statistic proposed by Janson and Nowicki (1991). Here, this assumption takes the form of Assumption 5.1: the difference is that we focus only on vertices, and perform a decomposition in an L_ν space with $\nu > 2$ (see Remark 5.10). (ii) For Gaussian approximation and for $\nu = 3$ (i.e. existence of the appropriate third moments), we obtain a rate of $n^{-1/8}$ in Kolmogorov distance plus an additional variance domination error of $\rho_{m,n;1}^{1/3}$. In comparison, Kaur and Röllin (2021) obtains a bound at the rate of $n^{-1/(2(m+2))}$ in a convex set distance; we expect that their approximation error to be related to our $\rho_{m,n;1}^{1/3}$ and that their bound, obtained by Stein's method, is sharper.

As before, the size of $\sigma^2_{m,n;M}$ – the variance of the M-th order Hoeffding decomposition – determines whether $\kappa_1(U)$ can be approximated by a degree-M polynomial of Gaussians, This is known in the literature as the $n^{m-\frac{M}{2}}$-th order fluctuation (Kaur and

Röllin, 2021), as evident from the $\binom{n}{m}\binom{n-M}{m-M} = O(n^{2m-M})$ factor in $\sigma^2_{m,n;M}$. A guiding example is when $H = K_2$, i.e. $\kappa(Y)$ is the edge count: In this case, the variances of the Gaussian component and the quadratic-form-of-Gaussian component are respectively

$$\sigma^2_{m,n;1} = n(n-1)^2 \operatorname{Var} \mathbb{E}[w(U_1, U_2)|U_1] \quad \text{and} \quad \sigma^2_{m,n;2} = \frac{n(n-1)}{2} \operatorname{Var}[w(U_1, U_2)] \,,$$

which are the variances of the $n^{3/2}$-th order fluctuation and the n-th order fluctuation.

The next result gives conditions under which variance domination by the Gaussian component ($M = 1$, or $n^{m-\frac{1}{2}}$-th order fluctuation) fails.

Lemma 5.8. *Let w, H and $m \geq 2$ be fixed. Then $\sigma^2_{m,n;r} = O(n^{2m-r})$ for all $r \in [m]$. Moreover, the following statements are equivalent:*

(i) $\rho^2_{m,n;1} = \Omega(1)$;
(ii) $\sigma^2_{m,n;1} = 0$;
(iii) $\sum_{H' \subseteq \mathcal{G}_H(\{1,\dots,m\})} \mathbb{E}\Big[\prod_{(i_s,i_t)\in E(H')} w(U_{i_s}, U_{i_t}) \,\Big|\, U_1 \Big]$ *is constant almost surely;*
(iv) For almost every $x \in [0,1]$,

$$\frac{1}{m} \sum_{i=1}^{m} \mathbb{E}\Big[\prod_{(i_s,i_t)\in E(H)} w(U_{i_s}, U_{i_t}) \,\Big|\, U_i = x \Big] = \mathbb{E}\Big[\prod_{(i_s,i_t)\in E(H)} w(U_{i_s}, U_{i_t}) \Big] \,.$$

Lemma 5.8(i) is the condition derived from variance domination, whereas Lemma 5.8(iv) is the definition of H-regularity of w, introduced by Hladký et al. (2021) for a complete subgraph H and extended to a general subgraph H by Bhattacharya et al. (2023). Indeed, Theorem 1.2 of Hladký et al. (2021) and Theorem 2.9 of Bhattacharya et al. (2023) establish that $\kappa_1(U)$ exhibit fluctuations of an order higher than $n^{m-\frac{1}{2}}$ if and only if H-regularity holds – a geometric condition that is recovered by variance domination. Meanwhile, variance domination provides more: We can now characterise all fluctuations in $\kappa_1(U)$ on the orders $n^{m-\frac{1}{2}}, n^{m-1}, \dots, n^{\frac{m}{2}}$, all with finite-sample bounds. Corollary 5.7 may also be applied to a setup where w and H are allowed to vary in n.

Remark 5.13. In the case of $H = K_2$, Lemma 5.8(iii) says that $\mathbb{E}[w(U_1, U_2) \,|\, U_1]$ is constant almost surely. In other words, the edge count statistic $\kappa(Y)$ exhibits n-th order fluctuations if and only if the random graph is regular on average.

Edge-level fluctuations. In practice, one may want to analyse edge properties of the random graph, which requires us to study $\kappa_2(\bar{Y})$. Kaur and Röllin (2021) provide error bounds on the Gaussian approximation of $\kappa_2(\bar{Y})$ by using the orthogonal decomposition of a generalised U-statistic of U and V proposed by Janson and Nowicki (1991). We consider a variant of their setup. For convenience, we re-index $(\bar{Y}_{ij})_{i,j\in E(K_n)}$ as $\bar{Y}_1, \dots, \bar{Y}_{n_*}$, where $n_* = |E(K_n)| = \binom{n}{2}$; recall that k is the number of edges of H. Conditioning on

$U = (U_i)_{i \le n}$, let $Z = (Z_i)_{i \le n_*}$ be conditionally normal variables defined by

$$Z_i \,|\, U \overset{i.i.d.}{\sim} \mathcal{N}\big(\mathbb{E}[\bar{Y}_i \,|\, U]\,,\, \mathrm{Var}[\bar{Y}_i \,|\, U]\big)\ .$$

The penultimate approximation for $\kappa_2(\bar{Y})$ is the incomplete U-statistic

$$\kappa_2(Z) \;=\; \sum_{\substack{\{e_1,\dots,e_k\} \subseteq E(K_n) \\ \text{is a set of distinct edges}}} \delta_H\big(\{e_s\}_{s\in[k]}\big) \prod_{s=1}^{k} Z_{e_s}\ .$$

Proposition 5.9. *If* $\mathrm{Var}[\kappa_2(Z) \,|\, U] > 0$ *almost surely, there exist some absolute constant* $C > 0$ *such that for every* $n \ge 2$ *and* $m, k \in \mathbb{N}$*, almost surely,*

$$\sup_{t\in\mathbb{R}} \Big| \mathbb{P}\big(\kappa_2(\bar{Y}) \le t \,\big|\, U\big) - \mathbb{P}\big(\kappa_2(Z) \le t \,\big|\, U\big) \Big| \;\le\; \Delta_{\bar{Y}}(U)\ ,$$

where $\Delta_{\bar{Y}}(U) \coloneqq C\, m\, n^{-\frac{\nu-2}{\nu k+1}}\, \rho_{\bar{Y}}(U)^{\frac{1}{\nu k+1}}$ *and*

$$\rho_{\bar{Y}}(U) \coloneqq \frac{\max_{i_1,\dots,i_k \in [n_*] \text{ distinct}} \prod_{l=1}^{k} \mathbb{E}\big[|\bar{Y}_{i_l}|^{\nu} \,|\, U\big]}{\min_{i_1,\dots,i_{k'} \in [n_*] \text{ distinct}} \prod_{l=1}^{k'} \mathbb{E}\big[|\bar{Y}_{i_l}|^{2} \,|\, U\big]^{\nu/2}}\ .$$

Moreover, writing Φ *as the c.d.f. of* $\mathcal{N}(0,1)$*, we have that almost surely*

$$\sup_{t\in\mathbb{R}} \Big| \mathbb{P}\big(\kappa_2(Y) - \mathbb{E}[\kappa_2(Y) \,|\, U] \le t \,\big|\, U\big) - \Phi\big(\mathrm{Var}[\kappa_2(Z) \,|\, U]^{-1/2}\, t\big) \Big|$$
$$\le \Delta_{\bar{Y}}(U) + \Big(\frac{4m-4}{3m} \Big|\mathrm{Kurt}[\kappa_2(Z) \,|\, U]\Big|\Big)^{1/2}\ ,$$

where we defined $\mathrm{Kurt}[\kappa_2(Z)|U] \coloneqq \mathbb{E}[(\kappa_2(Z) - \mathbb{E}[\kappa_2(Z)|U])^4|U] \,/\, \mathrm{Var}[\kappa_2(Z)|U]^2 - 3$.

Remark 5.14. (i) Observe that $\kappa(Y) - \mathbb{E}[\kappa(Y)|U] = \kappa_2(\bar{Y})$ almost surely, since $\kappa_1(U)$ is fixed under the conditioning on U. (ii) $\kappa_2(\bar{Y})$ is an incomplete U-statistic, where the asymmetry arises from δ_H, and $(\bar{Y}_i)_{i\le n_*}$ conditioning on U are non-identically distributed. This is an example of how our main results can be applied to asymmetric estimators and non-identically distributed data.

Remark 5.15. The max-min ratio $\rho_{\bar{Y}}$ can be avoided by directly using the error bound from Theorem 4.1: In exchange, the bound is given as a sum of non-identical moment terms and the convergence rate cannot be read off directly. This tradeoff is common for asymmetric estimators and non-identically distributed data, although simplifications may be possible for specific statistics.

Overall fluctuations. We now consider the asymptotic distribution of $\kappa(Y) = \kappa_1(U) + \kappa_2(\bar{Y})$. Specifically, we investigate whether the vertex-level fluctuations $\kappa_1(U)$ or the edge-level fluctuations $\kappa_2(\bar{Y})$ dominates in variance in $\kappa(Y)$. The variances of the different Hoeffding's decompositions of $\kappa_1(U)$ have been provided in Lemma 5.8, whereas the next result provides the variance of $\kappa_2(\bar{Y})$.

Lemma 5.10. *There are some absolute constants $c, C > 0$ such that*

$$\begin{aligned}\mathrm{Var}[\kappa_2(\bar{Y})] &\leq C^k \, |\mathcal{G}_H([n])| \, \mathbb{E}\Big[\max_{i_1,\ldots,i_k \in [n_*] \text{ distinct}} \prod_{l=1}^k \mathrm{Var}[\bar{Y}_{i_l} \,|\, U] \Big] \,, \\ \mathrm{Var}[\kappa_2(\bar{Y})] &\geq c^k \, |\mathcal{G}_H([n])| \, \mathbb{E}\Big[\min_{i_1,\ldots,i_k \in [n_*] \text{ distinct}} \prod_{l=1}^k \mathrm{Var}[\bar{Y}_{i_l} \,|\, U] \Big] \,.\end{aligned}$$

Consider again the setup with w, H, m and k fixed. Let $\mathrm{Aut(H)}$ be the set of all automorphisms of H. Provided that the expectation terms in Lemma 5.10 are $\Theta(1)$, we have

$$\mathrm{Var}[\kappa_2(\bar{Y})] \;=\; \Theta(|\mathcal{G}_H([n])|) \;=\; \Theta\Big(\binom{n}{m} \frac{m!}{|\mathrm{Aut(H)}|} \Big) \;=\; \Theta(n^m) \,.$$

Meanwhile, when the dominating term of $\kappa_1(U)$ is its M-th Hoeffding's decomposition, $\sigma^2_{m,n;M}$ is assumed to be non-zero, and Lemma 5.8 implies

$$\sigma^2_{m,n;M} \;=\; \Theta(n^{2m-M}) \,.$$

As a result, $\kappa_1(U)$ always dominates $\kappa_2(\bar{Y})$ in variance when $M < m$. This agrees with the observation in Kaur and Röllin (2021), who provide a detailed illustration for simple subgraphs. Consequently, analysing the distribution of $\kappa(Y)$ usually means that edge-level randomness is ignored except for the case with the highest level of degeneracy.

Some difficulties persist in the most general case: (i) Due to asymmetry and non-identically distributed data, the bounds can be loose or not immediately interpretable, as discussed in Remark 5.15. Under specific random graph models, bounds can simplify considerably (Hladký et al., 2021). (ii) Even if a penultimate approximation is established, the limiting distribution for an asymmetric polynomial of Gaussians is highly dependent on the weights and may not be immediately obvious. Different choices of such penultimate approximations can yield natural graph-based interpretations (Kaur and Röllin, 2021; Bhattacharya et al., 2023). In a concurrent work, Chatterjee, Dan, and Bhattacharya (2024) obtain limiting distributions for graphon models with higher-order degeneracies and for joint subgraph counts, and additional geometric insights.

Chapter 6

Effects of data augmentation via block dependence

The universality results developed so far focus on independent data, with a brief discussion of applicability to block dependence in remark (v) after Theorem 4.1. The main goal of this chapter is to develop universality results for analysing the effects of *data augmentation* — an ubiquitous technique in machine learning that exhibits both block dependence and high-dimensionality. Notably, while the examples in the preceding Chapters 3 and 5 are well-described by polynomials, this is not the case for the estimators we examine here. The main technical hurdle is to find the appropriate transformation under which the estimator behaves like a polynomial (see remark (iv) after Theorem 4.1), in the presence of block dependence and a growing dimensionality parameter. We also note that the results in this section consider functions with $\mathbb{R}^q$ output for q fixed (instead of $q = 1$ in the preceding chapters), and are finite-sample with respect to a smooth metric but asymptotic with respect to the Kolmogorov metric. To obtain a finite-sample bound in the Kolmogorov metric, one may apply our general results in Chapter 4 across the q coordinates; we do not focus on this here to avoid the need to consider anti-concentration bounds.

We first informally motivate the role of data augmentation in machine learning. The term data augmentation refers to a range of machine learning heuristics that synthetically enlarge a training data set[†]: Random transformations are applied to each training data point, and the transformed points are added to the training data (e.g. Taqi et al., 2018; Shorten and Khoshgoftaar, 2019). It has quickly become one of the most widely used heuristics in machine learning practice, and the scope of the term continues to evolve. One objective may be to make a neural network less sensitive to rotations of input images, by augmenting data with random rotations of training samples (e.g. Perez and Wang, 2017). In other cases, one may simply reason that "more data is always better".

The question how data augmentation affects learning rates remains open. It has been argued that augmentation reduces the variance of estimates (Zhang et al., 2021), that

[†]This meaning of the term data augmentation should not be confused with a separate meaning in statistics, which refers to the use of latent variables e.g. in the EM algorithm.

it increases the effective sample size (Balestriero et al., 2022b), and that it acts as a regulariser (Balestriero et al., 2022a), but none of these points have been rigorously established in great generality. Existing analysis studies the bias of estimates (Balestriero et al., 2022a), and shows a reduction of variance for certain *parametric* M-estimators under additional invariance assumptions (Chen et al., 2020). In the following, we study the limiting behaviour of augmentation methods. Two mathematical obstacles are (1) that augmentation makes independently distributed data dependent, and (2) that data may be high-dimensional. One may therefore expect the behaviour of augmented estimates to be highly sensitive to the input distribution. We show that, on the contrary, augmented statistics exhibit a form of universality under general stability conditions.

The key difference of the results in this chapter, as compared to the i.i.d. case, is that data augmentation introduces strong dependence that persists asymptotically. By approximating (appropriately transformed versions of) the data by Gaussians, the effect of such strong dependence manifests only through the first two moments, and the analysis reduces to understanding how the changes in the mean and variance affect properties of the estimator. Our findings show that a number of properties commonly attributed to data augmentation — variance reduction, increase in effective sample size, and regularisation — each occur in certain cases, but fail in others.

The rest of the chapter is organised as follows. Section 6.1 provides a high-level, non-technical overview of the results. Section 6.2 defines the setup and the concept of noise stability. Theoretical results—the main theorem and a number of consequences—follow in Section 6.3. The remaining sections apply these results to the variance analysis of plug-in estimators and ridge regression (Section 6.4) and the regularisation effects on an overparameterised model that exhibits double descent (Section 6.5). We also demonstrate the applicability of our universality results to other non-smooth and high-dimensional estimators in Section 6.6; in view of Theorem 4.1, these provide additional examples of how to find the "appropriate transformation" of the estimator to perform a low-degree polynomial approximation. All proofs are collected in Appendix D.

The examples we focus on in this chapter are on regression tasks, but the same universality approximation can be obtained for estimators in classification tasks. A follow-up work on logistic regression is considered in the joint work of Mallory, Huang, and Austern (2025), which is discussed in Chapter 7.

6.1 A non-technical overview

Here, we sketch the results informally to provide a high-level overview. Rigorous definitions follow in Section 6.2. Our general setup is as follows: Consider a dataset, consisting

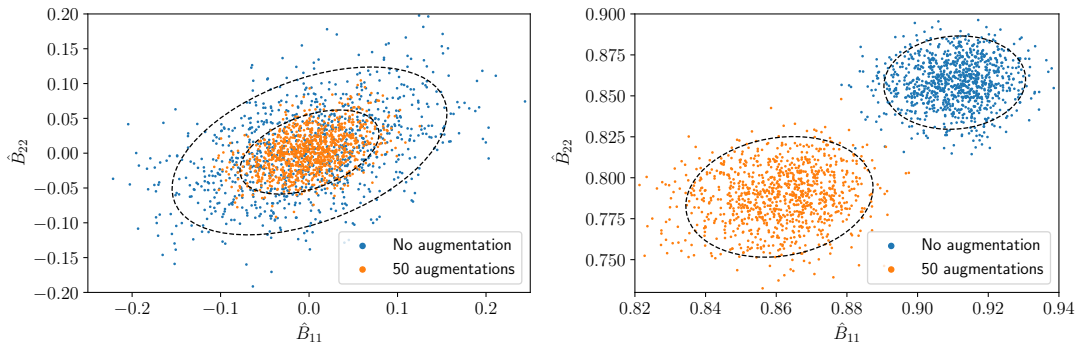


Figure 6.1: Effect of augmentation on the variability of estimates. *Left:* On an empirical average. *Right:* On a ridge regression estimator. Each point is an estimate computed from a single simulation experiment, and the dashed lines are the 95% 2d quantiles of the empirical distribution over 1000 simulations. Augmentation reduces the variability in the left plot, but increases the uncertainty of the estimate in the right plot. See Remark 6.3 in Section 6.4.5 for details on the plotted experiments.

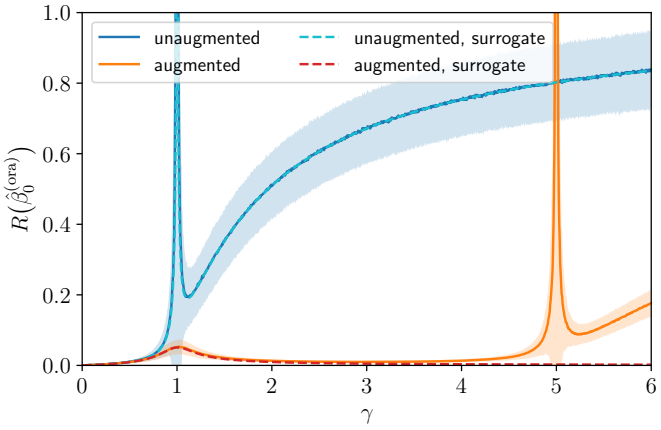


Figure 6.2: Effect of an oracle choice of augmentation on the limiting risk of a high-dimensional ridgeless regressor under the asymptotic $d/n \to \gamma$. A regularisation effect is observed around $\gamma = 1$, whereas a new double-descent peak shows up at $\gamma = 5 = k$, the number of augmentations. See Section 6.5.1 for the detailed setup.

of observations that we assume to be d-dimensional i.i.d. random vectors in $\mathcal{D} \subseteq \mathbb{R}^d$. We are interested in estimating a quantity $\theta \in \mathbb{R}^q$, for some q. This may be a model parameter, the value of a risk function or a statistic, and so forth. The data is augmented by applying k randomly generated transformations to each data point. That yields an augmented data set of size $n \cdot k$. An estimator for θ is then a function $f : \mathcal{D}^{nk} \to \mathbb{R}^q$, and we estimate θ as

$$\text{estimate of } \theta \;=\; f(\text{augmented data}) \;.$$

From a statistical perspective, this can be regarded as a form of sample randomisation. As for other randomisation techniques, such as the bootstrap or cross-validation, quantitative analysis of augmentation is complicated by the fact that randomised data points are not independent. To study such augmented estimates, we rely on the Linderberg method discussed in Chapter 2, and assume that our statistics f satisfy a "noise stability" condi-

tion (see Section 6.2). Informally, noise stability means that f is not too sensitive to small perturbations of any input coordinate. Examples of noise-stable statistics include sample averages (such as empirical risks or plug-in estimators), but also overparameterised linear regression, ridge regression, bagged estimators, and general M-estimators (Mei and Montanari, 2022; Soloff et al., 2024; Montanari and Saeed, 2022). Our Theorem 6.1 shows that the distribution of our augmented estimator is identical to the distribution of an estimator trained on some surrogate random variables. More precisely, for all h in a certain class $\mathcal{H}$ of smooth functions, we show that

$$\left| \mathbb{E}[h(f(\text{augmented data}))] - \mathbb{E}[h(f(\text{generic surrogate variables}))] \right| \leq \tau(n, k) \, .$$

The surrogates are variables completely determined by their mean and variance; depending on the problem, they may be Gaussian (e.g. for sample averages) or non-Gaussian (e.g. for ridge regression). Under general conditions, $\tau \to 0$, hence the limiting distribution of $f(\text{augmented data})$ is that of $f(\text{surrogates})$. In other words, the effect of augmentation on a noise-stable estimator is *completely determined by two moments* as n grows large. The theorem specifies these moments explicitly. That allows us to study the limiting estimator and its variance, and to read off the rate of convergence from τ. For sufficiently linear estimators, we can also draw consistent confidence intervals and evaluate their width.

Applications to specific models. The function τ is determined by terms that quantify the noise stability of f. For a given estimator, we can evaluate these terms to verify how fast τ converges to 0 as either n or k grows large. This establishes how fast the universality property happens, and we use this to gain insights into the effect of data augmentation for a few different models:

1) Underparameterised models. We analyse empirical averages (Section 6.4.2), plug-in estimators, the risk of M-estimators (Section 6.4.3) and ridge regression (Section 6.4.5). For empirical averages and risks, we characterise exactly when augmentation reduces variance. These results hold more generally for a class of linear sample statistics. For non-linear estimators, the behaviour can change significantly: Augmentation may increase rather than decrease variance. That can occur even in simple models, such as the ridge regression example (see the right plot of Figure 6.1).

2) Overparameterised models. As an example of an overparameterised model, we analyse the limiting risk of a high-dimensional ridgeless regressor. Without augmentation, this model is known to exhibit double descent (Hastie et al., 2022). We show that the behaviour under augmentation depends on an interplay of scales: If $d \approx n$, augmentation acts as a regulariser. For higher dimension, namely $d \approx nk$, it causes the risk to diverge to infinity. It can also shift the double-descent peak—see Figure 6.2.

Some key findings about the behaviour of data augmentation. To place our results in context, we note three hypotheses generally made in the existing literature of data augmentation and are either explicitly or implicitly required by proofs (e.g. Dao et al., 2019; Chen et al., 2020; Balestriero et al., 2022b): (i) Linearity or approximate linearity of the estimator, in the sense that f is linear in contributions of individual data points (typically, a sample average). (ii) Invariance of the data source, i.e. the transformations used to perform augmentation leave the data distribution invariant. (iii) The number of transformations applied to each data point diverges, i.e. $k \to \infty$. In the context of (iii), it is helpful to note that transformations can be applied once before fitting a model (*offline augmentation*), or repeatedly during each step of a training algorithm (*online augmentation*). Online augmentation is feasible if each transformation is computationally cheap (e.g. rotations in computer vision). Offline augmentation is particularly common in natural language processing, where more expensive transformations have emerged as useful (Feng et al., 2021). The assumption $k \to \infty$ is justified by choosing an online setup and arguing that the number of steps of the training algorithm is effectively infinite; offline augmentation implies $k < \infty$. Theorem 6.1 allows us to drop each of these assumptions, and overall, our results show that doing so can change the behaviour of augmentation decisively. In more detail, our results show the following:

1) Augmentation may or may not reduce variance. Augmentation is known to reduce variance under assumptions (i)—(iii) above, but empirical observations by Lyle et al. (2019) suggest this may not be true in practice. Theorem 6.1 allows us to make more detailed statements: If f is linear, augmentation reduces variance if the transformations do not increase the variance of the data distribution (Section 6.4.3). If f is non-linear, variance may increase, even if distributional invariance holds (Sections 6.4.4 and 6.4.5). More generally, the effects of augmentation depend not only on the data distribution, but also on the estimator f.

2) Invariance is not essential for augmentation, regardless of whether f is linear or non-linear. For linear f, the relevant criterion for variance reduction is that augmentation does not increase the variance of data variables (Section 6.4.2). The invariance assumption (ii) is one way to ensure this, but is not required: Invariance implies all moments are constant under transformation. What matters is that the second moment does not grow.

3) Augmentation and regularisation. It has been argued that data augmentation can be interpreted as a form of regularisation (e.g. Balestriero et al., 2022a). Our results show that augmentation can indeed act as a regulariser, but whether it does depends on details of the application—specifically, on how the sample size n, the dimension d, and the number k of augmentations per data point grow relative to each other (Section 6.5).

4) Whether augmentation is performed offline or online matters. If $k < \infty$, data

augmentation may not regularise (Section 6.5). This manifests for $d \approx nk$ in the double-descent peak of the risk in Figure 6.2.

In summary, Theorem 6.1 can be used to derive statistical guarantees for a range of augmented estimators. Several hypotheses on augmentation considered in machine learning turn out not to be either true or false, but rather depend on the data distribution, the properties of the estimator, and the interplay of sample size, number of augmentations, and dimension. The results may also be a step towards making data augmentation a viable technique for statisticians who seek guarantees for the methods they employ.

6.2 Definitions

Data and augmentation. Throughout, we consider a data set $\mathcal{X} := (\mathbf{X}_1, \dots, \mathbf{X}_n)$, where the $\mathbf{X}_i$ are i.i.d. random elements of some fixed convex subset $\mathcal{D} \subseteq \mathbb{R}^d$ that contains $\mathbf{0}$. The choice of $\mathbf{0}$ is for convenience and can be replaced by any other reference point. Let $\mathcal{T}$ be a set of (measurable) maps $\mathcal{D} \to \mathcal{D}$, and fix some $k \in \mathbb{N}$. We generate nk i.i.d. random elements $\phi_{11}, \dots, \phi_{nk}$ of $\mathcal{T}$, and abbreviate

$$\Phi_i := (\phi_{ij} | j \leq k) \qquad \Phi := (\phi_{ij} | i \leq n, j \leq k) \qquad \Phi_i \mathbf{X}_i := (\phi_{i1}\mathbf{X}_i, \dots, \phi_{ik}\mathbf{X}_i) \; .$$

The augmented data is then the ordered list

$$\Phi\mathcal{X} \;:=\; (\Phi_1\mathbf{X}_1, \dots, \Phi_n\mathbf{X}_n) \;=\; (\phi_{11}\mathbf{X}_1, \dots, \phi_{1k}\mathbf{X}_1, \dots, \phi_{n1}\mathbf{X}_n, \dots, \phi_{nk}\mathbf{X}_n) \; .$$

Here and throughout, we do not distinguish between a vector and its transpose, and regard the quantities above as vectors $\Phi_i \mathbf{X}_i \in \mathcal{D}^k$ and $\Phi\mathcal{X} \in \mathcal{D}^{nk}$ where convenient.

Estimates. An estimate computed from augmented data is the value

$$f(\Phi\mathcal{X}) \;=\; f(\phi_{11}\mathbf{X}_1, \dots, \phi_{nk}\mathbf{X}_n)$$

of a function $f : \mathcal{D}^{nk} \to \mathbb{R}^q$, for some $q \in \mathbb{N}$. An example is an empirical risk: If S is a regression function $\mathbb{R}^d \to \mathbb{R}$ (such as a statistic or a feed-forward neural network), and $C(\hat{y}, y)$ is the cost of a prediction $\hat{y}$ with respect to y, one might choose $\phi_{ij} = (\pi_{ij}, \tau_{ij})$ as a pair of transformations acting respectively on $\mathbf{v} \in \mathbb{R}^d$ and $y \in \mathbb{R}$ and $\mathbf{X}_i = (\mathbf{V}_i, \mathbf{Y}_i)$, in which case $f(\Phi\mathcal{X})$ is the empirical risk $\frac{1}{nk}\sum_{i \leq n, j \leq k} C(S(\pi_{ij}\mathbf{V}_i), \tau_{ij}\mathbf{Y}_i)$. However, we do *not* require that f is a sum, and other examples are given in Section 6.4.5, 6.5 and 6.6.

Norms. Three types of norms appear in what follows: For vectors and tensors, we use both a "flattened" Euclidean norm and its induced operator norm: If $\mathbf{x} \in \mathbb{R}^{d_1 \times \cdots \times d_m}$ and $A \in \mathbb{R}^{d \times d}$,

$$\|\mathbf{x}\| \;:=\; \Big(\sum\nolimits_{i_1 \leq d_1, \dots, i_m \leq d_m} |x_{i_1, \dots, i_m}|^2\Big)^{1/2} \qquad \text{and} \qquad \|A\|_{op} \;:=\; \sup_{\mathbf{v} \in \mathbb{R}^d} \frac{\|A\mathbf{v}\|}{\|\mathbf{v}\|} \; .$$

Thus, $\|\mathbf{v}\|$ is the Euclidean norm of $\mathbf{v}$ for $m = 1$, the Frobenius norm for $m = 2$, etc. For real-valued random variables X, we also use L_p-norms, denoted by $\|X\|_{L_p} := \mathbb{E}[|X|^p]^{1/p}$.

Covariance structure. For random vectors $\mathbf{Y}$ and $\mathbf{Y}'$ in $\mathbb{R}^m$, we define the $m \times m$ covariance matrices

$$\mathrm{Cov}[\mathbf{Y}, \mathbf{Y}'] \;:=\; (\mathrm{Cov}[Y_i, Y'_j])_{i,j \le m} \qquad \text{and} \qquad \mathrm{Var}[\mathbf{Y}] \;:=\; \mathrm{Cov}[\mathbf{Y}, \mathbf{Y}] \;.$$

Augmentation introduces dependence: Applying independent random elements ϕ and ψ of $\mathcal{T}$ to the same observation $\mathbf{X}$ results in dependent vectors $\phi(\mathbf{X})$ and $\psi(\mathbf{X})$. In the augmented data set, the entries of each vector $\Phi_i \mathbf{X}_i$ are hence dependent, whereas $\Phi_i \mathbf{X}_i$ and $\Phi_j \mathbf{X}_j$ are independent if $i \neq j$. That partitions the covariance matrix $\mathrm{Var}[\Phi\mathcal{X}]$ into $n \times n$ blocks of size $kd \times kd$, and makes it block-diagonal. This block structure is visible in all our results, and makes Kronecker notation convenient: For a matrix $A \in \mathbb{R}^{m \times n}$ and a matrix B of arbitrary size, define the Kronecker product

$$A \otimes B \;:=\; \big(A_{ij} B\big)_{i \le m,\, j \le n}$$

We write $A^{\otimes k} := A \otimes \cdots \otimes A$ for the k-fold product of A with itself. If $\mathbf{v}$ and $\mathbf{w}$ are vectors, $\mathbf{v} \otimes \mathbf{w} = \mathbf{v}\mathbf{w}^\top$ is the outer product. To represent block-diagonal or off-diagonal matrices, let $\mathbf{I}_k$ be the $k \times k$ identity matrix, and $\mathbf{1}_{k \times m}$ a $k \times m$ matrix all of whose entries are 1. Then

$$\mathbf{I}_k \otimes B \;=\; \begin{pmatrix} B & 0 & 0 & \cdots \\ 0 & B & 0 & \\ 0 & 0 & B & \\ \vdots & & & \ddots \end{pmatrix} \qquad \text{and} \qquad (\mathbf{1}_{k \times k} - \mathbf{I}_k) \otimes B \;=\; \begin{pmatrix} 0 & B & B & \cdots \\ B & 0 & B & \\ B & B & 0 & \\ \vdots & & & \ddots \end{pmatrix} .$$

Measuring noise stability. Our results require a control over the noise stability of f and smoothness of test function h, which we define next.

Write $\mathcal{F}_r(\mathcal{D}^a, \mathbb{R}^b)$ for the class of r times differentiable functions $\mathcal{D}^a \to \mathbb{R}^b$. To control how stable a function $f \in \mathcal{F}_r(\mathcal{D}^{nk}, \mathbb{R}^q)$ is with respect to random perturbation of its arguments, we regard it as a function of n arguments $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathcal{D}^k$. That reflects the block structure above—noise can only be added separately to components that are independent. We write $\mathcal{L}(\mathcal{A}, \mathcal{B})$ as the set of bounded linear functions $\mathcal{A} \to \mathcal{B}$, and denote by D_i^m the mth derivative with respect to the ith component,

$$D_i^m f(\mathbf{v}_1, \ldots, \mathbf{v}_n) \;:=\; \frac{\partial^m f}{\partial \mathbf{v}_i^m}(\mathbf{v}_1, \ldots, \mathbf{v}_n) \;\in\; \mathcal{L}\big((\mathcal{D}^k)^m, \mathbb{R}^q\big) \;\subseteq\; \mathbb{R}^{q \times (dk)^m} \;.$$

For instance, if $q = 1$ and g is the function $g(\,\bullet\,) := f(\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}, \bullet, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_n)$, then $D_i^1 f$ is the transposed gradient $\nabla g^\top$, and $D_i^2 f$ is the Hessian matrix of g. To measure the sensitivity of f with respect to each of its $d \times k$ dimensional arguments, we define

$$\mathbf{W}_i(\,\bullet\,) := (\Phi_1 \mathbf{X}_1, \ldots, \Phi_{i-1} \mathbf{X}_{i-1}, \bullet, \mathbf{Z}_{i+1}, \ldots, \mathbf{Z}_n) \;,$$

where $\mathbf{Z}_j$ are i.i.d. surrogate random vectors in $\mathcal{D}^k$ with first two moments matching those

of $\Phi_1\mathbf{X}_1$: Defining the $d \times d$ matrices $\Sigma_{11} := \mathrm{Var}[\phi_{11}\mathbf{X}_1]$ and $\Sigma_{12} := \mathrm{Cov}[\phi_{11}\mathbf{X}_1, \phi_{12}\mathbf{X}_1]$,

$$\mathbb{E}\mathbf{Z}_i \;=\; \mathbf{1}_{k\times 1} \otimes \mathbb{E}[\phi_{11}\mathbf{X}_1] \quad \text{and} \quad \mathrm{Var}\mathbf{Z}_i \;=\; \mathbf{I}_k \otimes \Sigma_{11} + (\mathbf{1}_{k\times k} - \mathbf{I}_k) \otimes \Sigma_{12}\,. \tag{6.1}$$

Write $f_s : \mathcal{D}^{nk} \to \mathbb{R}$ as the s-th coordinate of f. Noise stability is measured by

$$\alpha_r := \sum_{s\le q} \max_{i\le n} \max\Big\{ \Big\| \sup_{\mathbf{w}\in[\mathbf{0},\Phi_i\mathbf{X}_i]} \|D_i^r f_s(\mathbf{W}_i(\mathbf{w}))\| \Big\|_{L_6}, \Big\| \sup_{\mathbf{w}\in[\mathbf{0},\mathbf{Z}_i]} \|D_i^r f_s(\mathbf{W}_i(\mathbf{w}))\| \Big\|_{L_6} \Big\}\,, \tag{6.2}$$

where we have used $[\mathbf{a}, \mathbf{b}]$ to represent the set $\{c\,\mathbf{a} + (1-c)\mathbf{b} : c \in [0,1]\}$. This is a non-negative scalar, and large values indicate high sensitivity to changes of individual arguments (low noise stability). Our results also use test functions $h : \mathbb{R}^q \to \mathbb{R}$. For these, we measure smoothness simply as differentiability, using the scalar quantities

$$\gamma_r(h) \;:=\; \sup\{\|\partial^r h(\mathbf{v})\| \,|\, \mathbf{v} \in \mathbb{R}^q\}\,,$$

where ∂^r denotes the rth differential, i.e. $\partial^1 h$ is the gradient, $\partial^2 h$ the Hessian, etc. In the result below, these terms appear in the form of the linear combination

$$\lambda_h(n,k) \;:=\; \gamma_3(h)\alpha_1^3 + 3\gamma_2(h)\alpha_1\alpha_2 + \gamma_1(h)\alpha_3\,. \tag{6.3}$$

$\lambda_h(n,k)$ can then be computed explicitly for specific models. We note that the dependence on n and k is via the definition of α_r, and that derivatives appear up to 3rd order and moments up to 6th order. Notably, these conditions require that the effect of changing one data point on the first derivative of f is $o(n^{1/3})$.

Moment conditions. Our results also require the following 6th moments on data and the surrogate variables: Write $\mathbf{Z}_1 = (Z_{1jl})_{j\le k, l\le d}$ where $Z_{ijl} \in \mathbb{R}$, and define

$$c_X \;:=\; \frac{1}{6}\sqrt{\mathbb{E}\|\phi_{11}\mathbf{X}_1\|^6} \quad \text{and} \quad c_Z \;:=\; \frac{1}{6}\sqrt{\mathbb{E}\Big[\Big(\frac{|Z_{111}|^2 + \ldots + |Z_{1kd}|^2}{k}\Big)^3\Big]}\,. \tag{6.4}$$

6.3 Universality under block dependence

We now state our main theoretical result and several immediate consequences. With the definitions above, the error bound is a term that measures the noise stability of f and smoothness of h.

Theorem 6.1. (Main result) *Consider i.i.d. random elements $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of $\mathcal{D}$, and two functions $f \in \mathcal{F}_3(\mathcal{D}^{nk}, \mathbb{R}^q)$ and $h \in \mathcal{F}_3(\mathbb{R}^q, \mathbb{R})$. Let $\phi_{11}, \ldots, \phi_{nk}$ be i.i.d. random elements of $\mathcal{T}$ independent of $\mathcal{X}$, $\lambda_h(n,k)$ be defined as in (6.3), and moment terms c_X, c_Z be defined as in (6.4). Then, for any i.i.d. variables $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ in $\mathcal{D}^k$ satisfying (6.1),*

$$\big|\mathbb{E}h(f(\Phi\mathcal{X})) - \mathbb{E}h(f(\mathbf{Z}_1, \ldots, \mathbf{Z}_n))\big| \;\le\; nk^{3/2}\lambda_h(n,k)(c_X + c_Z)\,.$$

Hence if $nk^{3/2}\lambda_h(n,k)(c_X + c_Z) \to 0$, this means that the value $\mathbb{E}h(f(\Phi\mathcal{X}))$ only asymptotically depends on the mean and variance of the augmented samples. We will see that this implies that the distribution of the augmented estimator is universal.

Let v_l denote the l-th coordinate of a vector $v \in \mathbb{R}^q$. Note that if we choose the test function $h : \mathbb{R}^q \to \mathbb{R}$ to be the coordinate functions $h(v) = v_l$ for $1 \leq l \leq q$ as well as the product coordinate functions $h(v) = v_l v_{l'}$ for $1 \leq l, l' \leq q$, we can use the triangle inequality to establish:

Corollary 6.2 (Convergence of variance)**.** *Assume the conditions of Theorem 6.1. Then*

$$n\big\|\mathrm{Var}[f(\Phi\mathcal{X})] - \mathrm{Var}[f(\mathbf{Z}_1,\dots,\mathbf{Z}_n)]\big\| \;\leq\; 6n^2k^{3/2}(\alpha_0\alpha_3 + \alpha_1\alpha_2)(c_X + c_Z)\;.$$

Note that similar derivation can be made for many statistics of $f(\Phi\mathcal{X})$ such as the expectation. To compare the distributions on $\mathbb{R}^q$, we use all functions h in a suitable class $\mathcal{H}$ of test functions. In the context of the noise stability definitions above, we choose

$$\mathcal{H} \coloneqq \{h : \mathbb{R}^q \to \mathbb{R} \mid h \text{ is thrice-differentiable with } \gamma_1(h), \gamma_2(h), \gamma_3(h) \leq 1\}\;.$$

The distributions of two random elements $\mathbf{X}$ and $\mathbf{Y}$ of $\mathbb{R}^q$ are then compared by defining

$$d_{\mathcal{H}}(\mathbf{X},\mathbf{Y}) \coloneqq \sup_{h\in\mathcal{H}} |\mathbb{E}h(\mathbf{X}) - \mathbb{E}h(\mathbf{Y})|\;,$$

that is, the integral probability metric determined by $\mathcal{H}$. We note that it metrises weak convergence.

Lemma 6.3 ($d_{\mathcal{H}}$ metrises weak convergence)**.** *Let $\mathbf{Y}$ and $\mathbf{Y}_1, \mathbf{Y}_2, \dots$ be random variables in $\mathbb{R}^q$ with $q \in \mathbb{N}$ fixed. Then $d_{\mathcal{H}}(\mathbf{Y}_n, \mathbf{Y}) \to 0$ implies weak convergence $\mathbf{Y}_n \xrightarrow{d} \mathbf{Y}$.*

This metric is similar to the generalised Dudley distance of Grigorevskii and Shiganov (1976), but unlike the latter, $d_{\mathcal{H}}$ controls all three derivatives simultaneously. For a comparison of $d_{\mathcal{H}}$ to other probability metrics, see Appendix D.3.1. Since $\mathcal{H}$ is a subset of $\mathcal{F}_3(\mathbb{R}^q, \mathbb{R})$, replacing f with $\sqrt{n}f$ in Theorem 6.1 yields:

Corollary 6.4 (Convergence in $d_{\mathcal{H}}$)**.** *Under the conditions of Theorem 6.1,*

$$d_{\mathcal{H}}(\sqrt{n}f(\Phi\mathcal{X}), \sqrt{n}f(\mathbf{Z}_1,\dots,\mathbf{Z}_n)) \;\leq\; n^{3/2}k^{3/2}(n\alpha_1^3 + 3n^{1/2}\alpha_1\alpha_2 + \alpha_3)(c_X + c_Z)\;.$$

Thus, Theorem 6.1 exactly characterises the asymptotic variance and distribution of the augmented estimate $f(\Phi\mathcal{X})$ by showing universality of its distribution, as summarised in the next corollary. That allows us, for example, to compute consistent quantiles for $f(\Phi\mathcal{X})$.

Corollary 6.5. *Fix q. Assume the conditions of Theorem 6.1 hold, and that the bounds in Corollary 6.2 and 6.4 converge to zero as $n, k \to \infty$. Then*

$$d_{\mathcal{H}}(\sqrt{n}f(\Phi\mathcal{X}), \sqrt{n}f(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)) \ \to 0 \ ,$$
$$n\big\|\mathrm{Var}[f(\Phi\mathcal{X})] - \mathrm{Var}[f(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)]\big\| \ \to 0 \ .$$

The next lemma simplifies notation throughout—it shows that, if the scaling by $\sqrt{n}$ be dropped, one can still study convergence of both $\mathbb{E}[f(\Phi\mathcal{X})]$ and of the centred estimate. Results can hence be stated without explicitly centering terms.

Lemma 6.6. *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be random variables in* $\mathbb{R}^q$. *Suppose* $d_{\mathcal{H}}(\mathbf{X}, \mathbf{Y}) \leq \epsilon$ *for some constant* $\epsilon > 0$. *Then* $\|\mathbb{E}\mathbf{X} - \mathbb{E}\mathbf{Y}\| \leq q^{1/2}\epsilon$ *and* $d_{\mathcal{H}}(\mathbf{X} - \mathbb{E}\mathbf{X}, \mathbf{Y} - \mathbb{E}\mathbf{Y}) \leq (1 + q^{1/2})\epsilon$.

Remark 6.1 (Comments on the main theorem)**.** *(i) Gaussian surrogates*. In most of our examples, the data domain $\mathcal{D}$ is the entire space $\mathbb{R}^d$. If so, one may choose the $\mathbf{Z}_i$ as Gaussian vectors matching the first two moments of $\Phi_1\mathbf{X}_1$.

(ii) Generalisations. The proof techniques still apply if some conditions are relaxed. Generalised results are given in Appendix A, and appear in some of the applications we study below. For example, $\mathbf{Z}_i$ may be matrix-valued (e.g. in ridge regression, in Proposition 6.8). The range and domain of ϕ_{ij} may not agree (Theorem 13), and the ϕ_{ij} do not have to be i.i.d. We may also permit q to grow with n and k.

(iii) Distributional invariance. A common assumption in machine learning is that the data distribution is invariant under $\mathcal{T}$. That means that, for all $\phi \in \mathcal{T}$,

$$\phi\mathbf{X}_1 \stackrel{d}{=} \mathbf{X}_1 \qquad \text{or equivalently} \qquad \mathbb{E}[f(\mathbf{X}_1)] \ = \ \mathbb{E}[f(\phi\mathbf{X}_1)] \quad \text{for all } f \in \mathbf{L}_1(\mathbf{X}_1) \ .$$

From a statistical learning perspective, this is one way to ensure that augmentation does not alter the limiting estimator, although the speed of convergence to that limit may differ. In light of Theorem 6.1, invariance implies that the variance in (6.1) can be replaced by

$$\mathrm{Var}\mathbf{Z}_i = \mathbf{I}_k \otimes \mathbb{E}[\mathrm{Var}[\phi_{11}\mathbf{X}_1|\phi_{11}]] + (\mathbf{1}_{k\times k} - \mathbf{I}_k) \otimes \mathbb{E}[\mathrm{Cov}[\phi_{11}\mathbf{X}_1, \phi_{12}\mathbf{X}_1|\phi_{11}, \phi_{12}]] \ .$$

Note the off-diagonal terms are now covariance matrices that are smaller than those in (6.1) in the Loewner partial order.

(iv) *Different number of augmentations per data point.* At the cost of more cumbersome notation, our universality approximation can be extended to the case where the number of augmentations for the i-th data point, k_i, differs across $i \leq n$. To achieve this, we may first identify $k = \max_i k_i$, which allows us to write each i-th augmented data block as a size-$\mathbb{R}^{kd}$ vector by padding $(k - k_i)d$ many zeros. The problem is reduced to a Gaussian universality approximation for functions of n independent $\mathbb{R}^{kd}$ random vectors. This can be proved by Lindeberg's technique as was done in Chapter 4 for polynomial functions: the relaxation of i.i.d. assumption to independent assumption only results in a more cumbersome moment bound analogous to that of Theorem 4.1.

In conclusion, if the conditions of Theorem 6.1 hold and the bounds in Corollaries 6.2 and 6.4 converge to zero, then the asymptotic distribution of $\sqrt{n}f(\Phi\mathcal{X})$ only depends on the mean and covariance of the augmented samples $\Phi\mathcal{X}$. Hence, under general conditions, the effect of data augmentation on the learning rate only depends on how it affects the first few moments of the augmented variables, e.g. how strong the correlation between the augmented samples is. This universality greatly simplifies the asymptotic analysis of data augmentation.

6.4 Variance reduction and variance inflation

In this section, we consider estimators of the form

$$f(\mathbf{x}_{11}, \dots, \mathbf{x}_{nk}) = g\big(\tfrac{1}{nk} \textstyle\sum_{i \le n, j \le k} \mathbf{x}_{ij}\big) \tag{6.5}$$

for a smooth function g. The simplest is an empirical average, which we analyse first and which exhibit the known variance reduction effect in the literature (e.g. Chen et al. (2020)). The results we obtain for such averages still hold if f is approximately linear, in the sense that it can be approximated well by a first-order Taylor expansion. The risk of an M-estimator in fixed dimensions is an example. The behaviour changes, however, when f is non-linear. We illustrate this by a toy example in Section 6.4.4 and concretely in the example of ridge regression in moderate dimensions (Section 6.4.5).

6.4.1. Comparing limiting variances

A natural measure of the effect of data augmentation on the convergence rate is the variance ratio comparing estimates obtained with and without augmentation. To define a valid baseline for estimates without augmentation, we must replicate each input vector k times, since the number k of augmentations determines the number of arguments of f, and also enters in the upper bound. We denote such k-fold replicates by $\tilde{\mathbf{X}}_i := (\mathbf{X}_i, \dots, \mathbf{X}_i) \in \mathcal{D}^k$. No augmentation then corresponds to the case where $\mathcal{T}$ contains only the identity map of $\mathcal{D}^{nk}$. By setting each ϕ_{ij} to identity in Theorem 6.1, we can approximate the distribution of $f(\tilde{\mathbf{X}}_1, \dots, \tilde{\mathbf{X}}_n)$ by that of $f(\tilde{\mathbf{Z}}_1, \dots, \tilde{\mathbf{Z}}_n)$, where $\tilde{\mathbf{Z}}_1, \dots, \tilde{\mathbf{Z}}_n$ are any i.i.d. variables in $\mathcal{D}^k$ satisfying

$$\mathbb{E}\mathbf{Z}_i \;=\; \mathbf{1}_{k\times 1} \otimes \mathbb{E}\mathbf{X}_1 \qquad \text{and} \qquad \mathrm{Var}\mathbf{Z}_i \;=\; \mathbf{1}_{k\times k} \otimes \mathrm{Var}\mathbf{X}_1 \;, \tag{6.6}$$

and substituting into Theorem 6.1 shows

$$\big|\mathbb{E}h(f(\tilde{\mathbf{X}}_1, \dots, \tilde{\mathbf{X}}_n)) - \mathbb{E}h(f(\tilde{\mathbf{Z}}_1, \dots, \tilde{\mathbf{Z}}_n))\big| \;\le\; nk^{3/2}\lambda_h(n,k)(c_{\tilde{X}} + c_{\tilde{Z}}) \;. \tag{6.7}$$

The effect of augmentation versus no augmentation can now be compared by the ratio

$$\vartheta(f) \coloneqq \sqrt{\|\mathrm{Var} f(\tilde{\mathbf{Z}}_1, \ldots, \tilde{\mathbf{Z}}_n)\| \,/\, \|\mathrm{Var} f(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)\|} \,. \tag{6.8}$$

If $\vartheta(f) > 1$, augmentation is beneficial in the sense that it speeds up convergence of the estimator (though it may or may not introduce a bias). If $\vartheta(f) < 1$, it is detrimental, which is possible even if invariance holds.

Notation. We write $\Phi\mathcal{X}$ for augmented data, and $\mathcal{Z} \coloneqq \{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$ for i.i.d. surrogates satisfying (6.1). $\tilde{\mathcal{X}} = (\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_n)$ denotes the unaugmented, replicated data defined above, and $\tilde{\mathcal{Z}} \coloneqq \{\tilde{\mathbf{Z}}_1, \ldots, \tilde{\mathbf{Z}}_n\}$ surrogates satisfying (6.6). We refer to $\mathcal{Z}$ and $\tilde{\mathcal{Z}}$ as Gaussian if $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ and $\tilde{\mathbf{Z}}_1, \ldots, \tilde{\mathbf{Z}}_n$ are Gaussian vectors in $\mathbb{R}^d$.

6.4.2. Empirical averages

The arguably most common choice of f is an empirical average—augmentation is often used with empirical risk minimisation, and the empirical risk is such an average. By Remark 6.1(ii) above, empirical estimates of gradients can also be represented as empirical averages. An augmented empirical average is of the form

$$f(\mathbf{x}_{11}, \ldots, \mathbf{x}_{nk}) \coloneqq \tfrac{1}{nk} \textstyle\sum_{i=1}^{n} \sum_{j=1}^{k} \mathbf{x}_{ij} \;, \tag{6.9}$$

where $\mathcal{D} = \mathbb{R}^d$, and d and k are fixed. Specializing Theorem 6.1 yields:

Proposition 6.7 (Augmenting averages)**.** *Require that* $\mathbb{E}\|\mathbf{X}_1\|^6$ *and* $\mathbb{E}\|\phi_{11}\mathbf{X}_1\|^6$ *are finite. Let* $\mathcal{Z}$ *and* $\tilde{\mathcal{Z}}$ *be Gaussian. Then* f *as above satisfies*

$$d_{\mathcal{H}}(\sqrt{n} f(\Phi\mathcal{X}), \sqrt{n} f(\mathcal{Z})) \to 0 \quad \textit{and} \quad d_{\mathcal{H}}(\sqrt{n} f(\tilde{\mathcal{X}}), \sqrt{n} f(\tilde{\mathcal{Z}})) \to 0 \quad \textit{as } n \to \infty \,.$$

The Gaussian surrogates can be translated into asymptotic quantiles as follows: The ratio ϑ of standard deviations here takes the form

$$\vartheta \;=\; \sqrt{\left(\tfrac{1}{n}\mathrm{Var}[\mathbf{X}_1]\right) \Big/ \left(\tfrac{1}{nk}\mathrm{Var}[\phi_{11}\mathbf{X}_1] + \tfrac{k-1}{nk}\mathrm{Cov}[\phi_{11}\mathbf{X}_1, \phi_{12}\mathbf{X}_1]\right)} \,.$$

To keep notation simple, assume $d = 1$. To obtain $\alpha/2$-th asymptotic quantiles, for $\alpha \in [0, 1]$, denote by $z_{\alpha/2}$ the $(1 - \alpha/2)$-percentile of a standard normal. Then the lower and upper asymptotic quantiles of $f(\Phi\mathcal{X})$ and $f(\tilde{\mathcal{X}})$ are given respectively by

$$\mathbb{E}[\phi_{11}\mathbf{X}_1] \pm \frac{1}{\sqrt{\vartheta^2 n}} z_{\alpha/2} \sqrt{\mathrm{Var}[\mathbf{X}_1]} \quad \text{and} \quad \mathbb{E}[\mathbf{X}_1] \pm \frac{1}{\sqrt{n}} z_{\alpha/2} \sqrt{\mathrm{Var}[\mathbf{X}_1]} \;.$$

For empirical averages, the quantiles can be inverted to obtain asymptotic $(1 - \alpha)$-confidence intervals for $\mathbb{E}[\phi_{11}\mathbf{X}_1]$ and $\mathbb{E}[\mathbf{X}_1]$, given by

$$\left[f(\Phi\mathcal{X}) \pm \frac{1}{\sqrt{\vartheta^2 n}} z_{\alpha/2} \sqrt{\mathrm{Var}[\mathbf{X}_1]} \right] \quad \text{and} \quad \left[f(\tilde{\mathcal{X}}) \pm \frac{1}{\sqrt{n}} z_{\alpha/2} \sqrt{\mathrm{Var}[\mathbf{X}_1]} \right]$$

Remark 6.2. We note some implications of Proposition 6.7:

(i) In terms of confidence region width, computing the empirical average by augmenting n observations is equivalent to averaging over an unaugmented data set of size $\vartheta^2 n$.
(ii) Augmentation is hence beneficial for empirical averages if $\|\text{Var}[\phi_{11}\mathbf{X}_1]\| \leq \|\text{Var}\,\mathbf{X}_1\|$. To see this, observe that augmentation is beneficial if $\vartheta \geq 1$, and that

$$\|\text{Var} f(\mathcal{Z})\| = \|\frac{1}{k}\text{Var}[\phi_{11}\mathbf{X}_1] + \frac{k-1}{k}\text{Cov}[\phi_{11}\mathbf{X}_1, \phi_{12}\mathbf{X}_1]\| \leq \|\text{Var}[\phi_{11}\mathbf{X}_1]\| \,. \quad (6.10)$$

(iii) If the data distribution is invariant, in the sense that $\phi_{11}\mathbf{X}_1 \stackrel{d}{=} \mathbf{X}_1$, augmentation is always beneficial, since $\text{Var}\mathbf{X}_1 = \text{Var}[\phi_{11}\mathbf{X}_1] \succeq \text{Cov}[\phi_{11}\mathbf{X}_1, \phi_{12}\mathbf{X}_1]$.

6.4.3. Parametric plug-in estimators

Most of the observations for empirical averages still hold for plug-in estimators if the dimension is fixed, and more generally for any approximately linear function of averages, such as the risk of an M-estimator. To see this, note that if we choose g in (6.5) as a sufficiently smooth function, f can be approximated by a first-order Taylor expansion

$$f^T(\mathbf{x}_{11}, \ldots, \mathbf{x}_{nk}) \coloneqq g(\mathbb{E}[\phi_{11}\mathbf{X}_1]) + \partial g(\mathbb{E}[\phi_{11}\mathbf{X}_1])\big(\frac{1}{nk}\sum_{i\leq n, j\leq k}\mathbf{x}_{ij} - \mathbb{E}[\phi_{11}\mathbf{X}_1]\big) \,. \quad (6.11)$$

The key observation is that the only random contribution to f^T behaves exactly like an empirical average. Lemma 19 in the appendix shows that

$$d_{\mathcal{H}}(\sqrt{n}f(\Phi\mathcal{X}), \sqrt{n}f^T(\mathcal{Z})) \to 0 \qquad \text{and} \quad n\left(\|\text{Var}[f(\Phi\mathcal{X})]\| - \|\text{Var}[f^T(\mathcal{Z})]\|\right) \to 0 \,, \quad (6.12)$$

provided that g is sufficiently well-behaved and noise stability holds. That is even true if d grows (not too rapidly) with n.

The variance of f^T now depends additionally on $\partial g(\mathbb{E}[\phi_{11}\mathbf{X}_1])$. If the data distribution is not invariant under augmentation, it is possible that $\|\partial g(\mathbb{E}[\phi_{11}\mathbf{X}_1])\| > \|\partial g(\mathbb{E}\mathbf{X}_1)\|$. If so, the overall variance may increase even if augmentation decreases the variance of the empirical average. If invariance holds, augmentation reduces variance, as observed by Chen et al. (2020).

6.4.4. Non-linear estimators

We have seen above that, in the linear case, invariance guarantees that augmentation does not increase estimator variance. If the estimator (6.5) is not well-approximated by the linearisation (6.11), that need not be true, which can be seen as follows. Theorem 6.1

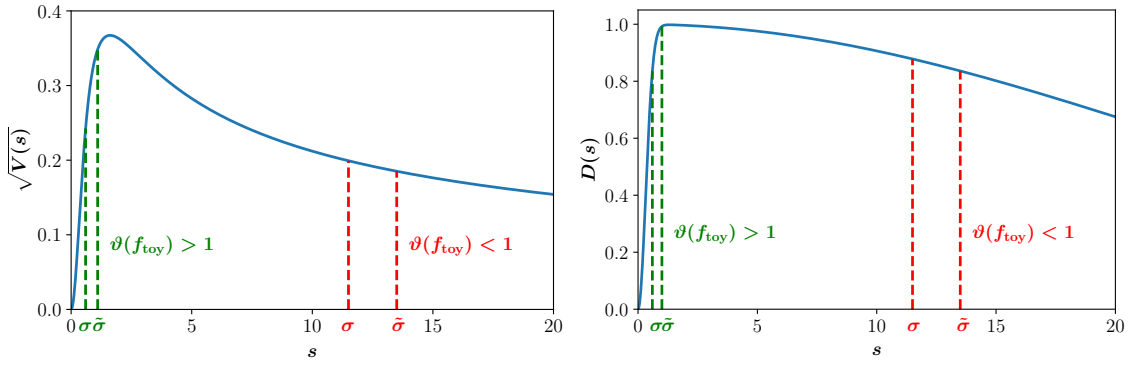


Figure 6.3: *Left*: The standard deviation $\sqrt{V(s)} := \sqrt{\text{Var}[f_{\text{toy}}(\mathcal{Z})]} = \sqrt{\text{Var}[g_{\text{toy}}(s\xi + \mathbb{E}[\mathbf{X}_1])]}$ as a function of s. *Right*: The difference $D(s)$ between the 0.025-th and the 0.975-th quantiles for $g_{\text{toy}}(s\xi + \mathbb{E}[\mathbf{X}_1])$ as a function of s. The functions are calculated analytically in Proposition 20. Since neither is monotonic, the parameter space contains regions where data augmentation is beneficial (green example), and where it is detrimental (red example). Notably, $\vartheta(f) < 1$ is possible even if σ, standard deviation of the augmented average, is smaller than $\tilde{\sigma}$, standard deviation of the unaugmented average.

shows that

$$\text{Var}[f(\Phi\mathcal{X})] \;\approx\; \text{Var}\Big[g\Big(\frac{\sqrt{\text{Var}[\mathbf{X}_1]}}{\sqrt{\vartheta^2 n}}\xi \;+\; \mathbb{E}[\phi_{11}\mathbf{X}_1]\Big)\Big] \quad \text{for } \xi \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)\ .$$

The same holds, with $\vartheta = 1$, for the unaugmented variance. Assume for simplicity that $d = 1$ and invariance holds, which implies $\mathbb{E}[\phi_{11}\mathbf{X}_1] = \mathbb{E}[\mathbf{X}_1]$ and $\vartheta \geq 1$. By a well-known result characterizing the variance of a function of a Gaussian (Proposition 3.1 of Cacoullos (1982)), we have

$$\sigma^2\mathbb{E}\big[\partial g(\sigma\xi + \mathbb{E}[\mathbf{X}_1])\big]^2 \leq \text{Var}\big[g\big(\sigma\xi + \mathbb{E}[\mathbf{X}_1]\big)\big] \leq \sigma^2\mathbb{E}\big[\partial g(\sigma\xi + \mathbb{E}[\mathbf{X}_1])^2\big]$$

for any $\sigma > 0$. When g is non-linear, ∂g is not constant, and $\text{Var}\big[g\big(\sigma\xi + \mathbb{E}[\mathbf{X}_1]\big)\big]$ is not necessarily monotonic in σ. Thus, in the non-linear case, invariance of the data distribution does not imply variance reduction. Figure 6.3 illustrates the variance and quantiles for a highly non-linear toy statistic, defined as

$$f_{\text{toy}}(x_{11}, \ldots, x_{nk}) \;:=\; g_{\text{toy}}\Big(\frac{1}{nk}\sum_{ij} x_{ij}\Big) \;=\; \exp\Big(-\Big(\frac{1}{\sqrt{nk}}\sum_{ij} x_{ij}\Big)^2\Big)\ . \tag{6.13}$$

In both plots of Figure 6.3, the behaviour of augmentation changes from one region of parameter space to another. See Appendix D.2.1 for formal statements and simulation results.

6.4.5. Ridge regression

This section studies the effect of augmentation on ridge regression in moderate dimensions. In light of the discussion in the previous section, this is an example of an estimator that is not approximately linear, which complicates the effect of augmentation on its variance.

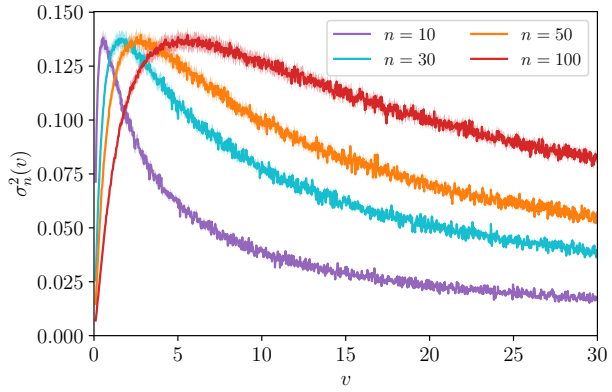


Figure 6.4: A simple ridge regression example, where variance of the risk is not monotonic in data variance despite invariance. Variance of R^Z in Lemma 6.9 is plotted as a function of the augmented covariance $\nu := \mathrm{Cov}[(\pi_{11}\mathbf{V}_1)^2, (\pi_{12}\mathbf{V}_1)^2]$ for $\lambda = 0.1$ and $\mathbb{E}[\mathbf{V}_1^2] = 0.1$. As no closed-form formula is available, the plot ie generated by a simulation over 10k random seeds.

In a regression problem, each data point $\mathbf{X}_i := (\mathbf{V}_i, \mathbf{Y}_i)$ consists of a covariate $\mathbf{V}_i$ with values in $\mathbb{R}^d$, and a response $\mathbf{Y}_i$ in $\mathbb{R}^b$. We hence consider pairs of transformations (π_{ij}, τ_{ij}) as augmentation, where π_{ij} acts on $\mathbf{V}_i$ and τ_{ij} acts on $\mathbf{Y}_i$. A transformed data point is then of the form $\phi_{ij}\mathbf{x}_i := ((\pi_{ij}\mathbf{v}_i)(\pi_{ij}\mathbf{v}_i)^\top, (\pi_{ij}\mathbf{v}_i)(\tau_{ij}\mathbf{y}_i)^\top)$, and hence an element of $\mathcal{D} := \mathbb{M}^d \times \mathbb{R}^{d\times b}$, where $\mathbb{M}^d$ denotes the set of positive semi-definite $d \times d$ matrices. For a fixed $\lambda > 0$, the ridge regression estimator on augmented data is therefore

$$\hat{B}(\phi_{11}\mathbf{x}_1, \ldots, \phi_{nk}\mathbf{x}_n) \;:=\; \Big(\frac{1}{nk}\sum\nolimits_{ij}(\pi_{ij}\mathbf{v}_i)(\pi_{ij}\mathbf{v}_i)^\top + \lambda\mathbf{I}_d\Big)^{-1}\frac{1}{nk}\sum\nolimits_{ij}(\pi_{ij}\mathbf{v}_i)(\tau_{ij}\mathbf{y}_i)^\top\,. \tag{6.14}$$

It takes values in $\mathbb{R}^{d\times b}$, and its risk is $R(\hat{B}) := \mathbb{E}[\|\mathbf{Y}_{new} - \hat{B}^\top\mathbf{V}_{new}\|_2^2 \mid \hat{B}]$.

The next result shows universality of the asymptotic distribution of the risk of a ridge estimator in a moderate-dimensional regime, for any choice of augmentation. In particular, one can study the effect of augmentation on the variance of the risk, which measures the speed of convergence of the risk to its infinite-data limit.

Proposition 6.8. *Suppose* $\max_{l\le d}\max\{(\pi_{11}\mathbf{V}_1)_l, (\tau_{11}\mathbf{Y}_1)_l\}$ *is almost surely bounded by* $Cd^{-1/2}$ *for some absolute constant* $C > 0$ *and that* $b = O(d)$*. Then there exist i.i.d. surrogate variables* $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ *such that*

$$d_{\mathcal{H}}(\sqrt{n}R^{\Phi\mathcal{X}}, \sqrt{n}R^Z) = O(n^{-1/2}d^9) \quad \textit{and} \quad n(\mathrm{Var}[R^{\Phi\mathcal{X}}] - \mathrm{Var}[R^Z]) = O(n^{-1}d^7)\,,$$

where $R^{\Phi\mathcal{X}} := R(\hat{B}(\Phi\mathcal{X}))$ *is the risk of the estimator trained on augmented data, and* $R^Z := R(\hat{B}(\mathcal{Z}))$ *the risk with surrogate variables.*

In this case, the surrogate variables $\mathbf{Z}_i$ are random elements of $(\mathbb{M}^d \times \mathbb{R}^{d\times b})^k$, whose first two moments match those of the augmented data. As part of the proof of the proposition, we also obtain convergence rates for the estimator $\hat{B}(\Phi\mathcal{X})$ (in addition to the rate

for its risk above); see Lemma D.32 in the appendix.

A detailed analysis of a simple illustrative example. We consider a special case in more detail, which illustrates that unexpected effects of augmentation can occur even in very simple models: Assume that

$$\mathbf{Y}_i \coloneqq \mathbf{V}_i + \varepsilon_i \quad \text{where} \quad \mathbf{V}_i \overset{i.i.d.}{\sim} \mathcal{N}(\mu \mathbf{1}_d, \Sigma) \quad \text{and} \quad \varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, c^2 \mathbf{I}_d) \; . \tag{6.15}$$

This is the setup used in Figure 6.1, where $d = 2$. Detrimental effects of augmentation can occur even in one dimension, though. To clarify that, we first show the following:

Lemma 6.9. *Consider the one-dimensional case ($d = 1$), with $c = 0$ and $\tau_{ij} = \pi_{ij}$. Assume that the augmentation leaves the covariate distribution invariant, $\pi_{ij}\mathbf{V}_i \overset{d}{=} \mathbf{V}_i$. Write the covariance $v_\pi = \mathrm{Cov}[(\pi_{11}\mathbf{V}_1)^2, (\pi_{12}\mathbf{V}_1)^2]$, and generate surrogate variables by drawing*

$$\mathbf{Z}_{111}, \ldots, \mathbf{Z}_{n11} \overset{i.i.d.}{\sim} \Gamma\Big(\frac{\mathbb{E}[\mathbf{V}_1^2]^2}{v_\pi}, \frac{\mathbb{E}[\mathbf{V}_1^2]}{v_\pi}\Big)$$

and setting $\mathbf{Z}_{ijl} \coloneqq \mathbf{Z}_{i11}$, for all $j \leq k$ and $l = 1, 2$. Then

$$d_{\mathcal{H}}(\sqrt{n}R^{\Phi\mathcal{X}}, \sqrt{n}R^{Z}) \to 0 \quad \textit{and} \quad n(\mathrm{Var}[R^{\Phi\mathcal{X}}] - \mathrm{Var}[R^{Z}]) \to 0 \qquad \textit{as } n, k \to \infty \; .$$

Moreover, denoting the Gamma random variable $X_n(v) \sim \Gamma(\frac{n\mathbb{E}[\mathbf{V}_1^2]^2}{v}, \frac{n\mathbb{E}[\mathbf{V}_1^2]}{v})$, we have

$$\mathrm{Var}[R^Z] \;=\; \sigma_n^2(v_\pi) \;=\; \mathbb{E}[\mathbf{V}_1^2]^2 \lambda^2 \mathrm{Var}\Big[\frac{1}{(X_n(v_\pi) + \lambda)^2}\Big] \; ,$$

where σ_n is a real-valued function that does not depend on the number of augmentations k, or on the law of the augmentations π_{ij}.

Note the surrogate distribution can be determined explicitly, and is non-Gaussian. The main object of interest is the variance σ_n^2 of the risk of an augmented ridge regressor. For any choice of augmentation, the augmented covariance ν_π is always bounded from above by the unaugmented variance $\mathrm{Var}[(\mathbf{V}_1)^2]$. This does not generally imply the the augmented ridge regressor is a better estimator—the simulation in Figure 6.4 shows that σ_n is non-monotonic, that is, even though augmentation reduces ν_π, it may increase the variance of the risk.

Remark 6.3 (Details on simulations)**.** (i) The simulation in Figure 6.5 uses the model (6.15) and two forms of augmentation are both adapted from image analysis:

(a) Random rotations. We represent the elements of the size-d cyclic group by matrices $C_1, \ldots, C_d$, generate random transformations

$$\phi_{ij} \;=\; \pi_{ij} \overset{i.i.d.}{\sim} \mathrm{Uniform}\{C_1, \ldots, C_d\} \; ,$$

and set $\phi_{ij}\mathbf{x}_i \coloneqq ((\pi_{ij}\mathbf{v}_i)(\pi_{ij}\mathbf{v}_i)^\top, (\pi_{ij}\mathbf{v}_i)(\tau_{ij}\mathbf{y}_i)^\top)$, i.e. we cycle through the d coordinates of $\mathbf{Y}_i$ and $\mathbf{V}_i$ simultaneously. The invariance $(\phi_{11}\mathbf{V}_1, \phi_{11}\mathbf{Y}_1) \overset{d}{=} (\mathbf{V}_1, \mathbf{Y}_1)$ holds.

(b) Random cropping for $d = 2$, where a uniformly chosen coordinate of both $\mathbf{Y}_i$ and $\mathbf{V}_i$

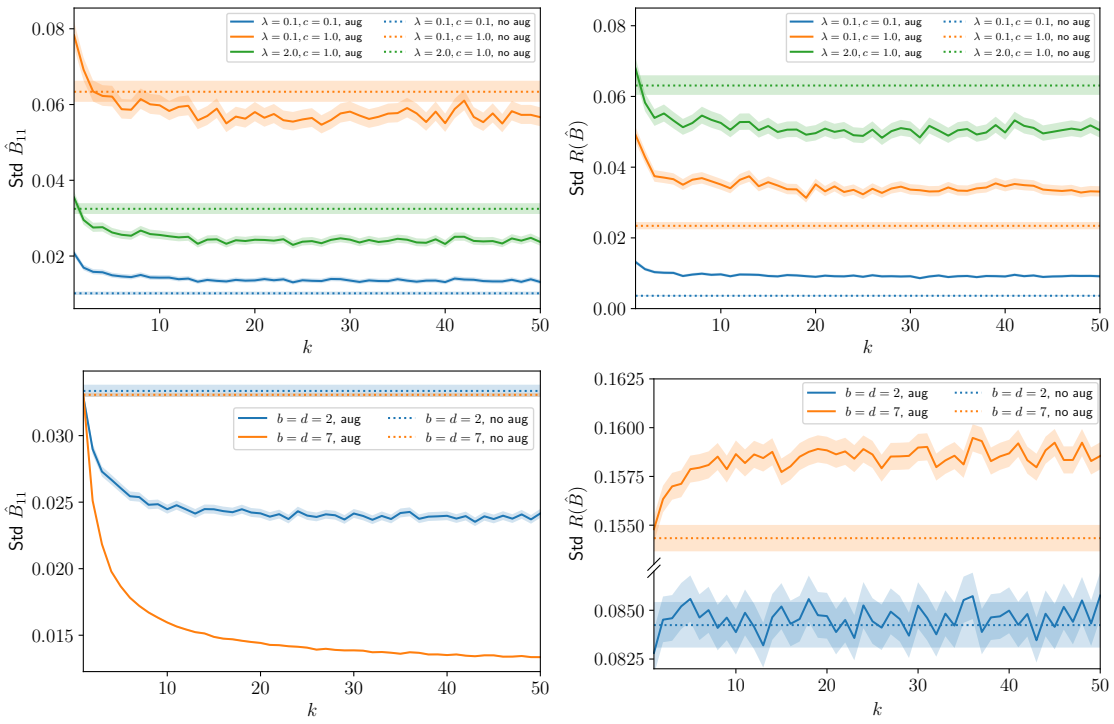


Figure 6.5: Augmentation can decrease the variance of an estimator, but at the same time increase the variance of its risk: Shown are simulations for ridge regression under (6.15) with $\mu = 0$ and varying k. The augmentations on each pair of $\mathbf{V}_{ij}$ and $\mathbf{Y}_{ij}$ are set to be the same, i.e. $\pi_{ij} = \tau_{ij}$. For random cropping, $n = 200$ and $\Sigma = \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$. For uniform rotations, $n = 50$ and $\Sigma = \mathbf{I}_d$, $c = 2$, $\lambda = 9$. *Top Left.* Standard deviation of $(\hat{B}(\Phi\mathcal{X}))_{11}$, first coordinate of ridge regression estimate under random cropping. *Top Right.* Standard deviation of $R(\hat{B}(\Phi\mathcal{X}))$ under random cropping. *Bottom Left.* Std $(\hat{B}(\Phi\mathcal{X}))_{11}$ under uniform rotations. *Bottom Right.* Std $R(\hat{B}(\Phi\mathcal{X}))$ under uniform rotations.

is set to 0, i.e. we have

$$\phi_{ij} \;=\; \pi_{ij} \overset{i.i.d.}{\sim} \text{Uniform}\{C_1 M, \ldots, C_d M\} \quad \text{where } M \;:=\; \begin{pmatrix} 0 & & \\ & 1 & \\ & & \ddots \\ & & & 1 \end{pmatrix}.$$

(ii) We can now specify the setting used in Figure 6.1 in the introduction: It shows the empirical average function and the ridge regression estimate computed on the random cropping setup in Figure 6.5, for $k = 50$ and $\lambda = c = 0.1$.

6.5 Non-regularisation in high-dimensional linear regression

We next consider the effect of data augmentation on the limiting risk of a ridgeless regressor in high dimensions. Without augmentation, such regressors are known to exhibit a double-descent phenomenon (Hastie et al., 2022). We show that augmentations can shift the double-descent peak of the risk curve, depending on the number of augmentations (see Figure 6.2 in the introduction). Such a shift has been observed empirically by

Dhifallah and Lu (2021).

Specifically, we consider the linear model where the univariate response variable Y_i is related to the covariate $\mathbf{V}_i$ in $\mathbb{R}^d$ by

$$Y_i \;=\; \mathbf{V}_i^\top \beta + \epsilon_i \quad \text{for } i = 1, \ldots, n\,, \tag{6.16}$$

where the variables $\mathbf{V}_i$ are i.i.d. mean-zero random (not necessarily Gaussian) vectors, and the noise variables ϵ_i are i.i.d. mean-zero with $\mathrm{Var}[\epsilon_i] = \sigma_\epsilon^2$ and a bounded fourth moment. The dimension d grows linearly with n, and the signal β and noise variance are assumed non-random with $\|\beta\| = \Theta(1)$ and $\sigma_\epsilon^2 = \Theta(1)$. Following standard assumptions in random matrix theory, we assume that $\mathbf{V}_i$ has independent coordinates $(V_{il})_{l \le d}$. For simplicity, we assume that $\mathbb{E}[V_{il}^3] = 0$ and $\mathbb{E}[V_{il}^4] = 3\mathrm{Var}[V_{il}]^2$, i.e. the first four moments of V_{il} matches those of its Gaussian surrogate; a similar assumption was used in Tao and Vu (2011) for applying the Lindeberg method to obtain universality of eigenvalue statistics of large matrices. We expect that the fourth moment condition can be replaced by a sub-exponential tail in view of known results on universality of covariance matrices, but this may require additional proof techniques involving the Dyson Brownian motion (see e.g. Theorem 5.1 and the subsequent discussion of Pillai and Yin (2014)) and we do not pursue it here.

6.5.1. Double descent shift under oracle augmentation

We first consider an oracle setup, where β is assumed known. This is a theoretical device, but we will see that it is informative. The setup is motivated by the fact that, once we have chosen transformations π_{ij} to augment the covariates $\mathbf{V}_i$, we must also specify a reasonable way to augment the responses Y_i. Since the covariates and responses are related via β, a known value of β allows us to "pass" transformations from the covariates to the responses according to the model, by defining

$$\tau_{ij}^{(\mathrm{ora})} Y_i \;:=\; Y_i + \big(\pi_{ij}\mathbf{V}_i - \mathbf{V}_i\big)^\top \beta \;=\; (\pi_{ij}\mathbf{V}_i)^\top \beta + \epsilon_i\,.$$

If invariance holds for the covariates, it extends to responses,

$$\pi_{ij}\mathbf{V}_i \overset{d}{=} \mathbf{V}_i \qquad \Longleftrightarrow \qquad (\pi_{ij}\mathbf{V}_i, \tau_{ij}^{(\mathrm{ora})}\mathbf{Y}_i) \overset{d}{=} (\mathbf{V}_i, \mathbf{Y}_i)\,. \tag{6.17}$$

The augmented estimator is then

$$\hat\beta_\lambda^{(\mathrm{ora})} \;:=\; \Big(\frac{1}{nk}\textstyle\sum_{ij}(\pi_{ij}\mathbf{V}_i)(\pi_{ij}\mathbf{V}_i)^\top + \lambda \mathbf{I}_d\Big)^\dagger \frac{1}{nk}\textstyle\sum_{ij}(\pi_{ij}\mathbf{V}_i)\,\tau_{ij}^{(\mathrm{ora})}\mathbf{Y}_i\,. \tag{6.18}$$

This is a ridge estimator for $\lambda > 0$, and ridgeless for $\lambda = 0$. Following Hastie et al. (2022), we study the risk

$$\hat L_\lambda^{(\mathrm{ora})} \;:=\; \mathbb{E}\big[\big((\hat\beta_\lambda^{(\mathrm{ora})} - \beta)^\top \mathbf{V}_{\mathrm{new}}\big)^2 \,\big|\, \mathcal{X}\big] \qquad \text{for } \lambda \ge 0\,. \tag{6.19}$$

in the asymptotic regime where

$$n, d \to \infty\,, \quad d/n \to \gamma \in [0, \infty)\,, \quad d/(kn) \to \gamma' \in [0, \infty)\,, \quad k = o(n^{1/4})\,, \tag{6.20}$$

and k is allowed to be fixed or grow with n. In the unaugmented case, $\hat{\beta}_\lambda^{(\mathrm{ora})}$ and $\hat{L}_\lambda^{(\mathrm{ora})}$ are precisely the quantities studied by Hastie et al. (2022), who show that for $\lambda = 0$, the risk reproduces the double-descent phenomenon also observed in neural networks.

To illustrate the effect of augmentations in a simple model, we focus on the augmentation

$$\pi_{ij}\mathbf{V}_i \;\coloneqq \mathbf{V}_i + \xi_{ij}\;, \tag{6.21}$$

where $(\xi_{ij})_{i,j}$ is a set of i.i.d. mean-zero noise vectors, each having independent coordinates $(\xi_{ijl})_{l\le d}$ with $\mathbb{E}[\xi_{ijl}^3] = 0$ and $\mathbb{E}[\xi_{ijl}^4] = 3\mathrm{Var}[\xi_{ijl}]^2$. This form of randomization is also known as *noise injection* in other contexts.

The main challenge in analyzing the risk is that the augmented risk depends on two strongly correlated high-dimensional sample covariance matrices,

$$\bar{\mathbf{X}}_1 \coloneqq \frac{1}{nk}\sum_{i\le n}\sum_{j\le k}(\pi_{ij}\mathbf{V}_i)(\pi_{ij}\mathbf{V}_i)^\top\,, \quad \bar{\mathbf{X}}_2 \coloneqq \frac{1}{n}\sum_{i\le n}\left(\frac{1}{k}\sum_{j\le k}(\pi_{ij}\mathbf{V}_i)\right)\left(\frac{1}{k}\sum_{j\le k}(\pi_{il}\mathbf{V}_i)\right)^\top.$$

For comparison, $\bar{\mathbf{X}}_1 = \bar{\mathbf{X}}_2$ in the unaugmented case, and therefore existing analysis of double descent only involves one such matrix (e.g. Hastie et al. (2022)). To address this, we consider the Gaussian surrogate vectors $\mathbf{Z}_i$'s, where

$$\mathbb{E}[\mathbf{Z}_i] \;= \mathbb{E}[\pi_{ij}\mathbf{V}_i] \qquad \text{and} \qquad \mathrm{Var}[\mathbf{Z}_i] \;= \mathrm{Var}[\pi_{ij}\mathbf{V}_i]\,.$$

We denote the corresponding sample covariance matrices by

$$\bar{\mathbf{Z}}_1 \;\coloneqq\; \frac{1}{nk}\textstyle\sum_{i=1}^n\sum_{j=1}^k \mathbf{Z}_{ij}\mathbf{Z}_{ij}^\top\,, \qquad \bar{\mathbf{Z}}_2 \;\coloneqq\; \frac{1}{n}\textstyle\sum_{i=1}^n\left(\frac{1}{k}\sum_{j=1}^k \mathbf{Z}_{ij}\right)\left(\frac{1}{k}\sum_{l=1}^k \mathbf{Z}_{ij}\right)^\top.$$

Applying Theorem 6.1 allows us to approximate $\bar{\mathbf{X}}_1$ and $\bar{\mathbf{X}}_2$ by $\bar{\mathbf{Z}}_1$ and $\bar{\mathbf{Z}}_2$, whose spectral distributions are in the universality regime of compound Marchenko-Pastur laws Marchenko and Pastur (1967). This can be used to investigate the limiting risk. The universality result requires several regularity assumptions, which we state next.

Assumption 6.1. There exists some absolute constant $c_0 > 0$ such that for all $n, k, d \in \mathbb{N}$, the following quantities are bounded from above by c_0:

$$\max_{i\le n, j\le k, l\le d}\|X_{ijl}\|_{L_{10}}\,, \quad \big\|\|\bar{\mathbf{X}}_2\|_{op}\big\|_{L_{60}}\,, \quad \big\|\|\bar{\mathbf{Z}}_2\|_{op}\big\|_{L_{60}}\,.$$

Remark. Note that the use of L_{60} norm arises from a crude Cauchy-Schwarz bound, and we expect this to be improvable.

Assumption 6.2. The following quantities are $O_{\gamma'}(1)$ with probability $1 - o_{\gamma'}(1)$:

$$\|\bar{\mathbf{X}}_1^\dagger\|_{op}\,, \quad \|\bar{\mathbf{Z}}_1^\dagger\|_{op}\,, \quad \|\bar{\mathbf{X}}_2\|_{op}\,, \quad \|\bar{\mathbf{Z}}_2\|_{op}\,,$$
$$\sum_{l=1}^{d} \mathbb{I}_{\{\lambda_l(\bar{\mathbf{X}}_1)=0\}}\big(v_l(\bar{\mathbf{X}}_1)^\top \bar{\mathbf{X}}_2\, v_l(\bar{\mathbf{X}}_1)\big)\,, \quad \sum_{l=1}^{d} \mathbb{I}_{\{\lambda_l(\bar{\mathbf{Z}}_1)=0\}}\big(v_l(\bar{\mathbf{Z}}_1)^\top \bar{\mathbf{Z}}_2\, v_l(\bar{\mathbf{Z}}_1)\big)\,,$$

where $(\lambda_l(A), v_l(A))$ denotes the l-th eigenvalue-eigenvector pair of a matrix $A \in \mathbb{R}^{d\times d}$, and $O_{\gamma'}(\,\bullet\,)$ and $o_{\gamma'}(\,\bullet\,)$ indicate that the bounding constants are allowed to depend on γ'.

Proposition 6.10. *Fix $\lambda > 0$ and assume Assumption 6.1 holds. Then under the asymptotic regime (6.20), we have*

$$d_{\mathcal{H}}\big(f_\lambda(\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2)\,,\, f_\lambda(\bar{\mathbf{Z}}_1, \bar{\mathbf{Z}}_2)\big) \;=\; O\Big(\frac{k^2 \max\{1, \lambda^{-7}\}}{n^{1/2}}\Big)\,.$$

If additionally Assumption 6.2 holds, then

$$d_P\big(f_0(\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2)\,,\, f_0(\bar{\mathbf{Z}}_1, \bar{\mathbf{Z}}_2)\big) \;=\; o(1)\,.$$

While the assumptions are complicated, Lemma D.17 in the appendix verifies them for the isotropic Gaussian case. For simplicity, we now focus on the isotropic setup: For some fixed $\sigma_A > 0$, let

$$\mathrm{Var}[\mathbf{X}_1] \;= \mathbf{I}_d \qquad \text{and} \qquad \mathrm{Var}[\xi_{ij}] \;= \sigma_A^2 \mathbf{I}_d\,. \tag{6.22}$$

We defer to Lemma D.16 in the appendix to show that, under (6.22), both $\bar{\mathbf{Z}}_1$ and $\bar{\mathbf{Z}}_2$ are simple functions of the same $d \times nk$ rectangular matrix with i.i.d. standard Gaussian entries, whose limiting spectral density is the Marchenko-Pastur law. However, the correlations introduced by augmentations mean that, even in the isotropic case (6.22), the limiting spectra of $\bar{\mathbf{Z}}_1$ and $\bar{\mathbf{Z}}_2$ obey some compound Marchenko-Pastur laws — typically found in the anisotropic setup without augmentation — and the limiting risk is cumbersome to state, as seen in Hastie et al. (2022). Nevertheless, the Gaussian matrices allow us to derive meaningful surrogates for the risk in settings where the compound Marchenko-Pastur laws do simplify to a simple Marchenko-Pastur law. To specify this surrogate risk, we define, for $\beta \in \mathbb{R}^d$ and $\sigma, \lambda, \gamma > 0$,

$$R(\beta, \sigma, \lambda, \gamma) \;:=\; \|\beta\|^2 \lambda^2\, \partial m_\gamma(-\lambda) + \sigma^2 \gamma\,\big(m_\gamma(-\lambda) - \lambda \partial m_\gamma(-\lambda)\big)\,,$$

where $m_\gamma(z) := \frac{1-\gamma-z-\sqrt{(1-\gamma-z)^2-4\gamma z}}{2\gamma z}$. For $\lambda = 0$ or $\gamma = 0$, we define the above as the respective limit as $\lambda \to 0^+$ or $\gamma \to 0^+$. Hastie et al. (2022) shows that this is the limiting risk of $\hat{\beta}_\lambda^{(\mathrm{ora})}$ in the unaugmented case ($k = 1$ and $\sigma_A = 0$). The next proposition shows that, under an additional asymptotic constraint, the limiting risk of the augmented estimator can be expressed through R. This is possible because the additional constraint allows the risk to be characterised only by $\bar{\mathbf{Z}}_2$, the Wishart-distributed surrogate of $\bar{\mathbf{X}}_2$; see the proof in Appendix D.7.2 for details and for an explicit bound on the approximation.

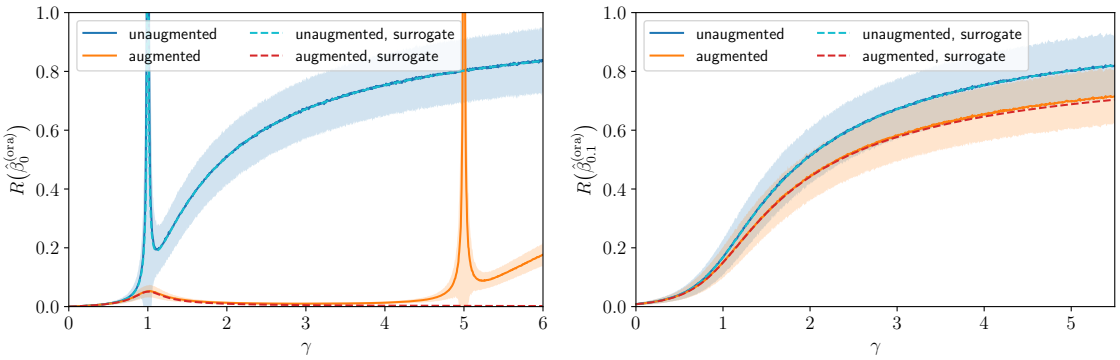


Figure 6.6: *Left.* Risk of the oracle ridgeless estimator $\hat{\beta}_0^{(\mathrm{ora})}$. *Right.* Risk of the oracle ridge estimator $\hat{\beta}_\lambda^{(\mathrm{ora})}$ with $\lambda = 0.1$. In both simulations, the data are generated as (6.16) with $n = 200$, varying d, $\|\beta\| = 1$ and $\sigma_\epsilon = 0.1$. The augmentations are noise injections defined in (6.21) with $k = 5$ and $\sigma_A = 0.1$. The risk used for simulation is defined in (6.25) while the theoretical risks are obtained from Proposition 6.11.

Proposition 6.11. *Consider the isotropic setup* (6.22) *and let* $k \geq 2$ *and* $\sigma_A^2 \leq 1$. *Write* $\lambda_k := \frac{(k-1)\sigma_A^2}{k} + \lambda$ *and* $\sigma_k^2 := \frac{k+\sigma_A^2}{k}$. *Consider the asymptotic regime* (6.20) *with* $\frac{\sigma_A^2}{\sqrt{k}}\frac{\sqrt{d}}{\sqrt{n}} = o(1)$ *and we allow* $\lambda \geq 0$. *Then*

$$f_\lambda(\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2) \xrightarrow{\mathbb{P}} \lim R\Big(\frac{\lambda}{\lambda_k}\beta, \frac{\sigma_\epsilon}{\sigma_k}, \frac{\lambda_k}{\sigma_k^2}, \gamma\Big) ,$$

where $\lim$ *denotes the limit under* (6.20) *with* $\frac{\sigma_A^2}{\sqrt{k}}\frac{\sqrt{d}}{\sqrt{n}} = o(1)$.

Proposition 6.11 is meaningful in two regimes: When $\sigma_A^2 \to 0^+$, i.e. little to no augmentations, or when $\gamma/k \to 0^+$, i.e. infinitely many augmentations compared to the dimension-to-sample-size ratio $\gamma = \lim d/n$. When the risk surrogates from Proposition 6.11 are valid, two effects of augmentation are visible: An additional regularization by $(k-1)\sigma_A^2/k$, and a shrinkage of effective size of β. The latter can be seen as a debiasing effect, as β only plays a role in the bias term of the risk. This mainly arises from the use of oracle augmentation, which introduces additional information on β. Section 6.5.2 shows that if we additionally need to estimate β in the augmentation, a bias term arises.

For the double-descent case $\lambda = 0$, the results can be interpreted as follows. As Hastie et al. (2022) explains, whether the unaugmented risk diverges to infinity is determined by the stability of the pseudoinverse. This stability is measured by the random quantity

$$\big\|\bar{\mathbf{X}}_1^\dagger\big\|_{op} = \Big\|\Big(\frac{1}{nk}\sum_{ij}(\mathbf{V}_i + \xi_{ij})(\mathbf{V}_i + \xi_{ij})^\top\Big)^\dagger\Big\|_{op} .$$

In the isotropic case, since both Gaussianity and the operator norm are invariant under orthogonal transformations, one may show (Lemma D.16 in the appendix) that the quantity above is distributed as

$$\Big\|\Big(\frac{1}{n}\sum_{i=1}^n \eta_{i1}\eta_{i1}^\top + \frac{\sigma_A^2}{nk}\sum_{i=1}^n\sum_{j=1}^k \eta_{ij}\eta_{ij}^\top\Big)^\dagger\Big\|_{op} =: \big\|\big(\mathbf{W}_1 + \frac{\sigma_A^2}{k}\mathbf{W}_2\big)^\dagger\big\|_{op} , \qquad (6.23)$$

where η_{ij} are i.i.d. standard Gaussians in $\mathbb{R}^d$ (see Lemma D.16 in the appendix for the derivation). The two matrices in (6.23) are differently scaled sample covariance matrices, one of n data and another of nk data. These matrices are correlated through $\{\eta_{i1}\}_{i=1}^n$. The behaviour of the risk can then be broken down as follows:

(i) If $\gamma = 1$ (i.e. $d \approx n$ asymptotically), the pseudoinverse of $\mathbf{W}_1$ is unstable, whereas since $\gamma' < 1$ (i.e. $d \lesssim kn$), $\mathbf{W}_2$ is asymptotically full-ranked and close to $\mathbb{E}\mathbf{W}_2$. Since $\mathbb{E}\mathbf{W}_2$ is a scaled identity matrix, it acts as a regularization of the pseudoinverse. The regularization effect is evident in Figure 6.6, where the risk curve of an augmented ridgeless regressor exhibits a small local maximum around $\gamma = 1$—similar to what is observed for a ridge regressor in Hastie et al. (2022)—instead of the spike towards infinity observed for the unaugmented risk curve. The same regularization effect can be seen from the surrogate risk formula from Proposition 6.11, computed based on the limiting Marchenko-Pastur law of $\mathbf{W}_1$; in Figure 6.6, the surrogate is a good approximation even when $\gamma = 1$ and $k = 5$, due to the small noise scale σ_A used.

(ii) If γ exceeds k, γ' exceeds 1, and d asymptotically exceeds kn. In this case, the sample covariance matrix $\mathbf{W}_2$ also becomes unstable, and is no longer regularises $\mathbf{W}_1$. That causes the risk to diverge, as illustrated in the left plot of Figure 6.6. The surrogate risk fails to be a good approximation in this regime, as the true risk is now characterised by a compound Marchenko-Pastur law arising from the limiting spectra of $\mathbf{W}_1 + \mathbf{W}_2$.

(iii) As this stability issue does not occur for $\lambda > 0$, no risk spikes are observed for ridge regression. When $\lambda > 0$, the pseudoinverse is also less sensitive to the minimum eigenvalue of the matrices, allowing for the surrogate risk from Proposition 6.11 to serve as a good approximation for larger range of values of γ. This is evident both in the improved rate of the approximation in Proposition 6.11 and in the right plot of Figure 6.6.

The analysis shows that the interpretation of augmentation as a regulariser suggested in the machine learning literature (Dao et al., 2019; Chen et al., 2020; Shorten and Khoshgoftaar, 2019; Balestriero et al., 2022b) depends on the interplay between the number of augmentations k, the number of data points n and the dimension d. Online augmentation (where the approximation $k = \infty$ can be justified) behaves like regulariser, as pointed out in previous work. In offline augmentation (where $k < \infty$), the risk still shows a spike towards infinity that is not regularised, although this spike now appears around $d \approx nk$ rather than $d \approx n$.

Remark 6.4 (Related work)**.** (i) The proofs of Hastie et al. (2022) use the fact that the random matrices in the unaugmented risk are all rescaled and shifted versions of $\bar{\mathbf{X}}_1$,

whose eigenspace align. That is a consequence of independence between data points, and no longer true if $k > 1$.

(ii) Noise injection is studied by Dhifallah and Lu (2021) for a small $\lambda > 0$, where double-descent is observed in a classification problem with a random feature model but not in regression. Although their work is phrased as a regularization approach, it can be regarded as augmentation. They employ a remarkable proof technique based on tools from convex analysis, and their results and ours are complementary: They assume Gaussian data and noise, and obtain two separate limiting expressions of the risk for an augmented estimator and an unaugmented estimator with a different regularization. Our analysis, on the other hand, shows that the shift in double-descent peak is in fact a combination of two effects: A regularization by noise injection around $d \approx n$, and a non-regularised instability around $d \approx nk$. Additionally, our results apply in the non-Gaussian case.

6.5.2. Double and triple descent for sample-splitting estimates

Augmenting the response variables requires knowledge of β. If we drop the oracle assumption, we can use a two-stage estimation process with sample splitting, where an initial estimate $\tilde{\beta}^{(m)}$ is computed on part of the data. On the remaining data, this value is used to augment both covariates and responses, and a final estimate $\hat{\beta}^{(m)}$ is computed. Consider m i.i.d. fresh draws of the data $\{\tilde{\mathbf{V}}_i, \tilde{Y}_i\}_{i=1}^m$ obtained e.g. via data splitting, and form an unaugmented estimator with parameter $\lambda \geq 0$:

$$\tilde{\beta}_\lambda^{(m)} \;\coloneqq\; \Big(\frac{1}{m}\sum_{i=1}^m \tilde{\mathbf{V}}_i\tilde{\mathbf{V}}_i^\top + \lambda \mathbf{I}_d\Big)^\dagger \frac{1}{m}\sum_{i=1}^m \tilde{\mathbf{V}}_i\tilde{\mathbf{Y}}_i \;.$$

In the case $m = 0$, we write $\tilde{\beta}_\lambda^{(0)} = \mathbf{0}$. The augmentations applied to Y_i's are given by

$$\tau_{ij}^{(m)} Y_i \;\coloneqq\; Y_i + \big(\pi_{ij}\mathbf{V}_i - \mathbf{V}_i\big)^\top \tilde{\beta}_\lambda^{(m)} \;=\; \tau_{ij}^{(\mathrm{ora})} Y_i + (\pi_{ij}\mathbf{V}_i - \mathbf{V}_i)^\top(\tilde{\beta}_\lambda^{(m)} - \beta) \;.$$

In this case, invariance of the covariates does not imply invariance of the entire data as in (6.17). The final augmented estimator is the two-stage estimator defined with $\tau_{ij}^{(m)}$ as

$$\hat{\beta}_\lambda^{(m)} \;\coloneqq (\bar{\mathbf{X}}_1 + \lambda \mathbf{I}_d)^\dagger \frac{1}{nk}\sum_{ij}(\pi_{ij}\mathbf{V}_i)\,\tau_{ij}^{(m)} Y_i \;.$$

Thus, $m = 0$ corresponds to not augmenting the response variables. Observe that the two-stage estimator is related to the oracle estimator by

$$\hat{\beta}_\lambda^{(m)} \;= \hat{\beta}_\lambda^{(\mathrm{ora})} + (\bar{\mathbf{X}}_1 + \lambda \mathbf{I}_d)^\dagger\, \bar{\mathbf{X}}_\Delta\, (\tilde{\beta}_\lambda^{(m)} - \beta) \;, \tag{6.24}$$

where the difference arises from the estimation error of the first-stage estimator, $\tilde{\beta}_\lambda^{(m)} - \beta$, as well as the difference arising from augmentation,

$$\bar{\mathbf{X}}_\Delta \;\coloneqq\; \frac{1}{n}\sum_{i=1}^n \Big(\frac{1}{k}\sum_{j=1}^k \pi_{ij}\mathbf{V}_i\Big)\Big(\frac{1}{k}\sum_{j=1}^k (\pi_{ij}\mathbf{V}_i - \mathbf{V}_i)\Big)^\top \;.$$

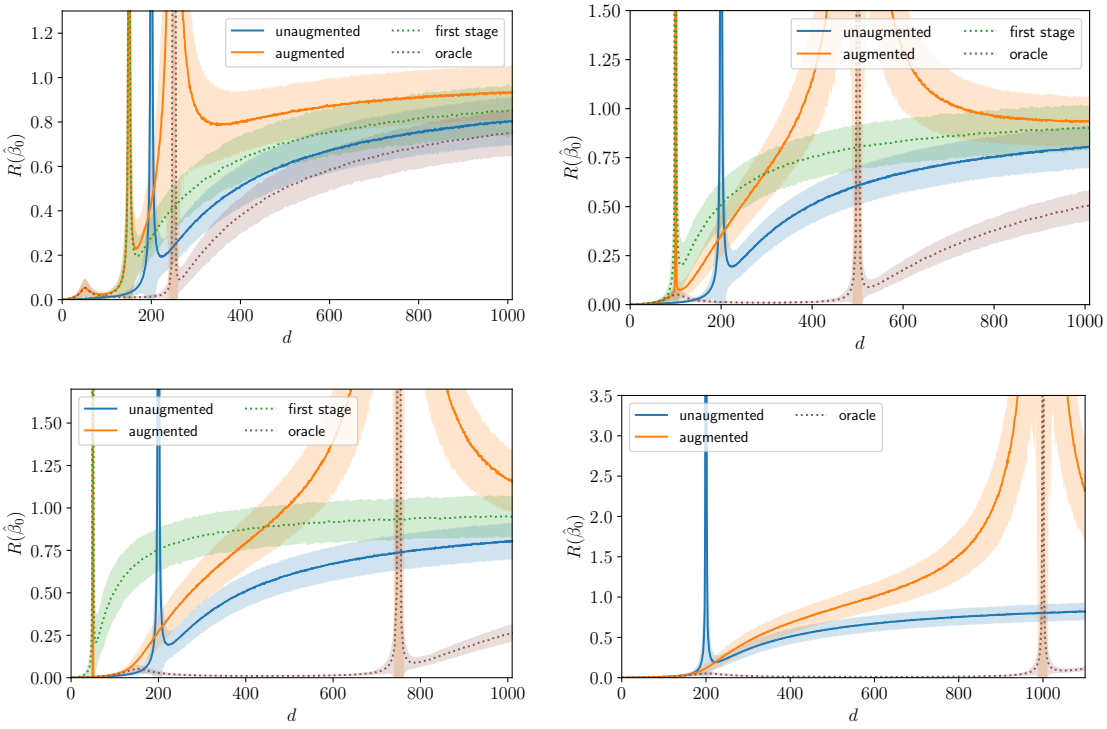


Figure 6.7: Risks of the two-stage ridgeless estimator $\hat{\beta}_0^{(m)}$. In all figures, $n_{\text{unaug}} = 200$ data are used for the unaugmented estimator and $k = 5$ augmentations are used for the augmented estimator. The number of data used for the two stages of the augmented estimator differ: *Top Left.* $m = n_{\text{aug}} = 100$; *Top Right.* $m = 50$ and $n_{\text{aug}} = 150$; *Bottom Left.* $m = 150$ and $n_{\text{aug}} = 50$; *Bottom Right.* $m = 0$ and $n_{\text{aug}} = 200$. In each figure, risk of the first-stage unaugmented estimator $\tilde{\beta}_0^{(m)}$ and risk of the oracle estimator $\hat{\beta}_0^{(\text{ora})}$ trained on $\{\mathbf{V}_i\}_{i=1}^{n_{\text{aug}}}$ are also plotted for comparison.

We consider the risk R defined in Section 6.4.5, which simplifies under the linear model (6.16) as

$$R(\hat{\beta}_\lambda^{(m)}) \;=\; \mathbb{E}[(Y_{\text{new}} - (\hat{\beta}_\lambda^{(m)})^\top \mathbf{V}_{\text{new}})^2 \,|\, \hat{\beta}_\lambda^{(m)}] \;=\; \|\hat{\beta}_\lambda^{(m)} - \beta\|^2 + \sigma_\epsilon^2 \,. \tag{6.25}$$

We are again interested in the double-descent case $\lambda = 0$.

Proposition 6.12. *Assume that* $\|\bar{\mathbf{X}}_1^\dagger\|_{op}$, $\|\bar{\mathbf{X}}_2\|_{op}$, $\|\bar{\mathbf{X}}_\Delta\|_{op}$ *and* $\|\tilde{\beta}_\lambda^{(m)} - \beta\|$ *are* $O(1)$ *with probability* $1 - o(1)$. *Then*

$$R(\hat{\beta}_0^{(m)}) - \big(\sigma_\epsilon^2 + \hat{L}_0^{(\text{ora})} + \big\|\bar{\mathbf{X}}_1^{-1}\bar{\mathbf{X}}_\Delta(\tilde{\beta}_\lambda^{(m)} - \beta)\big\|^2\big) \;\xrightarrow{\mathbb{P}}\; 0 \,.$$

In other words, the limiting risk $R(\hat{\beta}_0^{(m)})$ can be separated into the the risk $\hat{L}_0^{(\text{ora})}$ of the oracle estimator, a noise term σ_ϵ^2 that arises due to a different choice of the risk, and the term $\big\|\bar{\mathbf{X}}_1^{-1}\bar{\mathbf{X}}_\Delta(\tilde{\beta}_\lambda^{(m)} - \beta)\big\|^2$. Adapting our universality result allows one to show that $(\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_\Delta)$ behave like correlated matrices with Gaussian entries, and in the isotropic case, we expect delocalization of the eigenvectors of $\bar{\mathbf{X}}_1^{-1}\bar{\mathbf{X}}_\Delta$ in the sense that

$$\big\|\bar{\mathbf{X}}_1^{-1}\bar{\mathbf{X}}_\Delta(\tilde{\beta}_\lambda^{(m)} - \beta)\big\|^2 \;\approx\; \frac{1}{d}\text{Tr}\big(\bar{\mathbf{X}}_\Delta\bar{\mathbf{X}}_1^{-2}\bar{\mathbf{X}}_\Delta\big)\,\|\tilde{\beta}_\lambda^{(m)} - \beta\|^2 \,. \tag{6.26}$$

A formal justification requires developing anisotropic local laws similar to Knowles and Yin (2017) but for matrices of the form $\bar{\mathbf{X}}_1^\dagger \bar{\mathbf{X}}_\Delta$, which we leave to future work. Under (6.26), the main difference between the two-stage risk $R(\hat{\beta}_0^{(m)})$ and $\hat{L}_0^{(\mathrm{ora})}$ is a rescaled risk of the first-stage estimator. We expect $\hat{L}_0^{(\mathrm{ora})}$ to diverge near $\gamma' = 1$ (i.e. $d \approx kn$) and R_m to diverge near $\gamma/\rho = 1$ (i.e. $d \approx m$), leading to *two* spikes in the risk curve of $\hat{\beta}_0^{(m)}$. One spike is due to augmentation as discussed in Section 6.5.1, and hence not observed if $k \to \infty$. The other is due to the first-stage, unaugmented regressor on m data, and hence not observed if $m = 0$. Figure 6.7 shows empirical results for fixed k and $\lambda = 0$. Both double-descent (for $m = 0$) and triple-descent behaviours are clearly visible.

Remark 6.5. (i) The results above can be generalised from the ridgeless regressor considered here to two-layer linear networks. Indeed, Ba et al. (2019) and Chatterji et al. (2022) characterise the risk of such a network after training in terms of the pseudoinverse in (6.23). Our proof technique can be applied to this risk, at the price of more notation.

(ii) For simplicity, we have assumed the same value of λ is used in both stages, although our approach can be extended to distinct values. Since both stages use $\lambda = 0$, we see two peaks in the risk, and hence triple-descent. If a positive value is used in the first stage instead and $\lambda = 0$ in the second, one of the peaks would vanish.

6.6 Universality for other non-smooth and high-dimensional estimators

So far, we have demonstrated universality for properties of the ridge regression and linear regression. In this section, we present two additional results that show how the universality results also apply to a non-smooth statistic and a high-dimensional statistic. The second example also illustrates how augmentation may increase effective sample size.

6.6.1. Maximum of exponentially many correlated random variables

As an example of non-smooth statistic, consider the function

$$f(\mathbf{x}_{11}, \dots, \mathbf{x}_{nk}) \coloneqq \max_{1 \le l \le d_n} \frac{1}{nk} \sum_{i \le n} \sum_{j \le k} x_{ijl} \qquad \text{for} \quad \mathbf{x}_{11}, \dots, \mathbf{x}_{nk} \in \mathbb{R}^{d_n} ,$$

where the dimension d_n grows exponentially in n as specified below. This function occurs in the context of uniform confidence bands (Deng and Zhang, 2020) and high-dimensional central limit theorems (Chernozhukov et al., 2013).

Proposition 6.13. *Consider i.i.d. Gaussian variables* $\{\mathbf{Z}_i\}_{i \le n}$ *that satisfy* (6.1). *Suppose* $\mathbb{E}[\phi_{11}\mathbf{X}_1] = \mathbf{0}$ *and the moments* $\| \max_{l \le d_n} |(\phi_{11}\mathbf{X}_1)_l| \|_{L_6}$ *and* $\| \max_{l \le d_n} |(\mathbf{Z}_{11})_l| \|_{L_6}$ *are*

finite for each $d_n \in \mathbb{N}$. *Then*

$$d_{\mathcal{H}}(\sqrt{n}f(\Phi\mathcal{X}), \sqrt{n}f(\mathcal{Z})) \to 0 \qquad \textit{and} \qquad n\big\|\mathrm{Var}[f(\Phi\mathcal{X})] - \mathrm{Var}[f(\mathcal{Z})]\big\| \to 0$$

whenever d_n *grows as* $\log(d_n) = o(n^{1/10})$.

Thus, the maximum coordinate of the high-dimensional augmented average can be approximated by the maximum of a high-dimensional Gaussian. Similar results are available in the literature on high-dimensional central limit theorems, and suggest that the condition on d_n can be improved.

6.6.2. Softmax ensemble of exponentially many estimators

Let $\beta_1, \ldots, \beta_{m_n}$ be $\mathbb{R}^p$-valued functions, for some $m_n \in \mathbb{N}$. We think of these as estimators or predictors previously calibrated on a separate sample. Fix a loss function $L : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$ and a scalar t, and define

$$f(\mathbf{x}_{11}, \ldots, \mathbf{x}_{nk}) \coloneqq \sum_{r=1}^{m_n} \beta_r \frac{\exp\big(-t\frac{\log(m_n)}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k} L(\beta_r, \mathbf{x}_{ij})\big)}{\sum_{s=1}^{m}\exp\big(-t\frac{\log(m_n)}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k} L(\beta_s, \mathbf{x}_{ij})\big)} \qquad \text{for } \mathbf{x}_{ij} \in \mathbb{R}^d \,.$$

For $k = 1$, this is a softmax version of the *super learner* built on the base estimators β_r, see e.g. Van der Laan et al. (2007). For large values of t, the function f approximates

$$f \;\approx\; \mathrm{argmin}_{\{\beta_r | 1 \le r \le m\}} \frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k} L(\beta_r, \mathbf{x}_{ij}) \,.$$

The number m_n of estimators is typically permitted to grow exponentially with n. The result below shows that augmenting f can effectively increase sample size. To make that precise, we define an "effective sample size" $n^* = c_\phi n$, for some $c_\phi > 0$, and write $\mu_r \coloneqq \mathbb{E}[L(\beta_r, \mathbf{X}_1)]$. Consider

$$f^*(\mathbf{x}_1, \ldots, \mathbf{x}_n) \;\coloneqq\; \sum_{r=1}^{m_n} \beta_r \frac{\exp\big(-t\log(m_n)\big(\frac{1}{n^*}\sum_{i=1}^{n}(L(\beta_r, \mathbf{x}_i) - \mu_r) + \mu_r\big)\big)}{\sum_{s=1}^{m}\exp\big(-t\log(m_n)\big(\frac{1}{n^*}\sum_{i=1}^{n}(L(\beta_s, \mathbf{x}_i) - \mu_s) + \mu_s\big)\big)} \,.$$

If an augmented ensemble $f(\Phi\mathcal{X})$ behaves like $f^*(\mathcal{X})$ for some $c_\phi > 1$, then augmentation increases effective sample size.

Proposition 6.14 (ensemble of exponentially many estimators)**.** *Assume that the data distribution is invariant, i.e.* $\phi_{11}\mathbf{X}_1 \overset{d}{=} \mathbf{X}_1$, *and that* $\|\max_{r \le m_n} |L(\beta_r, \phi_{11}\mathbf{X}_1)|\|_{L_6}$ *and* $\|\max_{r \le m_n} |L(\beta_r, \mathbf{X}_1)|\|_{L_6}$ *are bounded. If*

$$\textstyle\sum_{l=1}^{p}(\sup_{r \le m_n} |(\beta_r)_l|) = O(1) \qquad \textit{and} \qquad \log m_n = o(n^{1/9})$$

then there exists $c_\phi \ge 1$ *such that*

$$d_{\mathcal{H}}(\sqrt{n}f(\Phi\mathcal{X}), \sqrt{n}f^*(\mathcal{X})) \to 0 \qquad \textit{and} \qquad n\big\|\mathrm{Var}[f(\Phi\mathcal{X})] - \mathrm{Var}[f^*(\mathcal{X})]\big\| \to 0 \,.$$

An explicit formula for c_ϕ is given in Appendix D.6.6, and shows that there are cases

where indeed $c_\phi > 1$. The scaling $\log(m_n)$ is justified in Lemma D.39 in the appendix.

Chapter 7

Implications of universality in optimisation analysis

In all applications considered so far, the estimator f in (1.1) admits an explicit closed-form formula in terms of the data. There, we have seen how universality greatly simplifies analyses, since one may extract various theoretical properties just by studying a closed-form expression of Gaussians. In day-to-day machine learning, however, many estimators do not admit a closed-form formula. A common example is some estimator $\hat{\beta}(X)$ that depends implicitly on the data via some optimisation

$$\hat{\beta}(X) \ \in \ \text{argmin}_{\beta \in \mathbb{R}^d} \, \text{loss}(\beta; X_1, \ldots, X_n) \ . \tag{7.1}$$

When there are non-unique minimisers, the estimate $\hat{\beta}(X)$ additionally depends on the specific choice of optimisation algorithm used for obtaining $\hat{\beta}(X)$.

Universality results have now been established for various 1-dimensional properties $f(X)$ of the estimators $\hat{\beta}(X)$ in (7.1), e.g. when $f(X)$ is the training or test risk of $\hat{\beta}(X)$, mostly by the Lindeberg method and its variants. A non-exhaustive list of examples includes random feature models (Hu and Lu, 2022), regularised regression (Han and Shen, 2023), block dependent linear models (Lahiry and Sur, 2024), generalised linear models (Dandi et al., 2023), max-margin classifiers (Montanari et al., 2023) and general classes of empirical risk minimisers (Montanari and Saeed, 2022). In a recent joint work (Mallory, Huang, and Austern, 2025), we also establish universality for the risks of high-dimensional logistic regression classifiers, where data are allowed to be dependent.

Once universality results are established, one can already empirically investigate the behaviour of the estimators by substituting the data X with a set of Gaussians Z, which are computationally fast to generate. However, the non-closed-form nature of f means that $f(Z)$ is still complicated to analyse, and obtaining a rigorous theoretical statement can be much more difficult than the closed-form cases.

This chapter discuss several tools that, in setups where Gaussian universality does hold, can aid the analysis of quantities that arise in optimisation:

- Section 7.1 examines the convex Gaussian min-max theorem (CGMT), a theoretical

tool that is particularly well-suited for analysing optimisation problems on Gaussian data. We shall develop an extension of known CGMT results to accommodate dependence both across dimensions and across different data points, and briefly discuss its usage in Mallory, Huang, and Austern (2025) for analysing the effects of data augmentation on high-dimensional classifiers;

- Section 7.2 examines the stability analysis of stochastic optimisation algorithms used for training $\hat{\beta}(X)$ in the context of two specific examples.
  - The first is contrastive divergence (CD), an ML algorithm used for training energy-based models. We briefly discuss how universality plays a role in obtaining the necessary moment control for a multi-step stability analysis and obtaining the consistency of $\hat{\beta}(X)$. The discussion is used in a particular setting as part of the broader analysis of a joint work (Glaser, Huang, and Gretton, 2024), which provides near-optimal error bounds for the convergence of CD in various settings.
  - The second is variational Monte Carlo (VMC), an algorithm used in training large-scale neural network solvers to the Schrödinger equation. We briefly examine how the maximum CLT (a tight version of Proposition 6.13) can be applied to analyse the stability of a high-dimensional gradient update under data augmentation. This result is used as part of the broader analysis of a joint work (Huang, Zhan, Ertekin, Orbanz, and Adams, 2025), which examines the effects of different symmetrisations in neural network VMC solvers under a particularly intricate case of symmetry in crystals.

7.1 Dependent convex Gaussian min-max theorem

Many modern tools have been developed for a system of equations or an optimisation problem that involves only Gaussian random variables, such as the cavity method (Opper et al., 2001), approximate message passing method (Donoho et al., 2009), the replica method (Mézard et al., 1987) and the convex Gaussian min-max theorem (CGMT) (Gordon, 1985; Thrampoulidis et al., 2014). Among them, the CGMT is a framework that converts a complex optimisation problem on Gaussian data to a much more analytically tractable auxiliary problem. The auxiliary optimisation is often further simplified into a deterministic equation involving only a few scalars, and under the CGMT, its solution completely characterises that of the original problem.

In this section, we first provide an informal overview of the standard CGMT recipe for risk analysis in the case when $X_1, \dots, X_n$ are i.i.d. random vectors with i.i.d. coordinates (Thrampoulidis et al., 2014; Thrampoulidis, 2016); we refer interested readers to

Thrampoulidis (2016) for a detailed technical introduction. We proceed to present our extension to the dependent setup, and briefly discuss its usage in Mallory, Huang, and Austern (2025) for analysing the effects of data augmentation in logistic regression. A diagrammatic illustration of the universality-CGMT receipe is included at the end of the section in Figure 7.2.

7.1.1. An informal sketch of the universality-CGMT recipe

Let $X_1, \dots, X_n$ be mean-zero $\mathbb{R}^d$-random vectors (not assumed to be i.i.d. or isotropic for now). Suppose the corresponding labels $y_1, \dots, y_n$ are generated as $y_i = y(X_i^\top \beta_*)$ for some link function $y : \mathbb{R} \to \mathbb{R}$ and some unknown vector $\beta_* \in \mathbb{R}^d$ to be estimated. Consider the optimisation problem

$$\min_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum\nolimits_{i \le n} l\big(X_i^\top \beta\,,\, y(X_i^\top \beta_*)\big)\ ,$$

where $l : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is some loss function. We first discuss the analysis of the training risk, i.e. the global minimum of the above optimisation; it shall become clear in Section 7.1.2 how this can be utilised to analyse the test risk or other 1d properties of the minimiser.

The first step is to reformulate the optimisation problem, such that most of the technical difficulties are captured by a multilinear term involving a high-dimensional Gaussian matrix. CGMT then allows one to, informally, replace this Gaussian matrix by Gaussian vectors, which allows for easy downstream processing.

To this end, denote the concatenated $\mathbb{R}^{d \times n}$ data matrix $\mathbf{X} := (X_1, \dots, X_n)$ and $\mathbf{y} : \mathbb{R}^n \to \mathbb{R}$ as the coordinate-wise application of y, i.e.

$$\mathbf{y}(\mathbf{X}^\top \beta_*) \ = (y(X_1^\top \beta_*), \dots, y(X_n^\top \beta_*))^\top \ = \ (y_1, \dots, y_n)^\top\ .$$

Then for some appropriately chosen $L : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, we can express the original optimisation (OO) as

$$\min\nolimits_{\beta \in \mathbb{R}^d} \ L\big(\mathbf{X}^\top \beta, \mathbf{y}(\mathbf{X}^\top \beta_*)\big)\ . \tag{OO}$$

When Gaussian universality holds, we can WLOG study (OO) via the Gaussian optimisation

$$\min\nolimits_{\beta \in \mathbb{R}^d} \ L\big(\mathbf{Z}^\top \beta, \mathbf{y}(\mathbf{Z}^\top \beta_*)\big)\ , \tag{GO}$$

where $\mathbf{Z} = (Z_1, \dots, Z_n)$ is the matrix of Gaussian surrogates with the same mean and covariance structure as $\mathbf{X}$. (GO) can be rewritten as a constrained optimisation $\min_{\beta \in \mathbb{R}^d, \nu \in \mathbb{R}^n} L(\nu, \mathbf{y}(\mathbf{Z}^\top \beta_*))$ subject to $\nu = \mathbf{Z}^\top \beta$. By introducing the Lagrange mul-

tiplier $u \in \mathbb{R}^n$, we can further rewrite (GO) as

$$\min_{\beta\in\mathbb{R}^d,\nu\in\mathbb{R}^n} \max_{u\in\mathbb{R}^n} L(\nu, \mathbf{y}(\mathbf{Z}^\top\beta_*)) + (\beta^\top\mathbf{Z} - \nu^\top)u \ .$$

So far, we have successfully extracted a multilinear term involving the Gaussian matrix $\mathbf{Z}$. However, a subtle technical difficulty persists due to the dependence of the $\mathbb{R}^n$ random vector $\mathbf{y}(\mathbf{Z}^\top\beta_*)$ on $\mathbf{Z}$. In the case where the data are i.i.d. with i.i.d. coordinates, zero mean and identity covariance, a standard trick (see e.g. Thrampoulidis (2016); Salehi et al. (2019)) is to consider the projection $P_* \coloneqq \beta_*\beta_*^\top/\|\beta_*\|^2 \in \mathbb{R}^{d\times d}$: Since $\mathbf{Z}$ has i.i.d. $\mathcal{N}(0,1)$ entries, the projected matrix $\mathbf{H} \coloneqq (I - P_*)\mathbf{Z}$ is independent of $P_*\mathbf{Z}$ and $\mathbf{y}(\mathbf{Z}^\top\beta_*)$. This allows us to rewrite (GO) further as

$$\min_{\beta\in\mathbb{R}^d,\nu\in\mathbb{R}^n} \max_{u\in\mathbb{R}^n} \beta^\top\mathbf{H}u + \beta^\top P_*\mathbf{Z}u - \nu^\top u + L(\nu, \mathbf{y}(\mathbf{Z}^\top\beta_*)) \ .$$

Assuming that L is convex in the first argument, and writing $\psi(\beta, u) \coloneqq \min_{\nu\in\mathbb{R}^n} \beta^\top P_*\mathbf{Z}u - \nu^\top u + L(\nu, \mathbf{y}(\mathbf{Z}^\top\beta_*))$, one can apply the min-max theorem of Rockafellar (1970) to obtain

$$\min_{\beta\in\mathbb{R}^d,\nu\in\mathbb{R}^n} \max_{u\in\mathbb{R}^n} \beta^\top\mathbf{H}u + \psi(\beta, u) \ . \tag{PO}$$

Note that the only source of randomness in $\psi(\beta, u)$ is from a high-dimensional $\mathbb{R}^n$ vector $\mathbf{Z}^\top\beta_*$, which by construction is independent of $\mathbf{H}$. Optimisations of the above form are called the primary optimisation (PO) in CGMT. A direct analysis of (PO) can be rather cumbersome: One needs to consider how the limiting spectrum of the high-dimensional random matrix $\mathbf{H}$ interacts with the function ψ over the high-dimensional spaces $\mathbb{R}^d$ and $\mathbb{R}^n$.

Suppose for simplicity that $\mathbf{H}$ is a matrix with i.i.d. $\mathcal{N}(0,1)$ entries, and ignore the stochasticity in $\psi(\beta, u)$. The standard CGMT (Thrampoulidis et al., 2014; Thrampoulidis, 2016) says that, under mild conditions on ψ and by restricting the domains of optimisation appropriately, one may study the minimised value and properties of the minimisers of (PO) via the auxiliary optimisation

$$\min_{\beta\in\mathbb{R}^d,\nu\in\mathbb{R}^n} \max_{u\in\mathbb{R}^n} \|\beta\|\mathbf{h}^\top u + \beta^\top\mathbf{g}\|u\| + \psi(\beta, u) \ , \tag{AO}$$

where $\mathbf{h}$ and $\mathbf{g}$ are two independent standard Gaussian vectors, each taking values in $\mathbb{R}^n$ and $\mathbb{R}^d$. The formal result is proved by applying Gordon's Gaussian comparison inequality (Gordon, 1985) to two suitably constructed Gaussian processes, and we defer the technical discussion to the proof of Theorem 7.1 in Appendix E. As a simple heuristic, we note that the formulation (AO) is expected: As the matrix $\mathbf{H}$ has i.i.d. $\mathcal{N}(0,1)$ entries and are thereby invariant to left-multiplication of $\mathbb{R}^{d\times d}$ rotations and right-multiplication of $\mathbb{R}^{n\times n}$ rotations, one may expect β and u to only contribute to the term $\beta^\top\mathbf{H}u$ only

through the Euclidean norms $\|\beta\|$ and $\|u\|$. (AO) essentially formalises this, since

$$\|\beta\|\mathbf{h}^\top u\,,\ \beta^\top \mathbf{g}\|u\| \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \|\beta\|^2\|u\|^2)\ .$$

The main practical advantage of analysing (AO) in lieu of (PO) is that (AO) no longer involves a high-dimensional matrix and only depends on high-dimensional vectors. Indeed, even when one considers the stochasticity of ψ, (AO) only depends on the three high-dimensional Gaussian vectors $\mathbf{h}$, $\mathbf{g}$ and $\mathbf{Z}^\top\beta_*$, all of which appear in the loss through some 1d quantities. Without the maximum and the minimum over the high-dimensional spaces $\mathbb{R}^d$ and $\mathbb{R}^n$, one would be able to apply the law of large numbers directly to obtain a deterministic formulation of the objective. In the CGMT literature, the min-max operators are handled on a case-by-case basis depending on L. A common pattern is that they typically involve introducing additional auxiliary variables e.g. $r = \|\nu\|$ to reduce the min-max operators to be over a low-dimensional space, after which an appropriate notion of law of large numbers is applied. The end result is typically a low-dimensional, deterministic (but not necessarily convex) optimisation, which may be investigated numerically or used to prove theoretical statements about (OO).

A diagrammatic illustration of the above recipe is included in (7.2). We do not focus on the direct analysis of (AO) in this thesis, but refer interested readers to the many successful applications of CGMT to the risk analysis of high-dimensional models (Stojnic, 2013a,b; Thrampoulidis et al., 2015; Thrampoulidis, 2016; Thrampoulidis et al., 2018; Mignacco et al., 2020; Dhifallah and Lu, 2021; Aolaritei et al., 2022; Javanmard and Soltanolkotabi, 2022; Akhtiamov et al., 2024a,b). Our focus will be to generalise the standard CGMT, which only works for $\mathbf{H}$ with i.i.d. entries, to a setup that allows for dependent rows and columns in the Gaussian matrix $\mathbf{H}$. In particular, this enables us to extend the above recipe to allow dependence both across data points and across coordinates.

7.1.2. Dependent CGMT

We shall develop a more general CGMT framework that accommodates a "low-rank assumption" on the dependence structure of $\mathbf{H}$. More formally, we assume the following:

Assumption 7.1 (Low-rank dependence)**.** Let $\mathbf{H}$ be an $\mathbb{R}^{d\times n}$ Gaussian matrix. There exist $M \in \mathbb{N}$ and symmetric positive semi-definite matrices $(\Sigma^{(l)}, \tilde{\Sigma}^{(l)})_{l\le M}$, with $\Sigma^{(l)} \in \mathbb{R}^{d\times d}$ and $\tilde{\Sigma}^{(l)} \in \mathbb{R}^{n\times n}$, such that

$$\mathrm{Cov}[\mathbf{H}_{ji}, \mathbf{H}_{j'i'}] \ = \ \textstyle\sum_{l=1}^{M} \Sigma^{(l)}_{jj'}\tilde{\Sigma}^{(l)}_{ii'} \qquad \text{for all } i, i' \le n \text{ and } j, j' \le d\ .$$

Remark 7.1. (i) When $M = 1$, we can re-express $\mathbf{H} = (\Sigma^{(1)})^{1/2}\mathbf{H}'(\tilde{\Sigma}^{(1)})^{1/2}$, where $\mathbf{H}'$ is an $\mathbb{R}^{d\times n}$ matrix with i.i.d. $\mathcal{N}(0,1)$ entries. In this case, by redefining β and u in

(PO), the standard CGMT still applies. Assumption 7.1 can therefore be understood as a generalisation of this argument in the case where the re-expression is not possible, but the dependence structure of $\mathbf{H}$ is still fully specified by a sum of M matrix products, each involving an $\mathbb{R}^{d\times d}$ matrix and an $\mathbb{R}^{n\times n}$ matrix. (ii) We call Assumption 7.1 a low-rank assumption due to the restriction on the size of M in Theorem 7.1.

The primary optimisation we study takes the form

$$\Psi_{\mathcal{S}_d,\mathcal{S}_n} \coloneqq \min_{w\in\mathcal{S}_d}\max_{u\in\mathcal{S}_n} L_\Psi(w,u) \quad \text{with} \quad L_\Psi(w,u) \coloneqq w^\top \mathbf{H} u + f(w,u)\ , \tag{7.2}$$

where $\mathcal{S}_d \subset \mathbb{R}^d$ and $\mathcal{S}_n \subset \mathbb{R}^n$ are the domains of optimisation and $f : \mathcal{S}_d \times \mathcal{S}_n \to \mathbb{R}$ is some function that plays the role of ψ in (PO). Denote $\|v\|_{\Sigma'} = \sqrt{v^\top \Sigma' v}$. Under Assumption 7.1, we shall compare $\Psi_{\mathcal{S}_d,\mathcal{S}_n}$ to the risk

$$\psi_{\mathcal{S}_d,\mathcal{S}_n} \coloneqq \min_{w\in\mathcal{S}_d}\max_{u\in\mathcal{S}_n} L_\psi(w,u)\ , \tag{7.3}$$

$$\text{where} \quad L_\psi(w,u) \coloneqq \sum_{l=1}^M \left(\|w\|_{\Sigma^{(l)}} \mathbf{h}_l^\top \left(\tilde{\Sigma}^{(l)}\right)^{1/2} u + w^\top \left(\Sigma^{(l)}\right)^{1/2} \mathbf{g}_l \|u\|_{\tilde{\Sigma}^{(l)}} \right) + f(w,u).$$

Here, $(\mathbf{h}_l, \mathbf{g}_l)_{l\le M}$ are independent standard Gaussians respectively in $\mathbb{R}^n$ and $\mathbb{R}^d$. Notice that (7.3) is analogous to (AO), except that the M pairs of covariance matrices are retained in (AO).

Our next result formalises the equivalence of $\Psi_{\mathcal{S}_d,\mathcal{S}_n}$ and $\psi_{\mathcal{S}_d,\mathcal{S}_n}$, and additionally controls $\hat{w}_\Psi \in \mathcal{S}_d$, the minimiser of $\Psi_{\mathcal{S}_d,\mathcal{S}_n}$.

Theorem 7.1 (Dependent CGMT)**.** *Suppose Assumption 7.1 holds, and that $\mathcal{S}_d$ and $\mathcal{S}_n$ are compact and f is continuous on $\mathcal{S}_d \times \mathcal{S}_n$. Then the following statements hold:*

(i) For all $c \in \mathbb{R}$,

$$\mathbb{P}(\Psi_{\mathcal{S}_d,\mathcal{S}_n} \le c) \ \le\ 2^M \mathbb{P}(\psi_{\mathcal{S}_d,\mathcal{S}_n} \le c)\ .$$

(ii) If additionally $\mathcal{S}_d$ and $\mathcal{S}_n$ are convex and f is convex-concave on $\mathcal{S}_d \times \mathcal{S}_n$, then for all $c \in \mathbb{R}$,

$$\mathbb{P}(\Psi_{\mathcal{S}_d,\mathcal{S}_n} \ge c) \ \le\ 2^M \mathbb{P}(\psi_{\mathcal{S}_d,\mathcal{S}_n} \ge c)\ ,$$

and in particular, for all $\mu \in \mathbb{R}$ and $t > 0$,

$$\mathbb{P}(|\Psi_{\mathcal{S}_d,\mathcal{S}_n} - \mu| \ge t) \ \le\ 2^M\, \mathbb{P}(|\psi_{\mathcal{S}_d,\mathcal{S}_n} - \mu| \ge t)\ . \tag{7.4}$$

(iii) Assume the conditions of (ii). Let $\mathcal{A}_d$ be an arbitrary open subset of $\mathcal{S}_d$ and $\mathcal{A}_d^c \coloneqq \mathcal{S}_d \setminus \mathcal{A}_d$. If there exists constants $\bar{\psi}_{\mathcal{S}_d}$, $\bar{\psi}_{\mathcal{A}_d^c}$ and $\eta, \epsilon > 0$ such that $\bar{\psi}_{\mathcal{A}_d^c} \ge \bar{\psi}_{\mathcal{S}_d} + 3\eta$, $\mathbb{P}(\psi_{\mathcal{S}_d,\mathcal{S}_n} \le \bar{\psi}_{\mathcal{S}_d} + \eta) \ge 1-\epsilon$ and $\mathbb{P}(\psi_{\mathcal{A}_d^c,\mathcal{S}_n} \ge \bar{\psi}_{\mathcal{A}_d^c} - \eta) \ge 1-\epsilon$, then

$$\mathbb{P}(\hat{w}_\Psi \in \mathcal{A}_d) \ge 1 - 4\epsilon\ . \tag{7.5}$$

Remark 7.2 (Compact and convex domains $\mathcal{S}_d$ and $\mathcal{S}_n$)**.** CGMT operates on compact and convex sets $\mathcal{S}_d$ and $\mathcal{S}_n$. To extend the result to the case with $\mathcal{S}_d = \mathbb{R}^d$ and $\mathcal{S}_n = \mathbb{R}^n$, one typically shows that for the particular problem of interest, the minimiser and maximiser lie in some compact sets with high probability; see e.g. Montanari and Saeed (2022); Lahiry and Sur (2024).

Remark 7.3 (Comparison to existing CGMT results)**.** The standard CGMT in the i.i.d. isotropic case is exactly the same as above with $\Sigma^{(1)} = I_p$, $\tilde{\Sigma}^{(1)} = I_n$ and $M = 1$; see Theorem 3.3.1 of Thrampoulidis (2016). Our result also recovers the multivariate CGMT of Dhifallah and Lu (2021) by setting $\Sigma^{(l)}$ and $\tilde{\Sigma}^{(l)}$ as block diagonal matrices with M equal-sized subblocks, such that the l-th subblock is identity and the other blocks are zero. Akhtiamov et al. (2024a) generalises the block diagonal setup to allow non-identity subblocks, which is a special case of our Assumption 7.1, but they also allow for transforming w and u, which we do not address here.

Remark 7.4 (Comment on the 2^M factor)**.** Similar to the i.i.d. isotropic CGMT, our Theorem 7.1 is proved by applying Gordon's Gaussian comparison inequality to two suitably chosen Gaussian processes, one of which corresponds to (7.3) whereas the other corresponds to (7.2) with M additional univariate Gaussian terms. The factor 2^M arises from approximating the M univariate Gaussian terms away. See Theorem 3.3.1 of Thrampoulidis (2016) for the proof in the i.i.d. isotropic case, and Appendix E for the proof in the general case.

The interpretation of the results are similar to that of the standard CGMT. For readers unfamiliar with the CGMT literature, we note that for most practical purposes, (7.4) and (7.5) are the two key CGMT results, whereas the rest can be viewed as intermediate steps to obtain these results. Roughly speaking, they can be interpreted as follows:

- (7.4) implies that if the risk of the auxiliary optimisation $\psi_{\mathcal{S}_d,\mathcal{S}_n}$ is close to some value μ with high probability, then necessarily the risk of the primary optimisation $\Psi_{\mathcal{S}_d,\mathcal{S}_n}$ is close to μ with high probability. This allows us to analyse the risk $\psi_{\mathcal{S}_d,\mathcal{S}_n}$ in lieu of $\Psi_{\mathcal{S}_d,\mathcal{S}_n}$;
- (7.5) concerns an "important set" $\mathcal{A}_d \subseteq \mathcal{S}_d$. It says that, if the inclusion or exclusion of $\mathcal{A}_d$ results in a substantial change of the risk of the auxiliary optimisation (from $\bar{\psi}_{\mathcal{S}_d}$ to $\bar{\psi}_{\mathcal{A}_d^c}$), which automatically implies a substantial change of the risk of the *primary* optimisation, then the minimiser $\hat{w}_\Psi$ of the *primary* optimisation must lie in the set $\mathcal{A}_d$ with high probability. This allows us to use the analysis of the auxiliary optimisation to make a statement about the minimiser of the primary optimisation, and thereby the original optimisation.

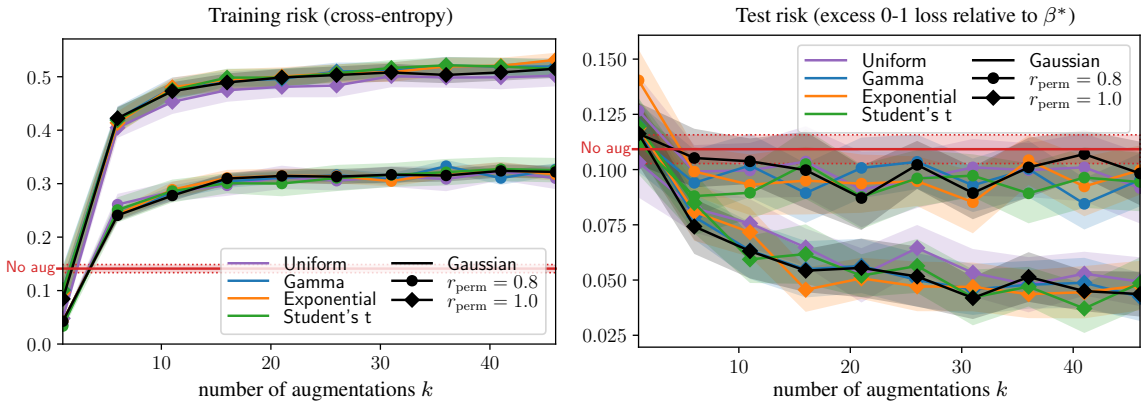


Figure 7.1: Figure 1 of Mallory, Huang, and Austern (2025). Universality of risks of a logistic regressor, trained with different number and amount of random permutations. See the paper for the detailed setup.

A more convenient version of (7.5) is the next asymptotic concentration statement. The proofs for both Theorem 7.1 and Corollary 7.2 are included in Appendix E.

Corollary 7.2 (Asymptotic CGMT)**.** *Assume the conditions of Theorem 7.1(ii) and let* $\mathcal{A}_d, \mathcal{A}_d^c$ *be defined as in Theorem 7.1(iii). If there exist constants* $\bar{\psi} < \bar{\psi}^c$ *s.t.* $\psi_{\mathcal{S}_d,\mathcal{S}_n} \xrightarrow{\mathbb{P}} \bar{\psi}$ *and* $\psi_{\mathcal{A}_d^c,\mathcal{S}_n} \xrightarrow{\mathbb{P}} \bar{\psi}^c$*, then*

$$\mathbb{P}(\hat{w}_\Psi \in \mathcal{A}_d) \;\to\; 1\,.$$

In practice, for both the standard CGMT and the dependent CGMT, one choose $\mathcal{A}_d$ depending on the desired properties of $\hat{w}_\Psi$ to investigate. For instance, if $\hat{w}_\Psi$ corresponds to some estimator $\hat{\beta}(X)$, one may let $\mathcal{A}_d$ be the set $\{\beta \in \mathbb{R}^d \,|\, |\mathbb{E}[l(X_{\text{new}}^\top \beta, Y_{\text{new}})] - \chi| \leq \epsilon\}$ for some test loss function $l : \mathbb{R}^2 \to \mathbb{R}$, some conjectured value of the test risk $\chi \in \mathbb{R}$, and an arbitrarily small $\epsilon > 0$. (7.2) then allows us to make high probability statements about the test risk of $\hat{\beta}(X)$ by analysing only the auxiliary optimisation, which can often be replaced by low-dimensional, deterministic optimisation, as discussed in Section 7.1.1. We also do not need to know the value of χ a priori, as this is typically deduced directly from the auxiliary optimisation analysis.

This approach of choosing $\mathcal{A}_d$ was used in the test risk analysis in the joint work of Mallory, Huang, and Austern (2025), which considers a high-dimensional logistic regression problem with dependent data. This is also the work where Theorem 7.1 and Corollary 7.2 were developed. In that work, Corollary 7.2 is used in conjunction with a training risk universality result under dependence to prove *test risk* universality. We also show that data augmentation (see e.g. Chapter 6 for the precise definition) corresponds exactly to Assumption 7.1 with $M = 2$. This enables us to derive a set of low-dimensional, deterministic fixed-point equations that completely characterise the effects of DA on high-dimensional logistic regression, which also recover the fixed-point equations derived by Salehi et al. (2019) for the isotropic i.i.d. setup as a special case. As a

practical application, we have applied this universality-CGMT recipe to investigate the effectiveness of data augmentation, and observe that the benefits of data augmentation can be highly reliant on the full knowledge of invariance. We refer interested readers to the paper for more details.

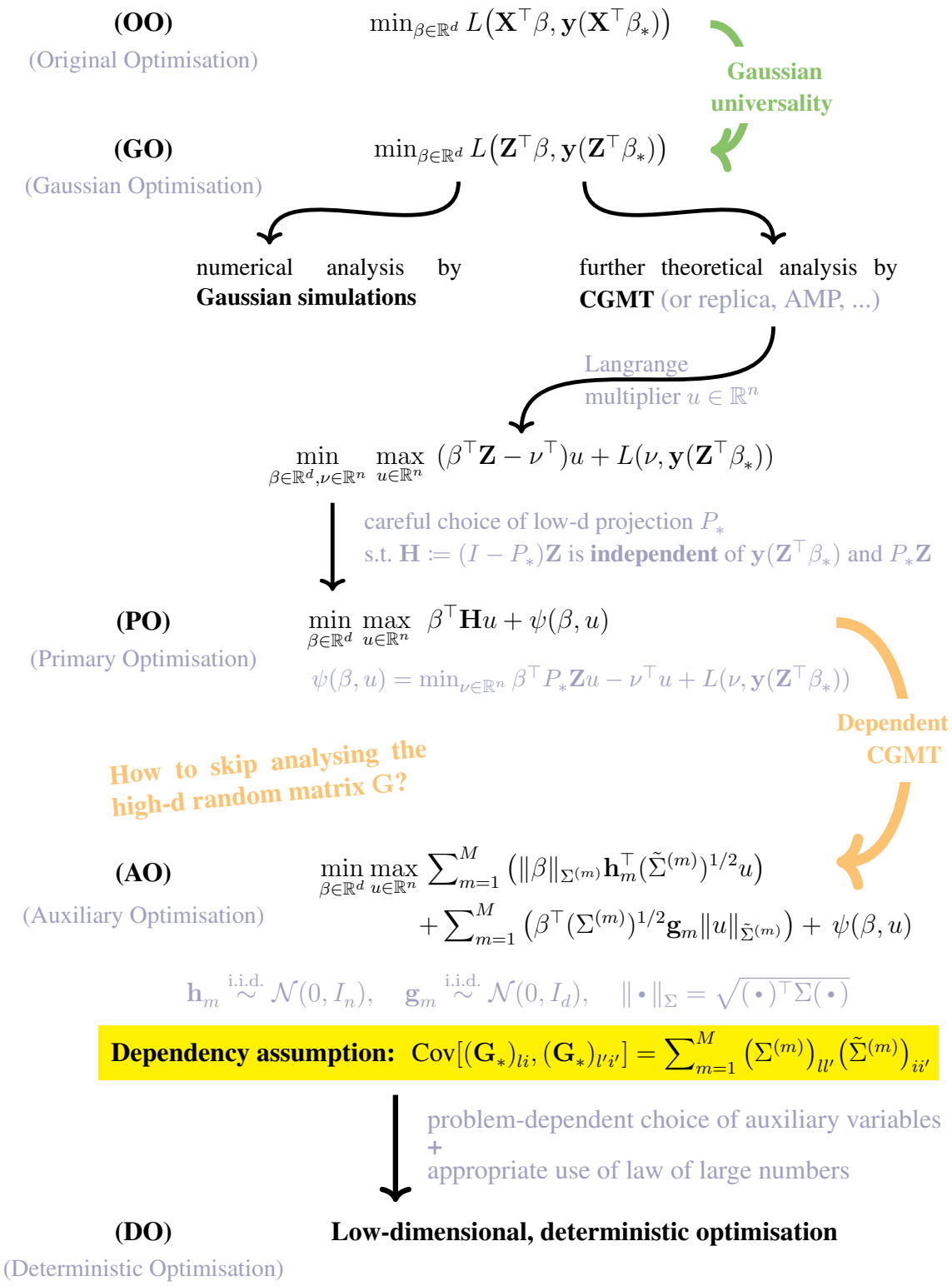


Figure 7.2: A diagrammatic illustration of the pipeline of high-dimensional risk analysis via Gaussian universality and our dependent CGMT.

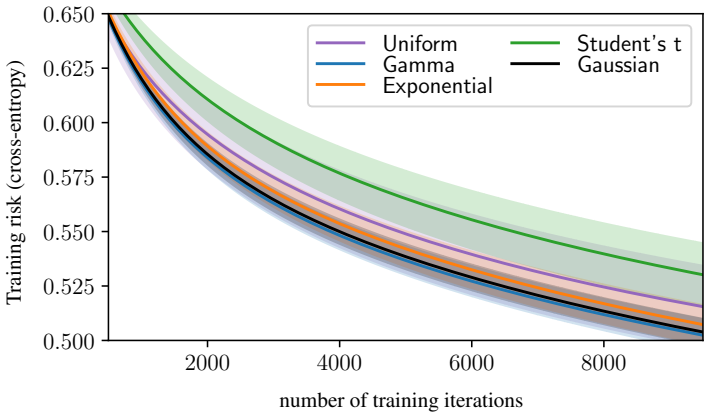


Figure 7.3: Figure 4 of Mallory, Huang, and Austern (2025): Initial training loss curves for the random permutation setup in Fig. 7.1 with $\rho_{\text{perm}} = 0.8$, $k = 11$ and learning rate $\text{LR} = 0.1$.

7.2 Stability analysis in stochastic optimisation

One intriguing observation of Mallory, Huang, and Austern (2025) is that in the same experiment setup, universality is observed for the global minimum of the risk (Figure 7.1) but *not* for the training trajectories (Figure 7.3). Specifically, we observe a discrepancy in training trajectories under the same set of hyperparameters (Figure 7.3), and a different learning rate is required for the Student's t and uniform setups to obtain convergence to the global minima. Underneath this observation is the crucial distinction that the global minimiser of a loss and the training iterates to optimise the same loss can be two different mathematical objects. The former takes the form $\hat{\beta}(X) \in \text{argmin}_{\beta \in \mathbb{R}^d} \text{loss}(\beta; X_1, \ldots, X_n)$; when the minimisers are unique, $\hat{\beta}(X)$ can be analysed directly via the loss function. The latter typically takes the form

$$\theta_T \;=\; f_{T,X,\xi_T} \circ f_{T-1,X,\xi_{T-1}} \circ \ldots \circ f_{1,X,\xi_1}(\theta_0)\;, \tag{7.6}$$

i.e. the composition of some $\mathbb{R}^d \to \mathbb{R}^d$ stochastic update functions f_{t,X,ξ_t}. The update functions are X-dependent, and $(\xi_t)_{t \le T}$ is some sequence of i.i.d random variables that represent per-step randomisation such as the use of stochastic mini-batches. The limit of θ_T as $T \to \infty$ may not exist, and even if it exists, it may not necessarily converge to $\hat{\beta}(X)$.

In this section, our choice of $f(X)$ in (1.1) is θ_T itself or some coordinate of θ_T. Proving the universality of θ_T, in general, can be a much more difficult task. As with the preceding chapters, we are faced with the obstacles of dependence and high-dimensionality, but now in the context of a composition of many stochastic functions:

- **Dependence.** Notice that, had the composition been formed by a sequence of deterministic functions, the asymptotic behaviour of (7.6) would be a well-studied problem in the dynamical systems literature (Collet and Eckmann, 2009). Alternatively, had

it been a composition of i.i.d. random functions, results from iterated random functions (Diaconis and Freedman, 1999) and Markov chains (Norris, 1998; Levin and Peres, 2017) would have been applicable. The dependence across all $f_{t;X;\xi_t}$, due to the presence of X, makes out-of-the-shelf applications of standard tools from those literatures challenging;

- **High-dimensionality.** In a machine learning setup, the update functions typically operate on a high-dimensional space $\mathbb{R}^d$. Even if one ignores X-dependence and studies (7.6) as a Markov chain, the problem of high dimensionality is a known challenging and ongoing area of research within e.g. the Markov Chain Monte Carlo community (Katafygiotis and Zuev, 2008; Girolami and Calderhead, 2011; Betancourt, 2017).

In this section, instead of tackling these challenges in full generality, we shall focus on two specific machine learning applications and briefly discuss how universality results and heuristics played a role in those analyses. Specifically, we review two of the authors' joint works, Glaser, Huang, and Gretton (2024) and Huang, Zhan, Ertekin, Orbanz, and Adams (2025), and highlight parts of their analyses where either the result or the heuristic of universality plays a role. In the former, universality is used to decouple the multi-step dependence across f_{t,X,ξ_t}'s with a small number of moment controls; in the latter, universality is used for one-step stability analysis of a high-dimensional gradient. As the nature of this section is a discussion, we will focus on sketching out the intuitions rather than presenting the full setups in both papers.

We emphasise that universality is not the key message of either work. We include a brief discussion of the key content of each work at the end of each subsection, and refer interested readers to both works for further reading.

7.2.1. Multi-step dependence decoupling in the contrastive divergence algorithm

Consider the problem of fitting the unknown natural parameter $\theta_* \in \mathbb{R}^p$ given n i.i.d. data $X_1, \dots, X_n$, drawn from an exponential family distribution (Brown, 1986; Wainwright et al., 2008)

$$p_\theta(dx) \;:=\; e^{\theta^\top \phi(x) - \log Z(\theta)} c(dx)\;, \qquad Z(\theta) \;:=\; \int_{\mathcal{X}} e^{\theta^\top \phi(x)} c(dx)\;.$$

Here, $\mathcal{X} \subseteq \mathbb{R}^d$ is the sample space, the function $\phi : \mathbb{R}^d \to \mathbb{R}^d$ is the sufficient statistic, and c is the base probability measure on $\mathcal{X}$. We also take d to be fixed throughout this subsection. A standard approach for estimating θ is by running the iterated update (7.6) with the maximum likelihood gradient update

$$f^{\mathrm{ML}}_{t,X}(\theta) \;:= \theta - \eta_t \Big(\frac{1}{n} \sum\nolimits_{i \le n} \phi(X_i) - \nabla \log Z(\theta) \Big)$$

$$= \theta - \eta_t \Big(\frac{1}{n} \sum_{i \le n} \phi(X_i) - \mathbb{E}_{X' \sim p_\theta}[\phi(X')] \Big) \, .$$

Here, $\eta_t > 0$ is the learning rate at step t. Note that this can be viewed as a finite-sample estimation of the population update

$$f_t^{\mathrm{pop}}(\theta) \; := \theta - \eta_t \big(\mathbb{E}_{X \sim p_{\theta_*}}[\phi(X)] - \mathbb{E}_{X' \sim p_\theta}[\phi(X')] \big) \, .$$

In many applications, however, the log-normaliser $\log Z(\theta)$ is not assumed to be admit a closed form. This allows for a flexible design of the model class, but prevents us from using the update scheme $f_{t,X}^{\mathrm{ML}}$.

The contrastive divergence (CD) algorithm (Hinton, 2002) is a popular method for fitting unnormalised models. In its most vanilla form, at each gradient update, CD simulates n i.i.d. Markov chains that target p_θ, initialised from $X_1, \ldots, X_n$ and run for m steps, to obtain the samples $\tilde{X}_1^{m,\theta}, \ldots, \tilde{X}_n^{m,\theta}$. Then, CD replaces the intractable gradient update $f_{t,X}^{\mathrm{ML}}$ by computing

$$f_{t,X}^{\mathrm{CD}}(\theta) \; := \theta - \eta_t \Big(\frac{1}{n} \sum_{i \le n} \phi(X_i) - \frac{1}{n} \sum_{i \le n} \phi(\tilde{X}_i^{m,\theta}) \Big) \, .$$

Let θ_T^{CD} be the final iterate of (7.6) under the CD updates by $f_{t,X}^{\mathrm{CD}}$. The goal is, informally, to obtain a high-probability bound or moment bound on the difference from the true parameter,

$$\| \theta_T^{\mathrm{CD}} - \theta_* \| \, . \tag{7.7}$$

If both the number of steps m and the number of data points n are large, under suitable conditions on the Markov chain sampling algorithms, CD yields a good approximation of the update f_t^{pop}. However in practice, since the Markov chains have to be run at every gradient step, m is never chosen to be large. This introduces a potential source of bias, which can affect the convergence rate of θ_T^{CD} to θ_*.

The prior work of Jiang et al. (2018) establishes an asymptotic $O(n^{-1/3})$ high probability bound on (7.7) for unnormalised exponential families for a finite m. In the work of Glaser, Huang, and Gretton (2024), one of our main goals was to obtain the parametric rate of $O(n^{-1/2})$-consistency (i.e. a high probability or moment bound on (7.7) with $O(n^{-1/2})$ error). Additionally, we aimed to obtain finite-sample moment bounds in (7.7) with explicit dependence on m, which allow one to interpret the interaction of m with other problem-dependent parameters e.g. the learning rate η_t and properties of p_{θ_*}.

One key technical challenge is the dependence of X_i's across $f_{t,X}^{\mathrm{CD}}$ for different t's, which manifest both explicitly through the data term $\frac{1}{n} \sum_{i \le n} \phi(X_i)$ and implicitly through the initialisations of the finite-length Markov chains. To address this, note that our goal is to prove a consistency result, which amounts to showing that the iterates produced by $f_{t,X}^{\mathrm{CD}}$ are close to the "oracle" iterates f_t^{pop}. As such, it suffices to *decouple* the dependence

across the $f_{t,X}^{\mathrm{CD}}$'s, i.e. to show that $f_{t,X}^{\mathrm{CD}}$'s are approximately independent functions.

The decoupling step, which appears in Section 4 of Glaser, Huang, and Gretton (2024) in the form of a tail condition, is achieved by universality. Notably, to show the independence of two generic variables, it does not suffice to consider only finitely many moments. In the case of the iterates of $f_{t,X}^{\mathrm{CD}}$, however, we are only concerned with functions of empirical averages of conditionally i.i.d. quantities, which allows for a conditional Gaussian approximation. In particular, we can achieve decoupling and thereby consistency by only assuming a ν-th moment control for $\nu > 2$ — analogous to the moment condition imposed in the earlier universality result of Theorem 4.1. Moreover, the more moments we can control, the stronger the decoupling result is and the sharper the rate we may obtain for CD. Section 4.1 of Glaser, Huang, and Gretton (2024) provides a range of rates for different fixed values of ν, and shows that under the sub-Gaussian tail condition assumed in Jiang et al. (2018), we can in fact achieve an $O(\sqrt{\log n}/\sqrt{n})$ bound on the L_2 norm of (7.7).

We do emphasise that the key messages and the many other results of Glaser, Huang, and Gretton (2024) are not tied to universality. For example, the regime discussed above, where the data X_i's are reused across iterates, is known as the *offline* CD. A substantial amount of work in Section 3 of Glaser et al. (2024) is devoted to the study of *online* CD, where data are not reused across iterates and where no decoupling is required. There, we show that one does achieve $O(n^{-1/2})$-consistency, and a CD iterate with Polyak-Ruppert averaging (Polyak and Juditsky, 1992) can even achieve an error bound that is close to the Cramér-Rao lower bound. Additionally, we also prove results for stochastic gradient descent versions of CD with and without replacement. The key tools behind these results are combinations of both known and new tools for stochastic optimisation analysis. Indeed even in the decoupling argument above, universality is not sufficient to yield the joint dependence control over *all* time steps, and the structure of the optimisation algorithm in CD plays a key role. We refer interested readers to the paper for the full picture.

7.2.2. One-step high-d stability analysis for large-scale neural network solvers to the many-body Schrödinger equation

Consider the problem of finding the ground state wavefunction $\psi : \mathbb{R}^{3n} \to \mathbb{C}$ to the n-electron Schrödinger equation. One essentially seeks the eigenfunction ψ with the minimal eigenvalue E for the Hamiltonian operator H,

$$H\psi(\mathbf{x}) = E\psi(\mathbf{x}), \quad \mathbf{x} \coloneqq (x_1, \ldots, x_n) \in \mathbb{R}^{3n}\,, \tag{7.8}$$

subject to some additional physical constraints that we do not specify here for simplicity. The Hamiltonian H is given by $H\psi(\mathbf{x}) \coloneqq -\frac{1}{2}\Delta\psi(\mathbf{x}) + V(\mathbf{x})\psi(\mathbf{x})$, Δ is the Laplacian operator representing the kinetic energy and $V : \mathbb{R}^{3n} \to \mathbb{R}$ is the potential energy of the physical system.

For many practical physical systems of interest, (7.8) is solved numerically by finding the best solution within a parametrised class of functions $\{\psi_\theta \,|\, \theta \in \mathbb{R}^q\}$, e.g. a class of large neural networks (Hermann et al., 2020; Pfau et al., 2020; Li et al., 2022). A brute-force search of ψ_θ is computationally difficult, since ψ_θ's are functions living in a high-dimensional space. To circumvent this, one typically takes advantage of a nice physical property of the wavefunction that $p_\psi(\mathbf{x}) \coloneqq \frac{|\psi(\mathbf{x})|^2}{\langle\psi,\psi\rangle}$ gives the probability distribution of the n electrons. Here, $\langle f, g\rangle \coloneqq \int f(\mathbf{x})^* g(\mathbf{x})\, d\mathbf{x}$ denotes the complex inner product.

One class of methods that utilises p_ψ is Variational Monte Carlo (VMC), which seeks to solve the minimum eigenvalue problem of (7.8) by the optimisation

$$\operatorname*{argmin}_{\theta\in\mathbb{R}^q} \frac{\langle\psi_\theta, H\psi_\theta\rangle}{\langle\psi_\theta,\psi_\theta\rangle} \;=\; \operatorname*{argmin}_{\theta\in\mathbb{R}^q} \mathbb{E}_{\mathbf{X}\sim p_{\psi_\theta}}[E_{\mathrm{local};\psi_\theta}(\mathbf{X})]\;, \tag{7.9}$$

where $E_{\mathrm{local};\psi_\theta}(\mathbf{x}) \coloneqq H\psi_\theta(\mathbf{X})/\psi_\theta(\mathbf{X})$ is called the local energy. In the simplest case, the optimisation may be performed by a first-order method, which can be represented by a generic function $F_{\mathbf{x};\psi_\theta} \equiv F(\psi_\theta(\mathbf{x}), \Delta\psi_\theta(\mathbf{x})) \in \mathbb{R}^q$ as

$$\theta \;\mapsto\; f_t(\theta) \;=\; \theta \,-\, \mathbb{E}_{\mathbf{X}\sim p_{\psi_\theta}}[F_{\mathbf{X};\psi_\theta}]\;.$$

Here we again use f_t from the notation of the composition (7.6) at the start of Section 7.2. Notably, the expectation formulation above converts the expensive integral over the entire space into an expectation, which can then be approximated by Monte Carlo averages computed on finitely many samples from p_{ψ_θ}.

A key distinction of VMC from standard machine learning problems is that there are *no* training data, and Monte Carlo samples are generated on-the-fly at every gradient step. More precisely, let ψ_θ being the iterate from the last training step. The new training step is performed by first running N MCMC chains for m steps, with p_{ψ_θ} as the target distribution, and initialised at the samples from the last iterate. Denote $p_{\psi_\theta}^{(m)}$ as the distribution of one of these m-th step MCMC chain (conditionally on initialisation and on ψ_θ). Then the obtained samples as $\mathbf{X}_1, \ldots, \mathbf{X}_N$ are i.i.d. drawn from $p_{\psi_\theta}^{(m)}$, and the original (OG) gradient update rule is given by

$$\theta \mapsto f_t^{(\mathrm{OG})}(\theta) \;=\; \theta - \delta\theta^{(\mathrm{OG})}\;,\quad \delta\theta^{(\mathrm{OG})} \coloneqq \frac{1}{N}\sum\nolimits_{i\le N} F_{\mathbf{X}_i;\psi_\theta}\;. \tag{7.10}$$

The main goal of the joint work of Huang, Zhan, Ertekin, Orbanz, and Adams (2025) is to examine a large class of difficult symmetries that arise naturally in modelling ψ_θ in infinite periodic crystalline solids, and to examine the strengths and limitations of vari-

ous standard symmetrisation techniques in machine learning. Among those, one of the approaches considered is *in-training data augmentation*. This is the same data augmentation (DA) approach discussed in Chapter 6. Perhaps surprisingly, however, we observe that DA can lead to variance inflation and instability in VMC even for empirical averages, in contrast to the observations in Section 6.4.

The main difference arises from a unique computational-statistical tradeoff in VMC. Specifically, the per-step training cost of VMC consists of two parts: (i) the sampling cost of each m-step MCMC chain, c_{sample}, and (ii) the cost of evaluating the gradient of a large neural network on each sample, c_{grad}. In particular, (i) typically involves computing only $\partial_x \psi_\theta$, whereas (ii) involves at least $\partial_\theta \psi_\theta$ and $\partial_x^2 \psi_\theta$ (due to the presence of Laplacian in the training objective). The computational cost of the original update rule (7.10) is therefore $N\, c_{\text{sample}} + N\, c_{\text{grad}}$. To implement data augmentation, one may sample N' DA samples $\mathbf{X}_1, \ldots, \mathbf{X}_{N'} \sim p_{\psi_\theta}^{(m)}$ and draw $N'k$ i.i.d. augmentations $\mathbf{g}_{i,j}$ (as described in Chapter 6), which gives the update rule

$$\theta \mapsto \theta - \delta\theta^{(\text{DA})} \;, \quad \delta\theta^{(\text{DA})} \;:=\; \frac{1}{N'k} \sum\nolimits_{i \le N'} \sum\nolimits_{j \le k} F_{\mathbf{g}_{i,j}(\mathbf{X}_i);\psi_\theta} \;. \tag{7.11}$$

This leads to a per-step computational cost of $N'\, c_{\text{sample}} + N'k\, c_{\text{grad}}$. Since the main bottleneck in training is the computational cost (e.g. the number of GPUs to parallelise sampling over), for a fair comparison, we need to choose N' and k such that

$$N\, c_{\text{sample}} + N\, c_{\text{grad}} \;\approx\; N'\, c_{\text{sample}} + N'k\, c_{\text{grad}} \;.$$

This, for example, necessitates $N' < N$ as the number of augmentations is typically $k \ge 2$. In the neural network considered in Huang, Zhan, Ertekin, Orbanz, and Adams (2025), we find that $c_{\text{grad}} \gg c_{\text{sample}}$ in general, and restrict our attention to the case $N' = N/k$, assuming the divisibility of N by k. In particular, the augmented batch size is the same as the original batch size N, which is in stark contrast to the augmented batch size Nk considered in Section 6.4. Note that this arises because in the standard ML setup considered in Chapter 6, the main bottleneck is the size N of the finitely many real-life data, and augmentations are assumed to be computationally cheap; here, the main bottleneck is computational rather than number of samples, and each augmentation can incur a similar computational cost to that of acquiring a new data point.

An immediate consequence of the choice $N' = N/k$ is that each step of (7.11) now involves an average of fewer i.i.d. summands. As expected, DA here leads to variance inflation and instability of the gradient, even for an empirical average, as observed in the normalised variance plot in Figure 7.4. Huang et al. (2025) then proceeds to validate empirically that the performance of DA does not improve from that of the original update; see the paper for the full experiment setup.

To derive a rigorous theoretical statement for this effect, one faces the additional tech-

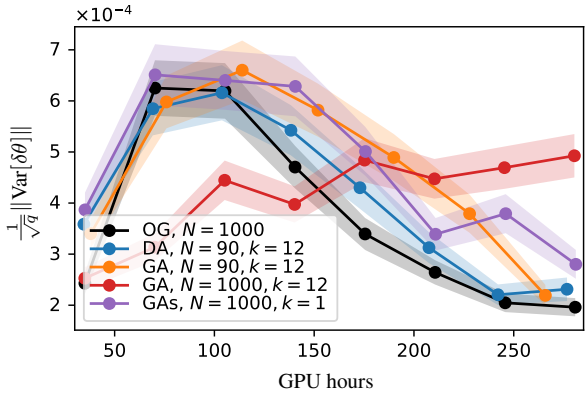


Figure 7.4: Figure 3 of Huang, Zhan, Ertekin, Orbanz, and Adams (2025): Normalised variance of differently symmetrized gradient updates against GPU hours.

nical challenge of analysing a high-dimensional empirical average, which is not expected to behave like a high-dimensional Gaussian vector in general. Nevertheless, universality results in the preceding chapters tell us that one may expect Gaussianity if the property of interest is a suitably stable univariate function of the high-dimensional inputs. Here, since we seek to analyse stability, given a random $\mathbb{R}^p$-valued gradient update $\delta\theta$, we may focus on the object

$$\max_{l \leq p} |(\delta\theta)_l - \mathbb{E}[(\delta\theta)_l]| \,, \tag{7.12}$$

i.e. the maximum deviation from the mean of the gradient update $\delta\theta$ over the p coordinates. Since $\delta\theta^{(\mathrm{DA})}$ takes the form of an i.i.d. augmented average, an analogous result to Proposition 6.13 applies. In Appendix D of Huang, Zhan, Ertekin, Orbanz, and Adams (2025), we directly used the finite-sample bound of the high-dimensional CLT from Chernozhukov et al. (2017) to show that the limiting distribution of (7.12) is completely captured by $\mathrm{Var}[\delta\theta]$, provided that $p = o(\exp(N^{1/7}))$ for (7.10) and $p = o(\exp((N')^{1/7}))$ for (7.11). Under an exact invariance condition, we further showed that $\mathrm{Var}[\delta\theta^{(\mathrm{DA})}] \gtrsim \mathrm{Var}[\delta\theta^{(\mathrm{OG})}]$, where $\gtrsim$ is the Loewner order of non-negative matrices; under an approximate invariance condition, we provided a corresponding error bound. These results helped to provide theoretical support for the instability observed for DA in VMC.

We conclude by stressing that the main messages and results of Huang, Zhan, Ertekin, Orbanz, and Adams (2025) are not tied to universality. A substantial amount of work in Huang et al. (2025) is devoted to analysing the type of symmetries that arise in a many-body wavefunction for crystalline solids, and the possible symmetrisation techniques one may employ to a large-scale VMC neural network. Besides in-training data augmentation, we also examine a group averaging and a smooth canonicalisation approach both during training and inference. A main discovery of Huang et al. (2025) is that post-hoc group averaging can be a simple and effective method for substantially improving the chemical accuracy of the learned wavefunction within the same amount of computational resources: We validated this finding across graphene, lithium hydride (LiH) and metallic

lithium systems, with a $10\times$ computational speedup in the LiH system compared to state-of-the-arts neural network solvers. As these results are obtained on large neural network solvers parallelised over many GPUs, a significant amount of work in Huang et al. (2025) is computational rather than theoretical in nature. We refer interested readers to the paper for the full picture.

Chapter 8

Conclusion and future directions

Consider, for one last time, an empirical average of i.i.d. zero-mean 1d variables,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \;=\; \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}.$$

One way to interpret the CLT is that the behaviour of the $\mathbb{R}^d$ random vector $X = (X_1, \ldots, X_n)^\intercal$, when viewed through the 1d projection $(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})$, is completely described by the first two moments of X and therefore substitutable by a Gaussian vector.

In the same spirit, Gaussian universality can be viewed as a set of general Gaussian approximation results on the input space of some function f: It says that the behaviour of a sequence of random vectors $(X_1, \ldots, X_n)$, when viewed through some possibly non-linear and complicated function f, is also completely captured by the first two moments. Various examples of such f's have been examined in this thesis in the context of high-dimensionality and dependence, and universality has been used to provide theoretical and practical intuitions in each of them. We have also developed a general set of universality results for the L_2 space spanned by degree-m polynomials of high-dimensional random vectors, and provide nearly optimal upper and lower bounds to show how the approximation error deteriorates as m grows.

The applications considered in this thesis are merely a subset of the rapidly expanding body of work on universality, where a constant quest is to push the boundary on the types of high-dimensional estimators and dependence structures one may accommodate. For instance, Section 5.4 and the references therewithin are part of a growing body of works that use Gaussian approximations to characterise the limiting behaviours of subgraph count statistics by Wiener chaos. An interesting extension is to consider whether Gaussian universality may be a useful perspective for unifying and extending results in other network-related statistics and in dynamic graph models such as the preferential attachment models (Barabási and Albert, 1999; Albert and Barabási, 2002; Berger et al., 2014; Peköz et al., 2017; Bloem-Reddy and Orbanz, 2018). On the other hand, the problem (7.6) of analysing the iterates of random functions in a high-dimensional space, is partially solved by a recent line of work on the limiting spectral behaviour of products of many large random matrices (Hanin and Paouris, 2021; Hanin and Jiang, 2025), and such

results have seen applications in the analysis of deep neural networks (Hanin and Nica, 2020; Li et al., 2021; Noci et al., 2023; Favaro et al., 2025).

An open question remains as to whether and when Gaussian universality can break in practical machine learning applications. The simple maximum example (2.4) in Section 2.1 and our lower bound construction in Section 4.5 provide theoretical examples, but neither of them yields an immediate connection to a practical machine learning setup. Meanwhile, several lines of works have emerged in the literature to examine negative results on universality. For example, Pesce et al. (2023) show, both theoretically and empirically, that the behaviour of a high-dimensional generalised linear model changes when the input Gaussian mixture data are replaced by their Gaussian surrogates. We do note that, in view of the discussion in Section 2.2, Gaussian universality does not always replace the input data of the estimator directly by Gaussians, and the Gaussian mixture case may still be viewed as a type of universality with appropriately transformed data. On the other hand, universality with respect to heavy-tailed rather than Gaussian distributions is also observed in neural network layer weights and training iterates under stochastic gradient descent (Martin and Mahoney, 2020; Hodgkinson and Mahoney, 2021; Gurbuzbalaban et al., 2021).

For systems that do exhibit Gaussian universality, another vital question is whether these systems can be too restrictive or harmful in practice. Broadly speaking, Gaussian universality results apply to systems in which the stochasticity is approximately Gaussian-generated. One particular example of such systems is the large family of deep generative models built on transformations of Gaussian distributions, such as variational autoencoders, generative adversarial networks and diffusion models. In this literature, an emerging line of work (Salmona et al., 2022; Pandey et al., 2024; Tam and Dunson, 2025; Ghane et al., 2025) has shown that the use of Gaussian distributions can in fact restrict the representative power of these models, and cause them to struggle with modelling multimodal and heavy-tailed data. Ghane et al. (2025) also empirically observe that, for the same diffusion model with a restrictive representation power, a notion of Gaussian universality is satisfied by the test risk.

Another interesting mathematical problem is whether explicit and sharp constants may be obtained in the universality approximation bounds. Such constants, if known, are useful in practice, as it allows for an explicit computation of the approximation error. For empirical averages, considerable efforts have been made to obtain and sharpen these constants; see e.g. Shevtsova (2011) and references therewithin, and see Austern and Mackey (2022) on how such explicit constants facilitate practically computable and efficient concentration bounds. To obtain explicit constants for the results of this thesis, one may use the explicit numeric constant in the proof of the anti-concentration inequal-

ity from Carbery and Wright (2001) and combine it with our proofs. However, since the constant obtained would apply to a large class of statistics, it is unlikely to be sharp for specific applications. It remains an interesting question whether one may obtain sharp and explicit constants if, for example, one can restrict our attention to particular classes of U-statistics, or if suitable tail conditions or coordinate-wise dependence conditions are imposed on the data distribution.

In view of these open questions, an interesting direction of future work is to build a better understanding of the limitations and failure modes of Gaussian universality. One may also ask if a more general universality framework is necessary, in order to obtain a better theory for the behaviours of modern statistical and machine learning algorithms.

References

M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

D. Akhtiamov, D. Bosch, R. Ghane, K. N. Varma, and B. Hassibi. A novel Gaussian min-max theorem and its applications. *arXiv preprint arXiv:2402.07356*, 2024a.

D. Akhtiamov, R. Ghane, and B. Hassibi. Regularized linear regression for binary classification. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 202–207. IEEE, 2024b.

R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74 (1):47, 2002.

H. Alzer. On Ramanujan's double inequality for the gamma function. *Bull. London Math. Soc.*, 35(5): 601–607, 2003.

C.-G. Ambrozie. Multivariate truncated moments problems and maximum entropy. *Anal. Math. Phys.*, 3 (2):145–161, 2013.

L. Aolaritei, S. Shafieezadeh-Abadeh, and F. Dörfler. The performance of Wasserstein distributionally robust M-estimators in high dimensions. *arXiv preprint arXiv:2206.13269*, 2022.

M. Austern and L. Mackey. Efficient concentration with Gaussian approximation. *arXiv preprint arXiv:2208.09922*, 2022.

M. Austern and P. Orbanz. Limit theorems for distributions invariant under a group of transformations. *Ann. Statist.*, 50(4):1960–1991, 2022.

J. Ba, M. Erdogdu, T. Suzuki, D. Wu, and T. Zhang. Generalization of two-layer neural networks: An asymptotic viewpoint. In *International conference on learning representations*, 2019.

R. Balestriero, L. Bottou, and Y. LeCun. The effects of regularization and data augmentation are class dependent. In *Conference on Neural Information Processing Systems*, 2022a.

R. Balestriero, I. Misra, and Y. LeCun. A data-augmentation is worth a thousand samples: Exact quantification from analytical augmented sample moments. In *Conference on Neural Information Processing Systems*, 2022b.

A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

A. Basak, N. Cook, and O. Zeitouni. Circular law for the sum of random permutation matrices. 2018.

M. Bayati, M. Lelarge, and A. Montanari. Universality in polytope phase transitions and message passing algorithms. *Ann. Appl. Probab.*, 25(2):753 – 822, 2015.

V. Bentkus and F. Götze. Optimal bounds in non-Gaussian limit theorems for U-statistics. *Ann. Probab.*, 27(1):454–521, 1999.

N. Berger, C. Borgs, J. T. Chayes, and A. Saberi. Asymptotic behavior and distributional limits of preferential attachment graphs. *The Annals of Probability*, 42(1):1–40, 2014. doi: 10.1214/12-AOP755.

S. Bernstein. Sur l'extension du théorème limite du calcul des probabilités aux sommes de quantités dépendantes. *Math. Ann.*, 97:1–59, 1927.

M. Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

B. B. Bhattacharya, S. Das, S. Mukherjee, and S. Mukherjee. Asymptotic distribution of random quadratic forms. *arXiv:2203.02850*, 2022.

B. B. Bhattacharya, A. Chatterjee, and S. Janson. Fluctuations of subgraph counts in graphon based random graphs. *Combin. Probab. Comput.*, 32(3):428–464, 2023.

P. Billingsley. *Probability and measure*. John Wiley & Sons, New York, 1995.

Y. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete multivariate analysis: Theory and practice*. Springer Science & Business Media, Berlin, 1974.

T. Blasco, J. S. Sánchez, and V. García. A survey on uncertainty quantification in deep learning for financial time series prediction. *Neurocomputing*, 576:127339, 2024.

B. Bloem-Reddy and P. Orbanz. Random-walk models of network formation and sequential Monte Carlo methods for graphs. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(5): 871–898, 2018.

D. A. Bodenham and N. M. Adams. A comparison of efficient approximations for a weighted sum of chi-squared random variables. *Statist. Comput.*, 26(4):917–928, 2016.

R. C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probab. Surveys*, 2:107–114, 2005.

T. Brailovskaya and R. van Handel. Universality and sharp matrix concentration inequalities. *arXiv preprint arXiv:2201.05142*, 2022.

W. Brock, J. Lakonishok, and B. LeBaron. Simple technical trading rules and the stochastic properties of stock returns. *J. Fin.*, 47(5):1731–1764, 1992.

L. D. Brown. *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*, volume 9 of *Lecture Notes–Monograph Series*. Institute of Mathematical Statistics, 1986.

D. L. Burkholder. Martingale transforms. *Ann. Math. Statist.*, 37(6):1494–1504, 1966.

T. Cacoullos. On upper and lower bounds for the variance of a function of a random variable. *Ann. Probab.*, 10(3):799–809, 1982.

F. Caravenna, R. Sun, and N. Zygouras. Polynomial chaos and scaling limits of disordered systems. *J. Eur. Math. Soc.*, 19(1):1–65, 2016.

A. Carbery and J. Wright. Distributional and l^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Math. Res. Lett.*, 8(3):233–248, 2001.

A. Chatterjee, S. Dan, and B. B. Bhattacharya. Higher-order graphon theory: Fluctuations, degeneracies, and inference. *arXiv preprint arXiv:2404.13822*, 2024.

S. Chatterjee. A generalization of the Lindeberg principle. *Ann. Probab.*, 34(6):2061–2076, 2006.

N. S. Chatterji, P. M. Long, and P. L. Bartlett. The interplay between implicit bias and benign overfitting in two-layer linear networks. *J. Mach. Learn. Res.*, 23(1):12062–12109, 2022.

L. H. Chen, L. Goldstein, and Q.-M. Shao. *Normal approximation by Stein's method*, volume 2. Springer, 2011.

S. Chen, E. Dobriban, and J. H. Lee. A group-theoretic framework for data augmentation. *J. Mach. Learn. Res.*, 21(245):1–71, 2020.

S. X. Chen and Y.-L. Qin. A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.*, 38(2):808 – 835, 2010.

V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.*, 41(6):2786–2819, 2013.

V. Chernozhukov, D. Chetverikov, and K. Kato. Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 45(4):2309 – 2352, 2017.

K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 2606–2615, 2016.

P. Collet and J.-P. Eckmann. *Iterated maps on the interval as dynamical systems*. Springer Science & Business Media, 2009.

G. Constantine and T. Savits. A multivariate Faa di Bruno formula with applications. *Trans. Amer. Math. Soc.*, 348(2):503–520, 1996.

T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to algorithms*. MIT press, 2009.

J. D. Cryer. *Time series analysis*, volume 286. Duxbury Press Boston, 1986.

R. Cuppens. *Decomposition of multivariate probabilities*, volume 29. Academic Press, Cambridge, 1975.

Y. Dandi, L. Stephan, F. Krzakala, B. Loureiro, and L. Zdeborová. Universality laws for Gaussian mixtures in generalized linear models. *Advances in Neural Information Processing Systems*, 36:54754–54768, 2023.

T. Dao, A. Gu, A. Ratner, V. Smith, C. De Sa, and C. Ré. A kernel theory of modern data augmentation. In *International Conference on Machine Learning*, pages 1528–1537, 2019.

J. Davidson. A central limit theorem for globally nonstationary near-epoch dependent functions of mixing processes. *Econometric Theory*, 8(3):313–329, 1992.

P. De Jong. A central limit theorem for generalized multilinear forms. *Journal of Multivariate Analysis*, 34(2):275–289, 1990.

A. de Moivre. *Approximatio ad summam terminorum binomii .a C b/n in seriem expansi*. 1733.

H. Deng and C.-H. Zhang. Beyond Gaussian approximation: Bootstrap for maxima of sums of independent random vectors. *Ann. Statist.*, 48(6):3643–3671, 2020.

M. Denker. *Asymptotic distribution theory in nonparametric statistics*. Springer, Berlin, 1985.

S. W. Dharmadhikari, V. Fabian, and K. Jogdeo. Bounds on the moments of martingales. *Ann. Math. Statist.*, 39(5):1719 – 1723, 1968a.

S. W. Dharmadhikari, V. Fabian, and K. Jogdeo. Bounds on the moments of martingales. *Ann. Math. Statist.*, pages 1719–1723, 1968b.

O. Dhifallah and Y. Lu. On the inherent regularization effects of noise injection during training. In *International Conference on Machine Learning*, pages 2665–2675. PMLR, 2021.

P. Diaconis and D. Freedman. Iterated random functions. *SIAM Rev.*, 41(1):45–76, 1999.

P. Diaconis and S. Janson. Graph limits and exchangeable random graphs. *Rend. Math.*, VII(28):33–61, 2008.

C. Döbler and G. Peccati. Quantitative de jong theorems in any dimension. *Electronic Journal of Probability*, 22(2):1–35, 2017.

C. Döbler and G. Peccati. Quantitative CLTs for symmetric U-statistics using contractions. *Electronic Journal of Probability*, 24:1–43, 2019. doi: 10.1214/19-EJP307.

C. Döbler, M. J. Kasprzak, and G. Peccati. Functional convergence of sequential U-processes with size-dependent kernels. *The Annals of Applied Probability*, 32(1):551–601, 2022. doi: 10.1214/21-AAP1698.

E. Dobriban and S. Liu. Asymptotics for sketching in least squares regression. *Advances in Neural Information Processing Systems*, 32, 2019.

D. L. Donoho, A. Maleki, and A. Montanari. Message-passing algorithms for compressed sensing. *Proc. Nat. Acad. Sci.*, 106(45):18914–18919, 2009.

P. Duchesne and P. L. De Micheaux. Computing the distribution of quadratic forms: Further comparisons between the liu–tang–zhang approximation and exact methods. *Comput. Statist. Data Anal.*, 54(4): 858–862, 2010.

S. Favaro, B. Hanin, D. Marinucci, I. Nourdin, and G. Peccati. Quantitative CLTs in deep neural networks. *Probab. Theory Related Fields*, pages 1–45, 2025.

S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy. A survey of data augmentation approaches for NLP. In *Findings of Assoc. Comput. Linguist.*, pages 968–988, 2021.

D. Ferger. Moment inequalities for U-statistics with degeneracy of higher order. *Sankhya A*, pages 142–148, 1996.

A. Filippova. Mises' theorem on the asymptotic behavior of functionals of empirical distribution functions and its statistical applications. *Theory Probab. Appl.*, 7(1):24–57, 1962.

H. Fischer. *A history of the central limit theorem: from classical to modern probability theory*, volume 4. Springer, 2011.

T. Freiesleben and T. Grote. Beyond generalization: a theory of robustness in machine learning. *Synthese*, 202(4):109, 2023.

H. Gao and X. Shao. Two sample testing in high dimension via maximum mean discrepancy. *J. Mach. Learn. Res*, 24(304):1–33, 2023.

R. E. Gaunt. Stein's method for functions of multivariate normal random variables. 2020.

R. E. Gaunt and H. Sutcliffe. Improved bounds in Stein's method for functions of multivariate normal random vectors. *J. Theoret. Probab.*, pages 1–29, 2023.

F. Gerace, F. Krzakala, B. Loureiro, L. Stephan, and L. Zdeborová. Gaussian universality of perceptrons with random labels. *Phys. Rev. E*, 109(3):034305, 2024.

R. Ghane, A. Bao, D. Akhtiamov, and B. Hassibi. Concentration of measure for distributions generated via diffusion models. *arXiv preprint arXiv:2501.07741*, 2025.

B. Giovanola and S. Tiribelli. Beyond bias and discrimination: redefining the ai ethics principle of fairness in healthcare machine-learning algorithms. *AI & society*, 38(2):549–563, 2023.

M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B*, 73(2):123–214, 2011.

P. Glaser, K. H. Huang, and A. Gretton. Near-optimality of contrastive divergence algorithms. *Advances in Neural Information Processing Systems*, 37:91036–91090, 2024.

Y. Gordon. Some inequalities for Gaussian processes and applications. *Israel J. Math.*, 50:265–289, 1985.

J. Gorham and L. Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning*, pages 1292–1301. PMLR, 2017.

F. Götze and A. Tikhomirov. Asymptotic expansions in non-central limit theorems for quadratic forms. *J. Theoret. Probab.*, 18(4):757–811, 2005.

F. Götze and A. Y. Zaitsev. Explicit rates of approximation in the CLT for quadratic forms. *Ann. Probab.*, 42(1):354–397, 2014.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *J. Mach. Learn. Res*, 13(1):723–773, 2012.

N. Grigorevskii and I. S. Shiganov. Some modifications of the dudley metric. *Zapiski Nauchnykh Seminarov POMI*, 61:17–24, 1976.

J. Grimmer, M. E. Roberts, and B. M. Stewart. Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, 24(1):395–419, 2021.

M. Gurbuzbalaban, U. Simsekli, and L. Zhu. The heavy-tail phenomenon in SGD. In *International Con-*

ference on Machine Learning, pages 3964–3975. PMLR, 2021.

L. Haan and A. Ferreira. *Extreme value theory: an introduction*, volume 3. Springer, 2006.

P. Hall. *The bootstrap and Edgeworth expansion*. Springer Science & Business Media, 2013.

Q. Han and Y. Shen. Universality of regularized regression estimators in high dimensions. *Ann. Statist.*, 51(4):1799–1823, 2023.

B. Hanin and T. Jiang. Global universality of singular values in products of many large random matrices. *arXiv preprint arXiv:2503.07872*, 2025.

B. Hanin and M. Nica. Products of many large random matrices and gradients in deep neural networks. *Comm. Math. Phys.*, 376(1):287–322, 2020.

B. Hanin and G. Paouris. Non-asymptotic results for singular values of Gaussian matrix products. *Geom. Funct. Anal.*, 31(2):268–324, 2021.

Z. Harchaoui, L. Liu, and S. Pal. Asymptotics of discrete schrödinger bridges via chaos decomposition. *arXiv preprint arXiv:2011.08963*, 2020.

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Ann. Statist.*, 50(2):949–986, 2022.

J. Hermann, Z. Schätzle, and F. Noé. Deep-neural-network solution of the electronic schrödinger equation. *Nature Chemistry*, 12(10):891–897, 2020.

G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 2002.

J. Hladký, C. Pelekis, and M. Šileikis. A limit theorem for small cliques in inhomogeneous random graphs. *J. Graph Theory*, 97(4):578–599, 2021.

L. Hodgkinson and M. Mahoney. Multiplicative noise and heavy tails in stochastic optimization. In *International Conference on Machine Learning*, pages 4262–4274. PMLR, 2021.

H. Hu and Y. M. Lu. Universality laws for high-dimensional learning with random features. *IEEE Trans. Inf. Theory*, 69(3):1932–1964, 2022.

K. H. Huang, X. Liu, A. Duncan, and A. Gandy. A high-dimensional convergence theorem for U-statistics with applications to kernel-based testing. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3827–3918. PMLR, 2023.

K. H. Huang, M. Austern, and P. Orbanz. Gaussian universality for approximately polynomial functions of high-dimensional data. *arXiv:2403.10711*, 2024.

K. H. Huang, N. Zhan, E. Ertekin, P. Orbanz, and R. P. Adams. Diagonal symmetrization of neural network solvers for the many-electron schrödinger equation. *arXiv preprint arXiv:2502.05318*, 2025.

I. Ibragimov. Independent and stationary sequences of random variables. *Wolters, Noordhoff Pub.*, 1975.

S. Janson and K. Nowicki. The asymptotic distributions of generalized U-statistics with applications to random graphs. *Probab. Theory Related Fields*, 90:341–375, 1991.

A. Javanmard and M. Soltanolkotabi. Precise statistical analysis of classification accuracies for adversarial training. *Ann. Statist.*, 50(4):2127–2156, 2022.

B. Jiang, T.-Y. Wu, Y. Jin, and W. H. Wong. Convergence of contrastive divergence algorithm in exponential family. *Ann. Statist.*, 46(6A):3067–3098, 2018.

L. P. Kadanoff. Scaling and universality in statistical physics. *Physica A: Statistical Mechanics and its Applications*, 163(1):1–14, 1990.

O. Kallenberg. *Foundations of Modern Probability*. Springer, 2nd edition, 2001.

L. S. Katafygiotis and K. M. Zuev. Geometric insight into the challenges of solving high-dimensional reliability problems. *Probabilistic Engineering Mechanics*, 23(2-3):208–218, 2008.

G. Kaur and A. Röllin. Higher-order fluctuations in dense random graph models. *Electron. J. Probab.*, 26: 1–36, 2021.

A. Knowles and J. Yin. Anisotropic local laws for random matrices. *Probab. Theory Related Fields*, 169: 257–352, 2017.

S. B. Korada and A. Montanari. Applications of the lindeberg principle in communications and statistical learning. *IEEE Trans. Inf. Theory*, 57(4):2440–2450, 2011.

S. Lahiry and P. Sur. Universality in block dependent linear models with applications to nonlinear regression. *IEEE Trans. Inf. Theory*, 2024.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 1991.

A. J. Lee. *U-statistics: Theory and Practice*. Routledge, Milton Park, 1990.

A. Leucht and M. H. Neumann. Dependent wild bootstrap for degenerate u-and V-statistics. *J. Multivariate Anal.*, 117:257–280, 2013.

D. A. Levin and Y. Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

M. Li, M. Nica, and D. Roy. The future is log-Gaussian: Resnets and their infinite-depth-and-width limit at initialization. *Advances in Neural Information Processing Systems*, 34:7852–7864, 2021.

X. Li, Z. Li, and J. Chen. Ab initio calculation of real solids via neural network ansatz. *Nature Communications*, 13(1):7895, 2022.

J. W. Lindeberg. Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung. *Math. Z.*, 15(1):211–225, 1922.

Q. Liu, J. Lee, and M. Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 276–284, 2016.

B. Loureiro, G. Sicuro, C. Gerbelot, A. Pacco, F. Krzakala, and L. Zdeborová. Learning Gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions. *Advances in Neural Information Processing Systems*, 34:10144–10157, 2021.

C. Lyle, M. van der Wilk, M. Kwiatkowska, Y. Gal, and B. Bloem-Reddy. On the benefits of invariance in neural networks. In *Conference on Neural Information Processing Systems: Workshop on Machine Learning with Guarantees*, 2019.

J. R. Magnus. The moments of products of quadratic forms in normal variables. *Statistica Neerlandica*, 32:201–210, 1978.

M. E. Mallory, K. H. Huang, and M. Austern. Universality of high-dimensional logistic regression and a novel cgmt under dependence with applications to data augmentation. *arXiv preprint arXiv:2502.15752*, 2025.

V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

C. H. Martin and M. W. Mahoney. Heavy-tailed universality predicts trends in test accuracies for very large pre-trained deep neural networks. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 505–513. SIAM, 2020.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Comm. Pure Appl. Math.*, 75(4):667–766, 2022.

M. Mézard, G. Parisi, and M. A. Virasoro. *Spin glass theory and beyond: An Introduction to the Replica*

Method and Its Applications, volume 9. World Scientific Publishing Company, 1987.

F. Mignacco, F. Krzakala, Y. Lu, P. Urbani, and L. Zdeborova. The role of regularization in classification of high-dimensional noisy Gaussian mixture. In *International conference on machine learning*, pages 6874–6883. PMLR, 2020.

A. Montanari and P.-M. Nguyen. Universality of the elastic net error. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2338–2342. IEEE, 2017.

A. Montanari and B. N. Saeed. Universality of empirical risk minimization. In *Conference on Learning Theory*, pages 4310–4312. PMLR, 2022.

A. Montanari, F. Ruan, B. Saeed, and Y. Sohn. Universality of max-margin classifiers. *arXiv preprint arXiv:2310.00176*, 2023.

E. Mossel, R. O'Donnell, and K. Oleszkiewicz. Noise stability of functions with low influences: invariance and optimality. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*, pages 21–30. IEEE, 2005.

E. Mossel, R. O'Donnell, and K. Oleszkiewicz. Noise stability of functions with low influences: invariance and optimality. 2010.

A. Müller. Integral probability metrics and their generating classes of functions. *Adv. Appl. Probab.*, 29 (2):429–443, 1997.

L. Noci, C. Li, M. Li, B. He, T. Hofmann, C. J. Maddison, and D. Roy. The shaped transformer: Attention models in the infinite depth-and-width limit. *Advances in Neural Information Processing Systems*, 36: 54250–54281, 2023.

J. R. Norris. *Markov chains*. Number 2. Cambridge university press, 1998.

I. Nourdin. Lectures on Gaussian approximations with Malliavin calculus. In *Séminaire de Probabilités XLV*, pages 3–89. Springer, Nancy, 2013.

I. Nourdin and G. Peccati. Universal Gaussian fluctuations of non-hermitian matrix ensembles: from weak convergence to almost sure CLTs. *ALEA*, 7:341–375, 2010. Electronic.

I. Nourdin and G. Peccati. *Normal approximations with Malliavin calculus: from Stein's method to universality*, volume 192. Cambridge University Press, Cambridge, 2012.

I. Nourdin and G. Peccati. The optimal fourth moment theorem. *Proc. Amer. Math. Soc.*, 143(7):3123–3133, 2015.

I. Nourdin, G. Peccati, and G. Reinert. Second order Poincaré inequalities and CLTs on wiener space. *Journal of Functional Analysis*, 257(2):593–609, 2009.

I. Nourdin, G. Peccati, and G. Reinert. Invariance principles for homogeneous sums: Universality of Gaussian wiener chaos. *The Annals of Probability*, 38(5):1947–1985, Sept. 2010. doi: 10.1214/10-AOP522.

D. Nualart and G. Peccati. Central limit theorems for sequences of multiple stochastic integrals. 2005.

M. Opper, O. Winther, et al. From naive mean field theory to the tap equations. *Advanced mean field methods: theory and practice*, pages 7–20, 2001.

K. Pandey, J. Pathak, Y. Xu, S. Mandt, M. Pritchard, A. Vahdat, and M. Mardani. Heavy-tailed diffusion models. *arXiv preprint arXiv:2410.14171*, 2024.

E. Peköz, A. Röllin, and N. Ross. Joint degree distributions of preferential attachment random graphs. *Advances in Applied Probability*, 49(2):368–387, 2017.

W. Peng, T. Coleman, and L. Mentch. Rates of convergence for random forests via generalized U-statistics. *Electron. J. Stat.*, 16(1):232–292, 2022.

L. Perez and J. Wang. The effectiveness of data augmentation in image classification using deep learning.

arXiv preprint arXiv:1712.04621, 2017.

L. Pesce, F. Krzakala, B. Loureiro, and L. Stephan. Are Gaussian data all you need? The extents and limits of universality in high-dimensional generalized linear estimation. In *International Conference on Machine Learning*, pages 27680–27708. PMLR, 2023.

D. Pfau, J. S. Spencer, A. G. Matthews, and W. M. C. Foulkes. Ab initio solution of the many-electron schrödinger equation with deep neural networks. *Phys. Rev. Res.*, 2(3):033429, 2020.

N. S. Pillai and J. Yin. Universality of covariance matrices. *The Annals of Applied Probability*, 24(3):935 – 1001, 2014.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855, 1992.

A. Ramdas. *Computational and statistical advances in testing and learning*. PhD thesis, Carnegie Mellon University, 2015.

A. Ramdas, S. J. Reddi, B. Póczos, A. Singh, and L. Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015.

C. R. Rao, C. R. Rao, M. Statistiker, C. R. Rao, and C. R. Rao. *Linear statistical inference and its applications*, volume 2. John Wiley & Sons, New York, 1973.

S. Reddi, A. Ramdas, B. Póczos, A. Singh, and L. Wasserman. On the high dimensional power of a linear-time two sample test under mean-shift alternatives. In *Artificial Intelligence and Statistics*, pages 772–780. PMLR, 2015.

A. C. Rencher and G. B. Schaalje. *Linear models in statistics*. John Wiley & Sons, 2008.

R. T. Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, N. J., 1970.

H. P. Rosenthal. On the subspaces of $l_p(p > 2)$ spanned by sequences of independent random variables. *Israel J. Math.*, 8(3):273–303, 1970.

N. Ross. Fundamentals of Stein's method. 2011.

V. I. Rotar. Limit theorems for multilinear forms and quasipolynomial functions. *Theory Probab. Appl.*, 20(3):512–532, 1976.

V. I. Rotar et al. Limit theorems for polylinear forms. *J. Multivariate Anal.*, 9(4):511–530, 1979.

H. Rubin and R. Vitale. Asymptotic distribution of symmetric statistics. *Ann. Statist.*, pages 165–170, 1980.

F. Salehi, E. Abbasi, and B. Hassibi. The impact of regularization on high-dimensional logistic regression. *Advances in Neural Information Processing Systems*, 32, 2019.

A. Salmona, V. De Bortoli, J. Delon, and A. Desolneux. Can push-forward generative models fit multimodal distributions? *Advances in Neural Information Processing Systems*, 35:10766–10779, 2022.

A. Schrab, I. Kim, M. Albert, B. Laurent, B. Guedj, and A. Gretton. Mmd aggregated two-sample test. *J. Mach. Learn. Res*, 24(194):1–81, 2023.

M. Schweinberger and M. S. Handcock. Local dependence in random graph models: characterization, properties and statistical inference. *J. Roy. Statist. Soc. Ser. B*, 77(3):647–676, 2015.

M. E. A. Seddik, C. Louart, M. Tamaazousti, and R. Couillet. Random matrix theory proves that deep learning representations of gan-data behave as Gaussian mixtures. In *International Conference on Machine Learning*, pages 8573–8582. PMLR, 2020.

V. V. Senatov. *Normal Approximation*. De Gruyter, 1998.

R. J. Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 1980.

I. Shevtsova. On the absolute constants in the Berry-Esseen type inequalities for identically distributed summands. *arXiv preprint arXiv:1111.6554*, 2011.

C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *J. Big Data*, 6(1):1–48, 2019.

J. A. Soloff, R. F. Barber, and R. Willett. Bagging provides assumption-free stability. *J. Mach. Learn. Res*, 25(131):1–35, 2024.

E. M. Stein and R. Shakarchi. *Functional analysis: introduction to further topics in analysis*, volume 4. Princeton University Press, Princeton, 2011.

I. Steinwart and C. Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and rkhss. *Constr. Approx.*, 35:363–417, 2012.

M. Stojnic. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*, 2013a.

M. Stojnic. Upper-bounding ℓ_1-optimization weak thresholds. *arXiv preprint arXiv:1303.7289*, 2013b.

E. Tam and D. B. Dunson. On the statistical capacity of deep generative models. *arXiv preprint arXiv:2501.07763*, 2025.

T. Tao and V. Vu. Random matrices: universality of local eigenvalue statistics. *Acta Math.*, 206:127–204, 2011.

T. Tao and V. Vu. Random matrices: universality of local spectral statistics of non-hermitian matrices. *Ann. Probab.*, 43:782–874, 2015.

A. M. Taqi, A. Awad, F. Al-Azzo, and M. Milanova. The impact of multi-optimizers and data augmentation on tensorflow convolutional neural network performance. In *Proc. of IEEE MIPR*, pages 140–145, 2018.

T. Temčinas, V. Nanda, and G. Reinert. Multivariate central limit theorems for random clique complexes. *Journal of Applied and Computational Topology*, 8(6):1837–1880, 2024.

C. Thrampoulidis. *Recovering structured signals in high dimensions via non-smooth convex optimization: Precise performance analysis*. PhD thesis, California Institute of Technology, 2016.

C. Thrampoulidis, S. Oymak, and B. Hassibi. The Gaussian min-max theorem in the presence of convexity. *arXiv preprint arXiv:1408.4837*, 2014.

C. Thrampoulidis, S. Oymak, and B. Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pages 1683–1709. PMLR, 2015.

C. Thrampoulidis, E. Abbasi, and B. Hassibi. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Trans. Inf. Theory*, 64(8):5592–5628, 2018.

H. Trotter. An elementary proof of the central limit theorem. *Arch. Math. (Basel)*, 10(1):226–234, 1959.

M. J. Van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Stat. Appl. Genet. Mol. Biol.*, 6(1), 2007.

A. J. van Es and R. Helmers. Elementary symmetric polynomials of increasing order. *Probab. Theory Related Fields*, 80(1):21–35, 1988.

R. Van Handel. Probability in high dimension. *Lecture Notes (Princeton University)*, 2014.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

B. von Bahr and C.-G. Esseen. Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *Ann. Math. Statist.*, pages 299–303, 1965.

H. Wackernagel. *Multivariate geostatistics: an introduction with applications*. Springer Science & Business Media, 2003.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge

university press, 2019.

M. J. Wainwright, M. I. Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

L. Wang and D. Paul. Limiting spectral distribution of renormalized separable sample covariance matrices when p/n→ 0. *J. Multivariate Anal.*, 126:25–52, 2014.

L. Wang, B. Peng, and R. Li. A high-dimensional nonparametric multivariate test for mean vector. *J. Amer. Statist. Assoc.*, 110(512):1658–1669, 2015.

R. Wang, C. Zhu, S. Volgushev, and X. Shao. Inference for change points in high-dimensional data via selfnormalization. *Ann. Statist.*, 50(2):781–806, 2022.

C.-K. Wen, G. Pan, K.-K. Wong, M. Guo, and J.-C. Chen. A deterministic equivalent for the analysis of non-Gaussian correlated mimo multiple access channels. *IEEE Trans. Inf. Theory*, 59(1):329–352, 2012.

P. M. Wood. Universality of the esd for a fixed matrix plus small random noise: A stability approach. 2016.

G. Wynne and A. B. Duncan. A kernel two-sample test for functional data. *J. Mach. Learn. Res*, 23(73): 1–51, 2022.

J. Yan and X. Zhang. Kernel two-sample tests in high dimension: interplay between moment discrepancy and dimension-and-sample orders. *Biometrika*, 110(2):411–430, 2022.

O. Yanushkevichiene. On bounds in limit theorems for some U-statistics. *Theory Probab. Appl.*, 56(4): 660–673, 2012.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Comm. ACM*, 64(3):107–115, 2021.

Appendix A

Additional results and proofs for Sections 3.2, 3.4

This appendix provide additional results and proofs that concern the upper bounds and the applications of degree-two U-statistics in Chapter 3. The appendix is organised as follows. The first two sections provide additional content:

Appendix A.1 states additional results for the Gaussian mean-shift setup of Section 3.4.3, including a demonstration of how Assumption 3.2 can be verified by a simple Taylor expansion.

Appendix A.2 presents auxiliary tools used in subsequent proofs.

The remaining sections consist of proofs:

Appendix A.3 proves the main upper bound result, Theorem 3.1. Appendix A.3.1 states a list of intermediate lemmas that provides a proof overview.

Appendix A.4 proves the remaining results in Section 3.2.

Appendix A.5 proves the results in Section 3.4.

Appendices A.6 and **A.7** present proofs for the results in Appendices A.1 and A.2 respectively.

To standardise notation in this appendix, unlike Section 3.2, we shall use $D_n \equiv u_2(Y)$ for both the general degree-two U-statistic defined in (3.1) and the specific U-statistic D_n arising from MMD and KSD.

A.1 Additional results for Gaussian mean-shift in Section 3.4.3

In this section, we consider the Gaussian mean-shift setup defined in Section 3.4.3, where $Q = \mathcal{N}(\mu, \Sigma)$ and $P = \mathcal{N}(0, \Sigma)$ with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. We derive analytical expressions of the moments of U-statistics for (i) KSD with RBF, (ii) MMD with RBF and (iii) MMD with linear kernel. We also verify Assumption 3.2 for the three cases, which confirm that our error bounds apply. We refer interested readers to Huang, Liu, Duncan, and Gandy (2023) for a discussion and results on the verification of Assumption 3.1.

A.1.1. A decomposition of the RBF kernel

For both MMD and KSD, the key in verifying the assumptions for the RBF kernel is a functional decomposition. The usual Mercer representation of the RBF kernel is available only with respect to a univariate zero-mean Gaussian measure and involves some cumbersome Hermite polynomials. Since we do not actually require orthogonality of the functions in Assumption 3.2, we opt for a simpler functional representation as given below. We also assume WLOG that the bandwidth $\gamma > 8$, since we only consider the case $\gamma = \omega(1)$ in our setup.

Lemma A.1. *Assume that $\gamma > 8$. Consider two independent d-dimensional Gaussian vectors $\mathbf{U} \sim \mathcal{N}(\mu_1, I_d)$ and $\mathbf{V} \sim \mathcal{N}(\mu_2, I_d)$ for some mean vectors $\mu_1, \mu_2 \in \mathbb{R}^d$. Then, for any $\nu \in (2, 4]$ and $\mu_1, \mu_2 \in \mathbb{R}^d$, we have that*

$$\mathbb{E}\Big[\Big|\exp\Big(-\frac{1}{2\gamma}\|\mathbf{U}-\mathbf{V}\|_2^2\Big) - \prod_{j=1}^d \Big(\sum_{k=0}^K \lambda_k^* \phi_k^*(U_j)\phi_k^*(V_j)\Big)\Big|^\nu\Big] \xrightarrow{K\to\infty} 0\,.$$

where $\phi_k^*(x) \coloneqq x^k e^{-x^2/(2\gamma)}$ and $\lambda_k^* \coloneqq \frac{1}{k!\,\gamma^k}$ for each $k \in \mathbb{N}\cup\{0\}$.

To see that Lemma A.1 indeed gives the functional decomposition we want in Assumption 3.2, we need to rewrite the product of sums into a sum. To this end, let g_d be the d-tuple generalisation of the Cantor pairing function from $\mathbb{N}$ to $(\mathbb{N}\cup\{0\})^d$ and $[g_d(k)]_l$ be the l-th element of $g_d(k)$. Given $\{\lambda_l^*\}_{l=0}^\infty$ and $\{\phi_l^*\}_{l=0}^\infty$ from Lemma A.1, we define, for every $k \in \mathbb{N}$ and $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$,

$$\alpha_k \coloneqq \prod_{l=1}^d \lambda^*_{[g_d(k)]_l} \qquad \text{and} \qquad \psi_k(\mathbf{x}) \coloneqq \prod_{l=1}^d \phi^*_{[g_d(k)]_l}(x_l)\,. \tag{A.1}$$

With this construction, for each $K \in \mathbb{N}$, we can now write

$$\begin{aligned}
&\prod_{j=1}^d \Big(\sum_{k=0}^K \lambda_k^*\phi_k^*(U_j)\phi_k^*(V_j)\Big) \\
&\qquad = \sum_{k_1,\ldots,k_d=0}^K (\lambda^*_{k_1}\ldots\lambda^*_{k_d})(\phi^*_{k_1}(U_1)\ldots\phi^*_{k_d}(U_d))(\phi^*_{k_1}(V_1)\ldots\phi_{k_d}(V_d)) \\
&\qquad = \sum_{k_1,\ldots,k_d=0}^K \alpha_{g_d^{-1}(k_1,\ldots,k_d)}\ \psi_{g_d^{-1}(k_1,\ldots,k_d)}(\mathbf{U})\ \psi_{g_d^{-1}(k_1,\ldots,k_d)}(\mathbf{V})\,.
\end{aligned}$$

Since the Cantor pairing function is such that $\min_{l \leq d}[g_d(K)]_l \to \infty$ as $K \to \infty$, Lemma A.1 indeed gives a functional decomposition in terms of $\{\alpha_k\}_{k=1}^{\infty}$ and $\{\psi_k\}_{k=1}^{\infty}$ as

$$\mathbb{E}\Big[\Big|\exp\Big(-\frac{1}{2\gamma}\|\mathbf{U}-\mathbf{V}\|_2^2\Big) - \sum_{k=1}^{K}\alpha_k\psi_k(\mathbf{U})\psi_k(\mathbf{V})\Big|^{\nu}\Big] \xrightarrow{K\to\infty} 0\,. \tag{A.2}$$

We remark that both α_k and ψ_k are independent of the mean vectors μ_1 and μ_2, which makes this representation useful for a generic mean-shift setting.

A.1.2. KSD U-statistic with RBF kernel

Under the Gaussian mean-shift setup with an identity covariance matrix, gradient of the log-density is given by $\nabla \log p(\mathbf{x}) = -\mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^d$ and the U-statistic for the RBF-kernel KSD is

$$\begin{aligned} u_P^{\mathrm{KSD}}(\mathbf{x},\mathbf{x}') &= \big(\nabla \log p(\mathbf{x})\big)^{\top}\big(\nabla \log p(\mathbf{x}')\big)\kappa(\mathbf{x},\mathbf{x}') + \big(\nabla \log p(\mathbf{x})\big)^{\top}\nabla_2\kappa(\mathbf{x},\mathbf{x}') \\ &\quad + \big(\nabla \log p(\mathbf{x}')\big)^{\top}\nabla_1\kappa(\mathbf{x},\mathbf{x}') + \mathrm{Tr}(\nabla_1\nabla_2\kappa(\mathbf{x},\mathbf{x}')) \\ &= \exp\Big(-\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{2\gamma}\Big)\Big(\mathbf{x}^{\top}\mathbf{x}' + \frac{1}{\gamma}\mathbf{x}^{\top}(\mathbf{x}'-\mathbf{x}) + \frac{1}{\gamma}(\mathbf{x}')^{\top}(\mathbf{x}-\mathbf{x}') + \Big(\frac{d}{\gamma} - \frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{\gamma^2}\Big)\Big) \\ &= \exp\Big(-\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{2\gamma}\Big)\Big(\mathbf{x}^{\top}\mathbf{x}' - \frac{\gamma+1}{\gamma^2}\|\mathbf{x}-\mathbf{x}'\|_2^2 + \frac{d}{\gamma}\Big)\,. \end{aligned} \tag{A.3}$$

We first verify that Assumption 3.2 holds by adapting $\{\alpha_k\}_{k=1}^{\infty}$ and $\{\psi_k\}_{k=1}^{\infty}$ from Appendix A.1.1.

Lemma A.2. *Assume that $\gamma > 24$. For $k' \in \mathbb{N}$, consider*

$$\begin{aligned} \lambda_{(k'-1)(d+3)+1} &= -\frac{\gamma+1}{\gamma^2}\alpha_{k'}\,, & \phi_{(k'-1)(d+3)+1}(\mathbf{x}) &= \psi_{k'}(\mathbf{x})(\|\mathbf{x}\|_2^2+1)\,, \\ \lambda_{(k'-1)(d+3)+2} &= \frac{\gamma+1}{\gamma^2}\alpha_{k'}\,, & \phi_{(k'-1)(d+3)+2}(\mathbf{x}) &= \psi_{k'}(\mathbf{x})\|\mathbf{x}\|_2^2\,, \\ \lambda_{(k'-1)(d+3)+3} &= \Big(\frac{d}{\gamma}+\frac{\gamma+1}{\gamma^2}\Big)\alpha_{k'}\,, & \phi_{(k'-1)(d+3)+3}(\mathbf{x}) &= \psi_{k'}(\mathbf{x})\,, \end{aligned}$$

and for $l = 1, \ldots, d$, define

$$\lambda_{(k'-1)(d+3)+3+l} = \frac{\gamma^2+2\gamma+2}{\gamma^2}\alpha_{k'}\,, \qquad \phi_{(k'-1)(d+3)+3+l}(\mathbf{x}) = \psi_{k'}(\mathbf{x})x_l\,.$$

Then Assumption 3.2 holds with any $\nu \in (2,3]$ for $u = u_P^{\mathrm{KSD}}$, $\{\lambda_k\}_{k=1}^{\infty}$ and $\{\phi_k\}_{k=1}^{\infty}$ defined above.

A.1.3. MMD U-statistic with RBF kernel

Under the Gaussian mean-shift setup with an identity covariance matrix, the MMD U-statistic with a RBF kernel has the form

$$u^{\mathrm{MMD}}(\mathbf{z},\mathbf{z}') = \kappa(\mathbf{x},\mathbf{x}') + \kappa(\mathbf{y},\mathbf{y}') - \kappa(\mathbf{x},\mathbf{y}') - \kappa(\mathbf{x}',\mathbf{y})$$

$$= \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{2\gamma}\right) + \exp\left(-\frac{\|\mathbf{y}-\mathbf{y}'\|_2^2}{2\gamma}\right) - \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}'\|_2^2}{2\gamma}\right) - \exp\left(-\frac{\|\mathbf{x}'-\mathbf{y}\|_2^2}{2\gamma}\right), \tag{A.4}$$

for $\mathbf{z} \coloneqq (\mathbf{x}, \mathbf{y}), \mathbf{z}' \coloneqq (\mathbf{x}', \mathbf{y}') \in \mathbb{R}^{2d}$. We first verify that Assumption 3.2 holds again by adapting $\{\alpha_k\}_{k=1}^{\infty}$ and $\{\psi_k\}_{k=1}^{\infty}$ from Appendix A.1.1.

Lemma A.3. *Assume that $\gamma > 8$. Then Assumption 3.2 holds with any value $\nu \in (2, 3]$ and the function $u((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')) = u^{\mathrm{MMD}}((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}'))$ for $\mathbf{x}, \mathbf{y}, \mathbf{x}', \mathbf{y}' \in \mathbb{R}^d$, with the sequences of values and functions given for each $k \in \mathbb{N}$ as $\gamma_k = \alpha_k$ and $\phi_k(\mathbf{x}, \mathbf{y}) = \psi_k(\mathbf{x}) - \psi_k(\mathbf{y})$.*

A.1.4. MMD U-statistic with linear kernel

In this section, we consider the mean-shift setup with a general covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$, i.e., $Q = \mathcal{N}(\mu, \Sigma)$ and $P = \mathcal{N}(0, \Sigma)$. The MMD with a linear kernel $\kappa(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}'$ has the form

$$u^{\mathrm{MMD}}(\mathbf{z}, \mathbf{z}') = \mathbf{x}^\top \mathbf{x}' + \mathbf{y}^\top \mathbf{y}' - \mathbf{x}^\top \mathbf{y}' - \mathbf{y}^\top \mathbf{x}' ,$$

where $\mathbf{z} \coloneqq (\mathbf{x}, \mathbf{y}), \mathbf{z}' \coloneqq (\mathbf{x}', \mathbf{y}') \in \mathbb{R}^{2d}$. In this case, Assumption 3.2 holds directly because we can represent u^{MMD} as

$$u^{\mathrm{MMD}}(\mathbf{z}, \mathbf{z}') = (\mathbf{x} - \mathbf{y})^\top (\mathbf{x}' - \mathbf{y}') = \sum_{l=1}^{d} (x_l - y_l)(x'_l - y'_l) = \sum_{l=1}^{d} \gamma_l \psi_l(\mathbf{z}) \psi_l(\mathbf{z}') , \tag{A.5}$$

where $\gamma_l = 1$, $\psi_l(\mathbf{z}) = x_l - y_l$ and $\psi_l(\mathbf{z}') = x'_l - y'_l$.

Details on the choice of linear kernel in simulations. In the last example in Section 3.4.3, we chose $\mu = (0, 10, \ldots, 0) \in \mathbb{R}^d$ and a diagonal Σ with $\Sigma_{11} = 0.5(d+1)$, $\Sigma_{ii} = 0.5$ for $i > 1$ and $\Sigma_{ij} = 0$ otherwise. Note that by the invariance of Gaussian distributions under orthogonal transformation, this is equivalent to choosing Σ as $0.5\mathbf{I}_d + 0.5\mathbf{J}_d$, where $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ is the identity matrix, $\mathbf{J}_d \in \mathbb{R}^{d\times d}$ is the all-one matrix and μ is transformed by an appropriate orthogonal matrix of eigenvectors. Notably, this choice ensures the limit of u^{MMD} remains non-Gaussian. Indeed, when Q and P are Gaussian, the statistic D_n^{MMD} can be written as a sum of shifted-and-rescaled chi-squares, where the scaling factors are $0.5(d+1), 0.5, \ldots, 0.5$, the eigenvalues of Σ. As d grows, the eigenvalue $0.5(d+1)$ dominates, and the limiting distribution is then dominated by the first summand, thereby yielding a chi-square limit up to shifting and rescaling. This is numerically demonstrated in the right figure of Figure 3.2. As a remark, we do not expect this exact setting to occur in practice; it should instead be treated as a toy setup to demonstrate the possibility of non-Gaussianity and convey an intuition of when this may occur.

A.2 Auxiliary tools

A.2.1. Generic moment bounds

We first present two-sided bounds on the moments of a martingale, which are useful in bounding ν-th moment terms of different statistics. The original result is due to Burkholder (1966), and the constant C_ν is provided by von Bahr and Esseen (1965) and Dharmadhikari et al. (1968b). We also include Burkholder's original upper bound in the second line, which is used when a finer control is required in Chapter 4. The boundedness of C'_ν over a bounded interval is because Burkholder (1966)'s constant arises from Marcinkiewicz interpolation theorem and the Khintchine inequality.

Lemma A.4. *Fix $\nu > 1$. For a martingale difference sequence $Y_1, \dots, Y_n$ taking values in $\mathbb{R}$,*

$$c_\nu \, n^{\min\{0,\nu/2-1\}} \sum_{i=1}^n \mathbb{E}[|Y_i|^\nu] \;\le\; \mathbb{E}\big[\big|\sum_{i=1}^n Y_i\big|^\nu\big] \;\le\; C_\nu \, n^{\max\{0,\nu/2-1\}} \sum_{i=1}^n \mathbb{E}[|Y_i|^\nu] \,,$$

for $C_\nu := \max\big\{2, (8(\nu-1)\max\{1, 2^{\nu-3}\})^\nu\big\}$ and a constant $c_\nu > 0$ depending only on ν. Moreover, there exists some constant $C'_\nu > 0$ depending only on ν such that

$$\mathbb{E}\big[\big|\sum_{i=1}^n Y_i\big|^\nu\big] \;\le\; C'_\nu \, \mathbb{E}\Big[\big|\sum_{i=1}^n Y_i^2\big|^{\nu/2}\Big] \,,$$

where $\sup_{\nu \in I} C'_\nu$ is bounded whenever $I \subset (1, \infty)$ is a fixed bounded interval.

The next moment computation for a quadratic form of Gaussians is used throughout the proof:

Lemma A.5 (Lemma 2.3, Magnus (1978))**.** *Given a standard Gaussian vector η in $\mathbb{R}^m$ and a symmetric $m \times m$ matrix A, we have that $\mathbb{E}[\eta^\top A \eta] = \mathrm{Tr}(A)$ and*

$$\mathbb{E}[(\eta^\top A \eta)^2] = \mathrm{Tr}(A)^2 + 2\mathrm{Tr}(A^2) \,, \quad \mathbb{E}[(\eta^\top A \eta)^3] = \mathrm{Tr}(A)^3 + 6\mathrm{Tr}(A)\mathrm{Tr}(A^2) + 8\mathrm{Tr}(A^3) \,.$$

A.2.2. Moment bounds for U-statistics

We first present a result that bounds the moments of a U-statistic $D_n = u_2(Y)$ defined as in (3.1).

Lemma A.6. *Fix $n \ge 2$ and $\nu \ge 2$. Then, there exist absolute constants $c_\nu, C_\nu > 0$ depending only on ν such that*

$$\begin{aligned}
\mathbb{E}[|D_n - \mathbb{E}D_n|^\nu] \;&\le\; C_\nu \, n^{\nu/2}(n-1)^{-\nu} M^\nu_{\mathrm{cond};\nu} + C_\nu \, (n-1)^{-\nu} M^\nu_{\mathrm{full};\nu} \,, \\
\mathbb{E}[|D_n - \mathbb{E}D_n|^\nu] \;&\ge\; c_\nu n (n-1)^{-\nu} M^\nu_{\mathrm{cond};\nu} + c_\nu n^{-(\nu-1)} (n-1)^{-(\nu-1)} M^\nu_{\mathrm{full};\nu} \,.
\end{aligned}$$

In other words,

$$\begin{aligned}\mathbb{E}[|D_n - \mathbb{E}D_n|^\nu] &= O(n^{-\nu/2}M^\nu_{\text{cond};\nu} + n^{-\nu}M^\nu_{\text{full};\nu}) \,, \\ \mathbb{E}[|D_n - \mathbb{E}D_n|^\nu] &= \Omega(n^{-(\nu-1)}M^\nu_{\text{cond};\nu} + n^{-2(\nu-1)}M^\nu_{\text{full};\nu}) \,.\end{aligned}$$

The next two results summarise how the moments of variables under the functional decomposition in Assumption 3.2 interact with the moments of the original statistic u under R:

Lemma A.7. *Let $\{\phi_k\}_{k=1}^\infty$, $\{\lambda_k\}_{k=1}^\infty$ and $\varepsilon_{K;\nu}$ be defined as in Assumption 3.2. For $\mathbf{X}_1, \mathbf{X}_2 \overset{i.i.d.}{\sim} R$, write $\mu_k := \mathbb{E}[\phi_k(\mathbf{X}_1)]$ and let the moment terms $D, M_{\text{cond};\nu}, M_{\text{full};\nu}$ be defined as in Chapter 3 and Section 3.2. Then we have the following:*

(i) $\left|\sum_{k=1}^K \lambda_k\mu_k^2 - D\right| \le \varepsilon_{K;1}$;

(ii) for any $\nu \in [1,3]$, we have that

$$\tfrac{1}{4}(M_{\text{cond};\nu})^\nu - \varepsilon^\nu_{K;\nu} \le \mathbb{E}\Big[\Big|\sum\nolimits_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1) - \mu_k)\mu_k\Big|^\nu\Big] \le 4((M_{\text{cond};\nu})^\nu + \varepsilon^\nu_{K;\nu}) \,;$$

(iii) there exist some absolute constants $c, C > 0$ such that

$$\begin{aligned}&\mathbb{E}\Big[\Big|\sum\nolimits_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1) - \mu_k)(\phi_k(\mathbf{X}_2) - \mu_k)\Big|^\nu\Big] \\ &\qquad \le 4C(M_{\text{full};\nu})^\nu - \frac{1}{2}(M_{\text{cond};\nu})^\nu + (4C+2)\varepsilon^\nu_{K;\nu} \,, \\ &\mathbb{E}\Big[\Big|\sum\nolimits_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1) - \mu_k)(\phi_k(\mathbf{X}_2) - \mu_k)\Big|^\nu\Big] \\ &\qquad \ge \frac{c}{4}(M_{\text{full};\nu})^\nu - 8(M_{\text{cond};\nu})^\nu - (c+8)\varepsilon^\nu_{K;\nu} \,.\end{aligned}$$

The next result assumes the notations of Lemma A.7, and additionally denotes

$$\Lambda^K := \text{diag}\{\lambda_1, \ldots, \lambda_K\} \in \mathbb{R}^{K\times K} \,, \quad \phi^K(x) := (\phi_1(x), \ldots, \phi_K(x))^\top \in \mathbb{R}^K \,.$$

Lemma A.8. *For $\mu^K := \mathbb{E}[\phi^K(\mathbf{X}_1)]$ and $\Sigma^K := \text{Cov}[\phi^K(\mathbf{X}_1)]$, we have*

$$\begin{aligned}\sigma^2_{\text{cond}} - 4\sigma_{\text{cond}}\varepsilon_{K;2} - 4\varepsilon^2_{K;2} \le (\mu^K)^\top\Lambda^K\Sigma^K\Lambda^K(\mu^K) &\le (\sigma_{\text{cond}} + 2\varepsilon_{K;2})^2 \,. \\ (\sigma_{\text{full}} - \varepsilon_{K;2})^2 \le \text{Tr}((\Lambda^K\Sigma^K)^2) &\le (\sigma_{\text{full}} + \varepsilon_{K;2})^2 \,.\end{aligned}$$

In particular, for $\nu \in [1,3]$ and two i.i.d. zero-mean Gaussian vector $\mathbf{Z}_1$ and $\mathbf{Z}_2$ in $\mathbb{R}^K$ with variance Σ^K, there exists some absolute constant $C > 0$ such that

$$\begin{aligned}&\mathbb{E}[|(\mu^K)^\top\Lambda^K\mathbf{Z}_1|^\nu] \le 7\big(\sigma^\nu_{\text{cond}} + 8\varepsilon^\nu_{K;2}\big) \,, \qquad \mathbb{E}[|\mathbf{Z}_1^\top\Lambda^K\mathbf{Z}_2|^\nu] \le 6\big(\sigma^\nu_{\text{full}} + \varepsilon^\nu_{K;2}\big) \,, \\ &\mathbb{E}\big[\big|(\phi^K(\mathbf{X}_1) - \mu^K)^\top\Lambda^K\mathbf{Z}_1\big|^\nu\big] \le 8C(M_{\text{full};\nu})^\nu - (M_{\text{cond};\nu})^\nu + (8C+4)\varepsilon^\nu_{K;\nu} \,.\end{aligned}$$

The next lemma gives an equivalent expression for W_n^K defined in (3.9) and also

controls the moments of W_n^K.

Lemma A.9. *Let $\{\eta_i^K\}_{i=1}^n$ be a sequence of i.i.d. standard Gaussian vectors in $\mathbb{R}^K$. Then*

(i) the distribution of W_n^K satisfies

$$W_n^K \overset{d}{=} \frac{1}{n^{3/2}(n-1)^{1/2}}\Big(\sum_{i,j=1}^n (\eta_i^K)^\top (\Sigma^K)^{1/2}\Lambda^K(\Sigma^K)^{1/2}\eta_j^K - n\mathrm{Tr}(\Sigma^K\Lambda^K)\Big) + D;$$

(ii) the mean satisfies $\mathbb{E}[W_n^K] = D$ for every $K \in \mathbb{N}$;

(iii) the variance is controlled as

$$\frac{2}{n(n-1)}(\sigma_{\text{full}} - \varepsilon_{K;2})^2 \;\le \mathrm{Var}[W_n^K] \;\le\; \frac{2}{n(n-1)}(\sigma_{\text{full}} + \varepsilon_{K;2})^2\,;$$

(iv) the third central moment is controlled as

$$\mathbb{E}\big[(W_n^K - D)^3\big] \;\le \frac{8\big(\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_1)] - M_{\text{full};3}^3 + (M_{\text{full};3}+\varepsilon_{K;3})^3\big)}{n^{3/2}(n-1)^{3/2}},$$
$$\mathbb{E}\big[(W_n^K - D)^3\big] \;\ge \frac{8\big(\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_1)] + M_{\text{full};3}^3 - (M_{\text{full};3}+\varepsilon_{K;3})^3\big)}{n^{3/2}(n-1)^{3/2}};$$

(v) the fourth central moment is controlled as

$$\mathbb{E}\big[(W_n^K - D)^4\big] \;\le \frac{12}{n^2(n-1)^2}\Big(4\,\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_4)u(\mathbf{X}_4,\mathbf{X}_1)] - 4M_{\text{full};4}^4 + 4(M_{\text{full};4}+\varepsilon_{K;4})^4 + (\sigma_{\text{full}}+\varepsilon_{K;2})^4\Big),$$
$$\mathbb{E}\big[(W_n^K - D)^4\big] \;\ge \frac{12}{n^2(n-1)^2}\Big(4\,\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_4)u(\mathbf{X}_4,\mathbf{X}_1)] + 4M_{\text{full};4}^4 - 4(M_{\text{full};4}+\varepsilon_{K;4})^4 + (\sigma_{\text{full}}-\varepsilon_{K;2})^4\Big);$$

(vi) we also have a generic moment bound: For $m \in \mathbb{N}$, there exists some absolute constant $C_m > 0$ depending only on m such that

$$\mathbb{E}\big[(W_n^K)^{2m}\big] \;\le \frac{C_m}{n^m(n-1)^m}(\sigma_{\text{full}}+\varepsilon_{K;2})^{2m} + C_m\,D^{2m}\,;$$

(vii) if Assumption 3.2 holds for some $\nu \ge 2$ then $\lim_{K\to\infty}\mathrm{Var}[W_n^K] = \frac{2}{n(n-1)}\sigma_{\text{full}}^2$. If Assumption 3.2 holds for some $\nu \ge 3$, then

$$\lim_{K\to\infty}\mathbb{E}\big[(W_n^K - D)^3\big] \;= \frac{8\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_1)]}{n^{3/2}(n-1)^{3/2}}\,,$$

and if Assumption 3.2 holds for some $\nu \ge 4$, then

$$\lim_{K\to\infty}\mathbb{E}\big[(W_n^K - D)^4\big] \;= \frac{12(4\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_4)u(\mathbf{X}_4,\mathbf{X}_1)] + \sigma_{\text{full}}^4)}{n^2(n-1)^2}\,.$$

A.2.3. Distribution bounds

The following is a standard approximation of an indicator function for bounding the probability of a given event; see e.g. the proof of Theorem 3.3, Chen et al. (2011).

Lemma A.10. *Fix any $m \in \mathbb{N} \cup \{0\}$, $\tau \in \mathbb{R}$ and $\delta > 0$. Then there exists an m-times differentiable $\mathbb{R} \to \mathbb{R}$ function $h_{m;\tau;\delta}$ such that $h_{m;\tau+\delta;\delta}(x) \leq \mathbb{I}_{\{x>\tau\}} \leq h_{m;\tau;\delta}(x)$. For $0 \leq r \leq m$, the r-th derivative $h^{(r)}_{m;\tau;\delta}$ is continuous and bounded above by δ^{-r}. Moreover, for every $\epsilon \in [0,1]$, $h^{(m)}$ satisfies that*

$$|h^{(m)}_{m;\tau;\delta}(x) - h^{(m)}_{m;\tau;\delta}(y)| \leq C_{m,\epsilon}\, \delta^{-(m+\epsilon)}\, |x-y|^{\epsilon} ,$$

with respect to the constant $C_{m,\epsilon} = \binom{m}{\lfloor m/2 \rfloor}(m+1)^{m+\epsilon}$.

The next bound is useful for approximating the distribution of a sum of two (possibly correlated) random variables X and Y by the distribution of X alone, provided that the influence of Y is small.

Lemma A.11. *For two real-valued random variables X and Y, any $a, b \in \mathbb{R}$ and $\epsilon > 0$, we have*

$$\begin{aligned}
\mathbb{P}(a \leq X+Y \leq b) &\leq \mathbb{P}(a-\epsilon \leq X \leq b+\epsilon) + \mathbb{P}(|Y| \geq \epsilon) , \\
\mathbb{P}(a \leq X+Y \leq b) &\geq \mathbb{P}(a+\epsilon \leq X \leq b-\epsilon) - \mathbb{P}(|Y| \geq \epsilon) .
\end{aligned}$$

Fact 4.4, i.e. Theorem 8 of Carbery and Wright (2001), gives an anti-concentration result for a polynomial of random variables drawn from a log-concave density. The next lemma restates the result in the case of a quadratic form of a K-dimensional standard Gaussian vector η.

Lemma A.12. *Let $p(\mathbf{x})$ be a degree-two polynomial of $\mathbf{x} \in \mathbb{R}^K$ taking values in $\mathbb{R}$. Then there exists an absolute constant C independent of p and η such that, for every $t \in \mathbb{R}$,*

$$\mathbb{P}\big(|p(\eta)| \leq t\big) \leq Ct^{1/2}(\mathbb{E}[|p(\eta)|^2])^{-1/4} \leq Ct^{1/2}(\mathrm{Var}[p(\eta)])^{-1/4} .$$

A.2.4. Weak Mercer representation

In Section 3.4.2, we have used the *weak Mercer representation* from Steinwart and Scovel (2012). We summarise their result below, which combines their Lemma 2.3, Lemma 2.12 and Corollary 3.2:

Lemma A.13. *Consider a probability measure R on $\mathbb{R}^b$, $\mathbf{V}_1, \mathbf{V}_2 \overset{i.i.d.}{\sim} R$ and a measurable kernel κ^* on $\mathbb{R}^b$. If $\mathbb{E}[\kappa^*(\mathbf{V}_1, \mathbf{V}_1)] < \infty$, there exists a sequence of functions $\{\phi_k\}_{k=1}^{\infty}$ in $L_2(\mathbb{R}^b, R)$ and a bounded sequence of non-negative values $\{\lambda_k\}_{k=1}^{\infty}$ with $\lim_{k\to\infty} \lambda_k = 0$, such that as K grows, $\left|\sum_{k=1}^{K} \lambda_k \phi_k(\mathbf{V}_1)\phi_k(\mathbf{V}_2) - \kappa^*(\mathbf{V}_1, \mathbf{V}_2)\right| \to 0$. The series converges $R \otimes R$ almost surely.*

A.3 Proof of the main result, Theorem 3.1

In this section, we prove Theorem 3.1. The proof is necessarily tedious as we seek to control "spectral" approximation errors (i.e. the error from a truncated functional decomposition) and multiple stochastic approximation errors at the same time. The section is organised as follows:

- In Appendix A.3.1, we list notations and key lemmas that formalise the steps for proving Theorem 3.1;
- In Appendix A.3.2, we present the proof body of Theorem 3.1, which directly combines results from the different lemmas;
- In Appendix A.3.3, A.3.4, A.3.5 and A.3.6, we present the proof of the key lemmas. Each section starts with an informal sketch of proof ideas followed by the actual proof of the result.

A.3.1. Auxiliary lemmas

Recall that our goal is to study the distribution of

$$D_n \coloneqq \frac{1}{n(n-1)} \sum\nolimits_{1 \le i \ne j \le n} u(\mathbf{X}_i, \mathbf{X}_j) \, .$$

The three results in this section form the key steps of the proof. We fix $\sigma > 0$ to be some normalisation constant to be chosen later.

1. "Spectral" approximation. For $K \in \mathbb{N}$, we define the truncated version of D_n by

$$\begin{aligned} D_n^K &\coloneqq \frac{1}{n(n-1)} \sum\nolimits_{1 \le i \ne j \le n} \sum\nolimits_{k=1}^{K} \lambda_k \phi_k(\mathbf{X}_i) \phi_k(\mathbf{X}_j) \\ &= \frac{1}{n(n-1)} \sum\nolimits_{1 \le i \ne j \le n} (\phi^K(\mathbf{X}_i))^\top \Lambda^K \phi^K(\mathbf{X}_j) \, . \end{aligned}$$

We also denote the rescaled statistics for convenience as

$$\tilde{D}_n \coloneqq \frac{\sqrt{n(n-1)}}{\sigma} D_n \, , \qquad \tilde{D}_n^K \coloneqq \frac{\sqrt{n(n-1)}}{\sigma} D_n^K \, .$$

The first lemma allows us to study the distribution of D_n^K in lieu of that of D_n up to some approximation error that vanishes as K grows.

Lemma A.14. *Fix $\delta, \sigma > 0$, $K \in \mathbb{N}$ and $t \in \mathbb{R}$. Then*

$$\mathbb{P}(\tilde{D}_n^K > t + \delta) - \varepsilon'_K \le \mathbb{P}(\tilde{D}_n > t) \le \mathbb{P}(\tilde{D}_n^K > t - \delta) + \varepsilon'_K \, , \; \varepsilon'_K \coloneqq \frac{3 n^{1/4} (n-1)^{1/4} \varepsilon_{K;1}^{1/2}}{\sigma^{1/2} \delta^{1/2}} \, .$$

2. Gaussian approximation via the Lindeberg method. The distribution of D_n^K is easier to handle, as it is a double sum of a simple quadratic form of K-dimensional

random vectors. Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ be i.i.d. Gaussian random vectors in $\mathbb{R}^K$ with mean and variance matching those of $\phi^K(\mathbf{X}_1)$, and denote Z_{ik} as the k-th coordinate of $\mathbf{Z}_i$. The goal is to replace D_n^K by the random variable

$$D_Z^K \;\coloneqq\; \frac{1}{n(n-1)} \sum\nolimits_{1 \le i \ne j \le n} \mathbf{Z}_i^\top \Lambda^K \mathbf{Z}_j \;=\; \frac{1}{n(n-1)} \sum\nolimits_{1 \le i \ne j \le n} \sum\nolimits_{k=1}^{K} \lambda_k Z_{ik} Z_{jk} \,.$$

Notice that D_Z^K takes the same form as D_n^K except that each $\phi^K(\mathbf{X}_i)$ is replaced by $\mathbf{Z}_i$. Analogous to $\tilde{D}_n$ and $\tilde{D}_n^K$, we also define a rescaled version as

$$\tilde{D}_Z^K \;\coloneqq\; \frac{\sqrt{n(n-1)}\, D_Z^K}{\sigma} \,.$$

The second lemma replaces the distribution $\tilde{D}_n^K$ by that of $\tilde{D}_Z^K$, up to some approximation error that vanishes as n grows:

Lemma A.15. *Fix $\delta, \sigma > 0$, $K \in \mathbb{N}$, $t \in \mathbb{R}$ and any $\nu \in (2, 3]$. Then*

$$\begin{aligned} \mathbb{P}(\tilde{D}_n^K > t - \delta) \;&\le \mathbb{P}(\tilde{D}_Z^K > t - 2\delta) + E_{\delta;K} \,, \\ \mathbb{P}(\tilde{D}_n^K > t + \delta) \;&\ge \mathbb{P}(\tilde{D}_Z^K > t + 2\delta) - E_{\delta;K} \,, \end{aligned}$$

where the approximation error is defined as, for some absolute constant $C > 0$,

$$E_{\delta;K} \;\coloneqq \frac{C}{\delta^\nu n^{\nu/2-1}} \Big(\frac{(M_{\mathrm{full};\nu})^\nu + \varepsilon_{K;\nu}^\nu}{\sigma^\nu} + \frac{(M_{\mathrm{cond};\nu})^\nu + \varepsilon_{K;\nu}^\nu}{(n-1)^{-\nu/2}\, \sigma^\nu} \Big) \,.$$

3. Replace D_Z^K by U_n^K. As in the statement of Theorem 3.1, let $\{\eta_i^K\}_{i=1}^n$ be the i.i.d. standard normal vectors in $\mathbb{R}^K$, and recall the notations $\mu^K \coloneqq \mathbb{E}[\phi^K(\mathbf{X}_1)]$ and $\Sigma^K \coloneqq \mathrm{Cov}[\phi^K(\mathbf{X}_1)]$. We can then express D_Z^K as

$$\begin{aligned} D_Z^K \;&= \frac{1}{n(n-1)} \sum\nolimits_{1 \le i \ne j \le n} \big((\Sigma^K)^{1/2} \eta_i^K + \mu^K \big)^\top \Lambda^K \big((\Sigma^K)^{1/2} \eta_j^K + \mu^K \big) \\ &= \frac{1}{n(n-1)} \sum\nolimits_{1 \le i \ne j \le n} (\eta_i^K)^\top (\Sigma^K)^{1/2} \Lambda^K (\Sigma^K)^{1/2} \eta_j^K + \frac{2}{n} \sum\nolimits_{i=1}^n (\mu^K)^\top \Lambda^K (\Sigma^K)^{1/2} \eta_i^K \\ &\quad + (\mu^K)^\top \Lambda^K \mu^K \,. \end{aligned}$$

This is similar to the desired variable U_n^K except for the third term:

$$\begin{aligned} U_n^K = \frac{1}{n(n-1)} \sum\nolimits_{1 \le i \ne j \le n} (\eta_i^K)^\top (\Sigma^K)^{1/2} \Lambda^K (\Sigma^K)^{1/2} \eta_j^K \\ + \frac{2}{n} \sum\nolimits_{i=1}^n (\mu^K)^\top \Lambda^K (\Sigma^K)^{1/2} \eta_i^K + D \,. \end{aligned}$$

As before, we denote $\tilde{U}_n^K \coloneqq \frac{\sqrt{n(n-1)} U_n^K}{\sigma}$. The next lemma shows that the distribution of $\tilde{D}_Z^K$ can be approximated by that of $\tilde{U}_n^K$, up to some approximation error that vanishes as $K \to \infty$.

Lemma A.16. *For any $a, b \in \mathbb{R}$ and $\epsilon > 0$, we have that*

$$\mathbb{P}(a \le \tilde{D}_Z^K \le b) \;\le\; \mathbb{P}\big(a - \epsilon \le \tilde{U}_n^K \le b + \epsilon\big) + \frac{\varepsilon_{K;1}}{\epsilon\, n^{-1/2} (n-1)^{-1/2} \sigma} \,,$$

$$\mathbb{P}(a \le \tilde{D}_Z^K \le b) \;\ge\; \mathbb{P}(a+\epsilon \le \tilde{U}_n^K \le b-\epsilon) - \frac{\varepsilon_{K;1}}{\epsilon\, n^{-1/2}(n-1)^{-1/2}\sigma} \,.$$

4. Bound the distribution of $\tilde{U}_n^K$ over a short interval. If we are to use Lemma A.14 and Lemma A.15 directly, we would end up comparing $\mathbb{P}(\tilde{D}_n > t)$ against the probabilities $\mathbb{P}(\tilde{U}_n^K > t+2\delta)$ and $\mathbb{P}(\tilde{U}_n^K > t-2\delta)$ for some small $\delta > 0$. It turns out these are not too different from $\mathbb{P}(\tilde{U}_n^K > t)$: As $\tilde{U}_n^K$ is a quadratic form of Gaussians, we can ensure it is "well spread-out" such that the probability mass of $\tilde{U}_n^K$ within a small interval $(t-2\delta, t+2\delta)$ is not too large. This is ascertained by the following lemma:

Lemma A.17. *For $a \le b \in \mathbb{R}$, there exists some absolute constant C such that*

$$\mathbb{P}(a \le \tilde{U}_n^K \le b) \;\le C(b-a)^{1/2}\Big(\frac{1}{\sigma^2}(\sigma_{\text{full}} - \varepsilon_{K;2})^2 + \frac{n-1}{\sigma^2}(\sigma_{\text{cond}}^2 - 2\sigma_{\text{cond}}\varepsilon_{K;2} - 4\varepsilon_{K;2})\Big)^{-1/4} \,.$$

A.3.2. Proof body of Theorem 3.1

Fix $\delta, \sigma > 0$, $K \in \mathbb{N}$ and $t \in \mathbb{R}$. By Lemma A.14, we have that for $\varepsilon'_K = \frac{3n^{1/4}(n-1)^{1/4}\varepsilon_{K;1}^{1/2}}{\sigma^{1/2}\delta^{1/2}}$,

$$\mathbb{P}(\tilde{D}_n^K > t+\delta) - \varepsilon'_K \;\le\; \mathbb{P}(\tilde{D}_n > t) \;\le\; \mathbb{P}(\tilde{D}_n^K > t-\delta) + \varepsilon'_K \,.$$

By Lemma A.15, we have

$$\begin{aligned}\mathbb{P}(\tilde{D}_n^K > t-\delta) \;&\le \mathbb{P}(\tilde{D}_Z^K > t-2\delta) + E_{\delta;K} \,,\\ \mathbb{P}(\tilde{D}_n^K > t+\delta) \;&\ge \mathbb{P}(\tilde{D}_Z^K > t+2\delta) - E_{\delta;K} \,,\end{aligned}$$

where the error term is defined as, for some absolute constant $C' > 0$,

$$E_{\delta;K} \;:=\; \frac{C'}{\delta^\nu n^{\nu/2-1}}\Big(\frac{(M_{\text{full};\nu})^\nu + \varepsilon_{K;\nu}^\nu}{\sigma^\nu} + \frac{(M_{\text{cond};\nu})^\nu + \varepsilon_{K;\nu}^\nu}{(n-1)^{-\nu/2}\,\sigma^\nu}\Big) \,.$$

To combine the two bounds, we consider the following decomposition:

$$\begin{aligned}\mathbb{P}(\tilde{D}_Z^K > t-2\delta) \;&=\; \mathbb{P}(\tilde{D}_Z^K > t) + \mathbb{P}(t-2\delta < \tilde{D}_Z^K \le t) \,,\\ \mathbb{P}(\tilde{D}_Z^K > t+2\delta) \;&=\; \mathbb{P}(\tilde{D}_Z^K > t) - \mathbb{P}(t < \tilde{D}_Z^K \le t+2\delta) \,.\end{aligned} \tag{A.6}$$

This allows us to combine the earlier two bounds as

$$\big|\mathbb{P}(\tilde{D}_n > t) - \mathbb{P}(\tilde{D}_Z^K > t)\big| \;\le\; \max\{\mathbb{P}(t-2\delta \le \tilde{D}_Z^K < t)\,, \mathbb{P}(t < \tilde{D}_Z^K \le t+2\delta)\} + E_{\delta;K} + \varepsilon'_K \,,$$

which gives the error of approximating the c.d.f. of $\tilde{D}_n$ by that of $\tilde{D}_Z^K$. Now fix some $\epsilon > 0$. By applying Lemma A.16 and taking appropriate limits of the endpoints to change $\le$ to $<$, $\ge$ to $>$ and taking the right endpoint to positive infinity, we can now approximate

the c.d.f. of $\tilde{D}_Z^K$ by that of $\tilde{U}_n^K$:

$$\begin{aligned}
\mathbb{P}(t-2\delta \le \tilde{D}_Z^K < t) &\le \mathbb{P}\big(t-2\delta-\epsilon \le \tilde{U}_n^K < t+\epsilon\big) + \frac{\varepsilon_{K;1}}{\epsilon\, n^{-1/2}(n-1)^{-1/2}\sigma}\,, \\
\mathbb{P}(t \le \tilde{D}_Z^K < t+2\delta) &\le \mathbb{P}(t-\epsilon \le \tilde{U}_n^K < t+2\delta+\epsilon) + \frac{\varepsilon_{K;1}}{\epsilon\, n^{-1/2}(n-1)^{-1/2}\sigma}\,, \\
\mathbb{P}(\tilde{D}_Z^K > t) &\le \mathbb{P}\big(\tilde{U}_n^K > t-\epsilon\big) + \frac{\varepsilon_{K;1}}{\epsilon\, n^{-1/2}(n-1)^{-1/2}\sigma}\,, \\
\mathbb{P}(\tilde{D}_Z^K > t) &\ge \mathbb{P}(\tilde{U}_n^K > t+\epsilon) - \frac{\varepsilon_{K;1}}{\epsilon\, n^{-1/2}(n-1)^{-1/2}\sigma}\,.
\end{aligned}$$

Substituting the bounds into the earlier bound and using a similar decomposition to (A.6), we get that the error of approximating the c.d.f. of $\tilde{D}_n$ by that of $\tilde{U}_n^K$ is

$$\begin{aligned}
\big|\mathbb{P}(\tilde{D}_n > t) - \mathbb{P}(\tilde{U}_n^K > t)\big| &\le \max\{\mathbb{P}(t-\epsilon \le \tilde{U}_n^K < t)\,, \mathbb{P}(t < \tilde{U}_n^K \le t+\epsilon)\} \\
&\quad + \max\{\mathbb{P}(t-2\delta-\epsilon \le \tilde{U}_n^K < t+\epsilon)\,, \\
&\qquad \mathbb{P}(t-\epsilon < \tilde{U}_n^K \le t+2\delta+\epsilon)\} \\
&\quad + E_{\delta;K} + \varepsilon'_K + \frac{4\varepsilon_{K;1}}{\epsilon\, n^{-1/2}(n-1)^{-1/2}\sigma}\,.
\end{aligned}$$

To bound the maxima, we recall that by Lemma A.17, there exists some absolute constant C'' such that for any $a \le b \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{P}(a \le \tilde{U}_n^K \le b) \le C''(b-a)^{1/2}\Big(&\frac{1}{\sigma^2}(\sigma_{\text{full}} - \varepsilon_{K;2})^2 \\
&+ \frac{n-1}{\sigma^2}(\sigma_{\text{cond}}^2 - 2\sigma_{\text{cond}}\varepsilon_{K;2} - 4\varepsilon)\Big)^{-1/4}\,.
\end{aligned}$$

Substituting this into the above bound while noting $(2\delta+2\epsilon)^{1/2} \le 2\delta^{1/2} + 2\epsilon^{1/2}$, we get that

$$\begin{aligned}
&\big|\mathbb{P}(\tilde{D}_n > t) - \mathbb{P}(\tilde{U}_n^K > t)\big| \\
&\quad \le C''\big(6\epsilon^{1/2} + 4\delta^{1/2}\big)\Big(\frac{1}{\sigma^2}(\sigma_{\text{full}} - \varepsilon_{K;2})^2 + \frac{n-1}{\sigma^2}(\sigma_{\text{cond}}^2 - 2\sigma_{\text{cond}}\varepsilon_{K;2} - 4\varepsilon_{K;2})\Big)^{-1/4} \\
&\qquad + E_{\delta;K} + \varepsilon'_K + \frac{4\varepsilon_{K;1}}{\epsilon\, n^{-1/2}(n-1)^{-1/2}\sigma}\,.
\end{aligned}$$

We now take $K \to \infty$. By Assumption 3.2, $\varepsilon_{K;2} \to 0$ in the first term and the two trailing error terms vanish. The second error term becomes

$$E_{\delta;K} \;\to\; \frac{C'}{\delta^\nu n^{\nu/2-1}}\Big(\frac{(M_{\text{full};\nu})^\nu}{\sigma^\nu} + \frac{(M_{\text{cond};\nu})^\nu}{(n-1)^{-\nu/2}\,\sigma^\nu}\Big)\,.$$

By additionally taking $\epsilon \to 0$ in the first term and taking a supremum over t on both sides, we then obtain

$$\begin{aligned}
\sup_{t\in\mathbb{R}}\Big|\mathbb{P}(\tilde{D}_n > t) - \lim_{K\to\infty}\mathbb{P}(\tilde{U}_n^K > t)\Big| &\le 4C''\delta^{1/2}\Big(\frac{\sigma_{\text{full}}^2}{\sigma^2} + \frac{\sigma_{\text{cond}}^2}{(n-1)^{-1}\sigma^2}\Big)^{-1/4} \\
&\quad + \frac{C'}{\delta^\nu n^{\nu/2-1}}\Big(\frac{(M_{\text{full};\nu})^\nu}{\sigma^\nu} + \frac{(M_{\text{cond};\nu})^\nu}{(n-1)^{-\nu/2}\,\sigma^\nu}\Big)\,.
\end{aligned}$$

Finally, we choose

$$\delta \;=\; n^{-\frac{\nu-2}{2\nu+1}}\Big(\frac{(M_{\mathrm{full};\nu})^\nu}{\sigma^\nu}+\frac{(M_{\mathrm{cond};\nu})^\nu}{(n-1)^{-\nu/2}\,\sigma^\nu}\Big)^{\frac{2}{2\nu+1}}$$

and $\sigma = \sigma_{\max} \coloneqq \max\{\sigma_{\mathrm{full}}, (n-1)^{1/2}\sigma_{\mathrm{cond}}\}$. Then $\big(\frac{\sigma_{\mathrm{full}}^2}{\sigma^2}+\frac{\sigma_{\mathrm{cond}}^2}{(n-1)^{-1}\sigma^2}\big)^{-1/4} \le 1$, and by redefining constants, we get that there exists some absolute constant $C>0$ such that

$$\begin{aligned}
&\sup_{t\in\mathbb{R}}\Big|\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\max}}D_n > t\Big) - \lim_{K\to\infty}\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\max}}U_n^K > t\Big)\Big| \\
&\le\; C\,n^{-\frac{\nu-2}{4\nu+2}}\Big(\frac{(M_{\mathrm{full};\nu})^\nu}{\sigma_{\max}^\nu}+\frac{(M_{\mathrm{cond};\nu})^\nu}{(n-1)^{-\nu/2}\,\sigma_{\max}^\nu}\Big)^{\frac{1}{2\nu+1}} \qquad \text{(A.7)}\\
&\le\; 2^{\frac{1}{2\nu+1}}C\,n^{-\frac{\nu-2}{4\nu+2}}\Big(\frac{M_{\max;\nu}}{\sigma_{\max}}\Big)^{\frac{\nu}{2\nu+1}}\,,
\end{aligned}$$

where we have recalled that $M_{\max;\nu} \coloneqq \max\{M_{\mathrm{full};\nu}, (n-1)^{1/2}M_{\mathrm{cond};\nu}\}$. This finishes the proof.

A.3.3. Proof of Lemma A.14

Proof overview. The proof idea is reminiscent of the standard technique for proving that convergence in probability implies weak convergence. We first approximate each probability by the expectation of a δ^{-1} Lipschitz function h that is uniformly bounded by 1. This introduces an approximation error of δ, while replaces the difference in probability by the difference $\mathbb{E}[h(\tilde{D}_n) - h(\tilde{D}_n^K)]$. The expectation can be further split by the events $\{|\tilde{D}_n - \tilde{D}_n^K| < \epsilon\}$ and $\{|\tilde{D}_n - \tilde{D}_n^K| \ge \epsilon\}$. In the first case, the expectation can be bounded by a Lipschitz argument; in the second case, we can use the boundedness of h to bound the expectation by $2\mathbb{P}(|\tilde{D}_n - \tilde{D}_n^K| \ge \epsilon)$, which is in turn bounded by a Markov argument to give the "spectral" approximation error. Choosing ϵ appropriately gives the above error term.

Proof of Lemma A.14. For any $\tau \in \mathbb{R}$ and $\delta > 0$, let $h_{\tau;\delta}$ be the function defined in Lemma A.10 with $m = 0$, which satisfies

$$h_{\tau+\delta;\delta}(x) \;\le\; \mathbb{I}_{\{x>\tau\}} \;\le\; h_{\tau;\delta}(x)\,.$$

By applying the above bounds with τ set to t and $t-\delta$, we get that

$$\mathbb{P}(\tilde{D}_n > t) - \mathbb{P}\big(\tilde{D}_n^K > t-\delta\big) \;= \mathbb{E}[\mathbb{I}_{\{\tilde{D}_n>t\}} - \mathbb{I}_{\{\tilde{D}_n^K>t-\delta\}}] \;\le\; \mathbb{E}[h_{t;\delta}(\tilde{D}_n) - h_{t;\delta}(\tilde{D}_n^K)]\,,$$

and similarly

$$\mathbb{P}\big(\tilde{D}_n^K > t+\delta\big) - \mathbb{P}(\tilde{D}_n > t) \;\le\; \mathbb{E}[h_{t+\delta;\delta}(\tilde{D}_n^K) - h_{t+\delta;\delta}(\tilde{D}_n)]\,.$$

Therefore, defining $\xi_\tau := |\mathbb{E}[h_{\tau;\delta}(\tilde{D}_n) - h_{\tau;\delta}(\tilde{D}_n^K)]|$, we get that

$$\mathbb{P}\big(\tilde{D}_n^K > t+\delta\big) - \xi_{t+\delta} \;\le\; \mathbb{P}(\tilde{D}_n > t) \;\le\; \mathbb{P}\big(\tilde{D}_n^K > t-\delta\big) + \xi_t \,.$$

To bound quantities of the form ξ_τ, fix any $\epsilon > 0$ and write $\xi_\tau = \xi_{\tau,1} + \xi_{\tau,2}$ where

$$\begin{aligned}
\xi_{\tau,1} &:= \left|\mathbb{E}\Big[\big(h_{\tau;\delta}(\tilde{D}_n) - h_{\tau;\delta}(\tilde{D}_n^K)\big)\mathbb{I}_{\{|\tilde{D}_n - \tilde{D}_n^K| \le \epsilon\}}\Big]\right| , \\
\xi_{\tau,2} &:= \left|\mathbb{E}\Big[\big(h_{\tau;\delta}(\tilde{D}_n) - h_{\tau;\delta}(\tilde{D}_n^K)\big)\mathbb{I}_{\{|\tilde{D}_n - \tilde{D}_n^K| > \epsilon\}}\Big]\right| .
\end{aligned}$$

The first term can be bounded by recalling from Lemma A.10 that $h_{\tau;\delta}$ is δ^{-1}-Lipschitz:

$$\xi_{\tau,1} \;\le \delta^{-1}\mathbb{E}\big[|\tilde{D}_n - \tilde{D}_n^K|\mathbb{I}_{\{|\tilde{D}_n - \tilde{D}_n^K| \le \epsilon\}}\big] \;\le\; \delta^{-1}\epsilon\,\mathbb{P}\big(|\tilde{D}_n - \tilde{D}_n^K| \le \epsilon\big) \;\le\; \delta^{-1}\epsilon \,.$$

The second term can be bounded by noting that $h_{\tau;\delta}$ is uniformly bounded above by 1 and applying Markov's inequality:

$$\xi_{\tau,2} \;\le\; 2\mathbb{E}[\mathbb{I}_{\{|\tilde{D}_n - \tilde{D}_n^K| > \epsilon\}}] \;=\; 2\mathbb{P}(|\tilde{D}_n - \tilde{D}_n^K| > \epsilon) \;\le\; 2\epsilon^{-1}\mathbb{E}\big[|\tilde{D}_n - \tilde{D}_n^K|\big] \,.$$

By the definition of $\tilde{D}_n$ and $\tilde{D}_n^K$, the triangle inequality and noting that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d. , the absolute moment term can be bounded as

$$\begin{aligned}
\mathbb{E}\big[|\tilde{D}_n - \tilde{D}_n^K|\big] &= \frac{\sqrt{n(n-1)}}{\sigma}\,\mathbb{E}\big[|D_n - D_n^K|\big] \\
&= \frac{1}{\sigma\sqrt{n(n-1)}}\Big\| \textstyle\sum_{1\le i\ne j\le n}\big(u(\mathbf{X}_i,\mathbf{X}_j) - \sum_{k=1}^K \lambda_k\phi_k(\mathbf{X}_i)\phi_k(\mathbf{X}_j)\big)\Big\|_{L_1} \\
&\le \frac{\sqrt{n(n-1)}}{\sigma}\Big\|u(\mathbf{X}_1,\mathbf{X}_2) - \textstyle\sum_{k=1}^K \lambda_k\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\Big\|_{L_1} \\
&= \sigma^{-1}\sqrt{n(n-1)}\,\varepsilon_{K;1} \,.
\end{aligned}$$

Combining the bounds on $\xi_{\tau,1}, \xi_{\tau,2}$ and $\mathbb{E}[|\tilde{D}_n - \tilde{D}_n^K|]$ and choosing

$$\epsilon \;=\; \big(\sqrt{n(n-1)}\sigma^{-1}\delta\varepsilon_{K;1}\big)^{1/2} \,,$$

we get that

$$\xi_\tau \;\le \delta^{-1}\epsilon + 2\sqrt{n(n-1)}\,\epsilon^{-1}\sigma^{-1}\varepsilon_{K;1} \;=\; \frac{3n^{1/4}(n-1)^{1/4}\varepsilon_{K;1}^{1/2}}{\sigma^{1/2}\delta^{1/2}} \;=:\; \varepsilon'_K \,,$$

which yields the desired bound

$$\mathbb{P}(\tilde{D}_n^K > t+\delta) - \varepsilon'_K \;\le\; \mathbb{P}(\tilde{D}_n > t) \;\le\; \mathbb{P}(\tilde{D}_n^K > t-\delta) + \varepsilon'_K \,.$$

□

A.3.4. Proof of Lemma A.15

For convenience, we denote $\mathbf{V}_i := \phi^K(\mathbf{X}_i)$ throughout this section.

Proof overview. The key idea in the proof rests on Lindeberg's telescoping sum argu-

ment for central limit theorem. We follow Chatterjee (2006)'s adaptaion of the Lindeberg idea for statistics that are not asymptotically normal. As before, the difference in probability is first approximated by a difference in expectation $\mathbb{E}[h(\tilde{D}_n^K) - h(\tilde{D}_Z^K)]$ with respect to some function h, which introduces a further approximation error δ. The next step is to note that both $\tilde{D}_n^K$ and $\tilde{D}_Z^K$ can be expressed in terms of some common function $\tilde{f}$, such that

$$\tilde{D}_n^K \;=\; \tilde{f}(\mathbf{V}_1, \ldots, \mathbf{V}_n)\;, \qquad\qquad \tilde{D}_Z^K \;=\; \tilde{f}(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)\;.$$

Denoting $g = h \circ \tilde{f}$, we can then write the difference in expectation in terms of Lindeberg's telescoping sum as

$$\begin{aligned}\mathbb{E}[h(\tilde{D}_n^K) - h(\tilde{D}_Z^K)] \;&= \mathbb{E}[g(\mathbf{V}_1, \ldots, \mathbf{V}_1) - g(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)] \\ &= \textstyle\sum_{i=1}^n \big(\mathbb{E}[g(\mathbf{V}_1, \ldots, \mathbf{V}_{i-1}, \mathbf{V}_i, \mathbf{Z}_{i+1}, \ldots, \mathbf{Z}_n) \\ &\qquad\qquad - g(\mathbf{V}_1, \ldots, \mathbf{V}_{i-1},\; \mathbf{Z}_i,\; \mathbf{Z}_{i+1}, \ldots, \mathbf{Z}_n)]\big)\;.\end{aligned}$$

Since each summand differs only in the i-th argument, we can perform a second-order Taylor expansion about the i-th argument provided that the function h such that h is twice-differentiable. The second-order remainder term is further "Taylor-expanded" to an additional ϵ-order for any $\epsilon \in [0, 1]$ by choosing h'' to be ϵ-Hölder. Write D_i as the differential operator with respect to the i-th argument and denote

$$\tilde{f}_i(\mathbf{x}) \;\coloneqq\; \tilde{f}(\mathbf{V}_1, \ldots, \mathbf{V}_{i-1}, \mathbf{x}, \mathbf{Z}_{i+1}, \ldots, \mathbf{Z}_n)\;.$$

Then informally speaking, the Taylor expansion argument amounts to bounding each summand as

$$\begin{aligned}\big|(\text{summand})_i\big| \;&\le \mathbb{E}[D_i(h \circ \tilde{f}_i)(0)(\mathbf{V}_i - \mathbf{Z}_i)] + \frac{1}{2}\mathbb{E}[D_i^2(h \circ \tilde{f}_i)(0)(\mathbf{V}_i^2 - \mathbf{Z}_i^2)] \\ &\quad + \frac{1}{6}\big(\text{Hölder constant of } h''\big) \times \mathbb{E}\big[\big|D_i\tilde{f}_i(0)\mathbf{V}_i\big|^{2+\epsilon} + \big|D_i\tilde{f}_i(0)\mathbf{Z}_i\big|^{2+\epsilon}\big]\;,\end{aligned}$$

where we have used the fact that $\tilde{f}_i$ is a linear function in expressing the last quantity. The first two terms vanish because $h \circ \tilde{f}_i$ is independent of $\big(\mathbf{V}_i, \mathbf{Z}_i\big)$ and the first two moments of $\mathbf{V}_i$ and $\mathbf{Z}_i$ match. The third term is bounded carefully by noting the moment structure of $\mathbf{V}_i$ and $\mathbf{Z}_i$ to give the error term $\frac{1}{n}E_{\delta;K}$. Summing the errors over $1 \le i \le n$ then gives the Gaussian approximation error bound in Lemma A.15.

Proof of Lemma A.15. For any $\tau \in \mathbb{R}$ and $\delta > 0$, let $h_{\tau;\delta}$ be the twice-differentiable function defined in Lemma A.10 (i.e. $m = 2$), which satisfies

$$h_{\tau+\delta;\delta}(x) \;\le\; \mathbb{I}_{\{x>\tau\}} \;\le\; h_{\tau;\delta}(x)\;.$$

By applying the above bounds with τ set to $t-\delta$ and $t-2\delta$, we get that

$$\begin{aligned}\mathbb{P}(\tilde{D}_n^K > t-\delta) - \mathbb{P}(\tilde{D}_Z^K > t-2\delta) &= \mathbb{E}[\mathbb{I}_{\{\tilde{D}_n^K > t-\delta\}} - \mathbb{I}_{\{\tilde{D}_Z^K > t-2\delta\}}] \\ &\leq \mathbb{E}[h_{t-\delta;\delta}(\tilde{D}_n^K) - h_{t-\delta;\delta}(\tilde{D}_Z^K)] \,,\end{aligned}$$

and similarly

$$\begin{aligned}\mathbb{P}(\tilde{D}_Z^K > t+2\delta) - \mathbb{P}(\tilde{D}_n^K > t+\delta) &= \mathbb{E}[\mathbb{I}_{\{\tilde{D}_Z^K > t+2\delta\}} - \mathbb{I}_{\{\tilde{D}_n^K > t+\delta\}}] \\ &\leq \mathbb{E}[h_{t+2\delta;\delta}(\tilde{D}_Z^K) - h_{t+2\delta;\delta}(\tilde{D}_n^K)] \,.\end{aligned}$$

Therefore, we obtain that

$$\begin{aligned}\mathbb{P}(\tilde{D}_n^K > t-\delta) &\leq \mathbb{P}(\tilde{D}_Z^K > t-2\delta) + E'_{\delta;K} \,, \\ \mathbb{P}(\tilde{D}_n^K > t+\delta) &\geq \mathbb{P}(\tilde{D}_Z^K > t+2\delta) - E'_{\delta;K} \,,\end{aligned} \tag{A.8}$$

where $E'_{\delta;K} \coloneqq \sup_{\tau\in\mathbb{R}} |\mathbb{E}[h_{\tau;\delta}(\tilde{D}_n^K) - h_{\tau;\delta}(\tilde{D}_Z^K)]|$. The next step is to bound $E'_{\delta;K}$, to which we apply the Lindeberg method for proving central limit theorem. We denote the scaled mean as

$$\tilde{\mu} \coloneqq \frac{\mathbb{E}[\mathbf{V}_1]}{\sigma^{1/2}(n(n-1))^{1/4}} = \frac{\mathbb{E}[\mathbf{Z}_1]}{\sigma^{1/2}(n(n-1))^{1/4}} \,,$$

and define the centred and scaled versions of $\mathbf{V}_i$ and $\mathbf{Z}_i$ respectively as

$$\tilde{\mathbf{V}}_i \coloneqq \frac{\mathbf{V}_i}{\sigma^{1/2}(n(n-1))^{1/4}} - \tilde{\mu} \,, \qquad \tilde{\mathbf{Z}}_i \coloneqq \frac{\mathbf{Z}_i}{\sigma^{1/2}(n(n-1))^{1/4}} - \tilde{\mu} \,.$$

We also define the function $f : (\mathbb{R}^K)^n \to \mathbb{R}$ by

$$f(\mathbf{v}_1, \ldots, \mathbf{v}_n) \coloneqq \sum\nolimits_{1\leq i\neq j\leq n} (\mathbf{v}_i + \tilde{\mu})^\top \Lambda^K (\mathbf{v}_j + \tilde{\mu})$$

where we recall $\Lambda^K \coloneqq \mathrm{diag}\{\lambda_1, \ldots, \lambda_K\}$. This allows us to express the random quantities in (A.8) as

$$\tilde{D}_n^K = f(\tilde{\mathbf{V}}_1, \ldots, \tilde{\mathbf{V}}_n) \,, \qquad \tilde{D}_Z^K = f(\tilde{\mathbf{Z}}_1, \ldots, \tilde{\mathbf{Z}}_n) \,.$$

By defining the random function

$$F_i(\mathbf{v}) \coloneqq f(\tilde{\mathbf{V}}_1, \ldots, \tilde{\mathbf{V}}_{i-1}, \mathbf{v}, \tilde{\mathbf{Z}}_{i+1}, \ldots, \tilde{\mathbf{Z}}_n) \qquad \text{for } \mathbf{v} \in \mathbb{R}^K \text{ and } 1 \leq i \leq n \,,$$

we can write $E'_{\delta;K}$ into Lindeberg's telescoping sum as

$$\begin{aligned}E'_{\delta;K} &= \sup_{\tau\in\mathbb{R}} |\mathbb{E}[h_{\tau;\delta} \circ f(\tilde{\mathbf{V}}_1, \ldots, \tilde{\mathbf{V}}_n) - h_{\tau;\delta} \circ f(\tilde{\mathbf{Z}}_1, \ldots, \tilde{\mathbf{Z}}_n)]| \\ &= \sup_{\tau\in\mathbb{R}} \left| \sum\nolimits_{i=1}^n \mathbb{E}[h_{\tau;\delta}(F_i(\tilde{\mathbf{V}}_i) - h_{\tau;\delta}(F_i(\tilde{\mathbf{Z}}_i))] \right| \\ &\leq \sup_{\tau\in\mathbb{R}} \sum\nolimits_{i=1}^n |\mathbb{E}[h_{\tau;\delta} \circ F_i(\tilde{\mathbf{V}}_i) - h_{\tau;\delta} \circ F_i(\tilde{\mathbf{Z}}_i)]| \,.\end{aligned}$$

Since $h_{\tau;\delta} \circ f$ is twice-differentiable, by a second-order Taylor expansion around $\mathbf{0} \in \mathbb{R}^K$,

there exists random values $\theta_V, \theta_Z \in (0,1)$ almost surely such that

$$\begin{aligned} h_{\tau;\delta} \circ F_i(\tilde{\mathbf{V}}_i) &= \frac{\partial h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}}\Big|_{\mathbf{x}=\mathbf{0}} \tilde{\mathbf{V}}_i + \frac{1}{2}\frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2}\Big|_{\mathbf{x}=\theta_V \tilde{\mathbf{V}}_i} \tilde{\mathbf{V}}_i^{\otimes 2} , \\ h_{\tau;\delta} \circ F_i(\tilde{\mathbf{Z}}_i) &= \frac{\partial h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}}\Big|_{\mathbf{x}=\mathbf{0}} \tilde{\mathbf{Z}}_i + \frac{1}{2}\frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2}\Big|_{\mathbf{x}=\theta_Z \tilde{\mathbf{Z}}_i} \tilde{\mathbf{Z}}_i^{\otimes 2} . \end{aligned}$$

Substituting this into the sum above gives

$$\begin{aligned} E'_{\delta;K} \leq \sup_{\tau\in\mathbb{R}} \Big(& \sum_{i=1}^n \Big| \mathbb{E}\Big[\frac{\partial h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}}\Big|_{\mathbf{x}=\mathbf{0}} (\tilde{\mathbf{V}}_i - \tilde{\mathbf{Z}}_i) \Big] \Big| \\ & + \frac{1}{2}\sum_{i=1}^n \Big| \mathbb{E}\Big[\frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2}\Big|_{\mathbf{x}=\theta_V \tilde{\mathbf{V}}_i} \tilde{\mathbf{V}}_i^{\otimes 2} - \frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2}\Big|_{\mathbf{x}=\theta_Z \tilde{\mathbf{Z}}_i} \tilde{\mathbf{Z}}_i^{\otimes 2} \Big] \Big| \Big) . \end{aligned}$$

The first sum vanishes because the only randomness of the derivative comes from F_i, who is independent of $(\tilde{\mathbf{V}}_i, \tilde{\mathbf{Z}}_i)$, and the mean of $\tilde{\mathbf{V}}_i$ and $\tilde{\mathbf{Z}}_i$ match. To handle the second sum, we make use of independence again and the fact that the second moment of $\tilde{\mathbf{V}}_i$ and $\tilde{\mathbf{Z}}_i$ also match: By subtracting and adding the term

$$\mathbb{E}\Big[\frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2}\Big|_{\mathbf{x}=\mathbf{0}} (\tilde{\mathbf{V}}_i)^{\otimes 2} \Big] = \mathbb{E}\Big[\frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2}\Big|_{\mathbf{x}=\mathbf{0}} (\tilde{\mathbf{Z}}_i)^{\otimes 2} \Big] ,$$

we can apply the triangle inequality to get that

$$\begin{aligned} E'_{\delta;K} \leq \tfrac{1}{2} \sup_{\tau\in\mathbb{R}} \Big(& \textstyle\sum_{i=1}^n \Big| \mathbb{E}\Big[\Big(\frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2}\Big|_{\mathbf{x}=\theta_V \tilde{\mathbf{V}}_i} - \frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2}\Big|_{\mathbf{x}=\mathbf{0}} \Big) \tilde{\mathbf{V}}_i^{\otimes 2} \Big] \Big| \\ & + \textstyle\sum_{i=1}^n \Big| \mathbb{E}\Big[\Big(\frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2}\Big|_{\mathbf{x}=\theta_Z \tilde{\mathbf{Z}}_i} - \frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2}\Big|_{\mathbf{x}=\mathbf{0}} \Big) \tilde{\mathbf{Z}}_i^{\otimes 2} \Big] \Big| \Big) . \end{aligned} \tag{A.9}$$

The final step is to bound the two sums by exploiting the derivative structure of $h_{\tau;\delta}$ and F_i. Note that F_i is a linear function: its first derivative is given by

$$\partial F_i(\mathbf{x}) = 2\textstyle\sum_{1\leq j<i} \Lambda^K \tilde{\mathbf{V}}_j + 2\sum_{i<j\leq n} \Lambda^K \tilde{\mathbf{Z}}_j + 2(n-1)\Lambda^K \tilde{\mu} \in \mathbb{R}^K ,$$

which is independent of $\mathbf{x}$, while its higher derivatives vanish. By a second-order chain rule, this implies that almost surely

$$\begin{aligned} & \Big| \Big(\frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2}\Big|_{\mathbf{x}=\theta_V \tilde{\mathbf{V}}_i} - \frac{\partial^2 h_{\tau;\delta} \circ F_i(\mathbf{x})}{\partial \mathbf{x}^2}\Big|_{\mathbf{x}=\mathbf{0}} \Big) \tilde{\mathbf{V}}_i^{\otimes 2} \Big| \\ &= \Big| \big(\partial^2 h_{\tau;\delta}\big(F_i(\theta_V \tilde{\mathbf{V}}_i)\big) - \partial^2 h_{\tau;\delta}\big(F_i(\mathbf{0})\big) \big) (\partial F_i(\mathbf{0})^\top \tilde{\mathbf{V}}_i)^2 \Big| \\ &\leq \big| \partial^2 h_{\tau;\delta}\big(F_i(\theta_V \tilde{\mathbf{V}}_i)\big) - \partial^2 h_{\tau;\delta}\big(F_i(\mathbf{0})\big) \big| \, \big| \partial F_i(\mathbf{0})^\top \tilde{\mathbf{V}}_i \big|^2 . \end{aligned}$$

For $\nu \in (2,3]$, by the Hölder property of $\partial^2 h_{\tau;\delta}$ from Lemma A.10, we get that almost surely,

$$\begin{aligned} \big| \partial^2 h_{\tau;\delta}(F_i(\theta_V \tilde{\mathbf{V}}_i)) - \partial^2 h_{\tau;\delta}(F_i(\mathbf{0})) \big| &\leq 18 \times 3^{\nu-2} \delta^{-\nu} |F_i(\theta_V \tilde{\mathbf{V}}_i) - F_i(\mathbf{0})|^{\nu-2} \\ &= 18 \times 3^{\nu-2} \delta^{-\nu} |\partial F_i(\mathbf{0})^\top (\theta_V \tilde{\mathbf{V}}_i)|^{\nu-2} \\ &\leq 54 \delta^{-\nu} |\partial F_i(\mathbf{0})^\top \tilde{\mathbf{V}}_i|^{\nu-2} . \end{aligned}$$

In the last inequality, we have used that θ_V takes value in $[0,1]$. Combining the results, we get that each summand in the first sum in (A.9) can be bounded as

$$\Big|\mathbb{E}\Big[\Big(\frac{\partial^2 h_{\tau;\delta}\circ F_i(\mathbf{x})}{\partial \mathbf{x}^2}\Big|_{\mathbf{x}=\theta_V\tilde{\mathbf{V}}_i} - \frac{\partial^2 h_{\tau;\delta}\circ F_i(\mathbf{x})}{\partial \mathbf{x}^2}\Big|_{\mathbf{x}=\mathbf{0}}\Big)\tilde{\mathbf{V}}_i^{\otimes 2}\Big]\Big| \;\le 54\delta^{-\nu}\mathbb{E}\big[|\partial F_i(\mathbf{0})^\top\tilde{\mathbf{V}}_i|^\nu\big] \;.$$

The exact same argument applies to the summands of the second sum to give

$$\Big|\mathbb{E}\Big[\Big(\frac{\partial^2 h_{\tau;\delta}\circ F_i(\mathbf{x})}{\partial \mathbf{x}^2}\Big|_{\mathbf{x}=\theta_Z\tilde{\mathbf{Z}}_i} - \frac{\partial^2 h_{\tau;\delta}\circ F_i(\mathbf{x})}{\partial \mathbf{x}^2}\Big|_{\mathbf{x}=\mathbf{0}}\Big)\tilde{\mathbf{Z}}_i^{\otimes 2}\Big]\Big| \;\le 54\delta^{-\nu}\mathbb{E}\big[|\partial F_i(\mathbf{0})^\top\tilde{\mathbf{Z}}_i|^\nu\big] \;,$$

so a substitution back into (A.9) gives

$$E'_{\delta;K} \;\le 27\delta^{-\nu}\sum_{i=1}^n\Big(\mathbb{E}\big[|\partial F_i(\mathbf{0})^\top\tilde{\mathbf{V}}_i|^\nu\big] + \mathbb{E}\big[|\partial F_i(\mathbf{0})^\top\tilde{\mathbf{V}}_i|^\nu\big]\Big) \;.$$

We defer to Lemma A.18 to show that there exists an absolute constant $C'>0$ such that the moment terms can be bounded as

$$\mathbb{E}\big[|\partial F_i(\mathbf{0})^\top\tilde{\mathbf{V}}_i|^\nu\big] + \mathbb{E}\big[|\partial F_i(\mathbf{0})^\top\tilde{\mathbf{Z}}_i|^\nu\big] \;\le \frac{C'}{n^{\nu/2}}\Big(\frac{(M_{\mathrm{full};\nu})^\nu+\varepsilon_{K;\nu}^\nu}{\sigma^\nu} + \frac{(M_{\mathrm{cond};\nu})^\nu+\varepsilon_{K;\nu}^\nu}{(n-1)^{-\nu/2}\,\sigma^\nu}\Big) \;. \tag{A.10}$$

Combining with (A.8) and defining $E_{\delta;K}$ to be the upper bound for $E'_{\delta;K}$, we get that

$$\begin{aligned}\mathbb{P}(\tilde{D}_n^K > t-\delta) \;&\le \mathbb{P}(\tilde{D}_Z^K > t-2\delta) + E_{\delta;K} \;,\\ \mathbb{P}(\tilde{D}_n^K > t+\delta) \;&\ge \mathbb{P}(\tilde{D}_Z^K > t+2\delta) - E_{\delta;K} \;,\end{aligned}$$

where we have made the K-dependence explicit and define, for $C \coloneqq 27C'$,

$$E_{\delta;K} \;\coloneqq \frac{C}{\delta^\nu n^{\nu/2-1}}\Big(\frac{(M_{\mathrm{full};\nu})^\nu+\varepsilon_{K;\nu}^\nu}{\sigma^\nu} + \frac{(M_{\mathrm{cond};\nu})^\nu+\varepsilon_{K;\nu}^\nu}{(n-1)^{-\nu/2}\,\sigma^\nu}\Big) \;.$$

□

Lemma A.18. (A.10) *holds.*

Proof of Lemma A.18. We seek to bound $\mathbb{E}[|\partial F_i(\mathbf{0})^\top\tilde{\mathbf{V}}_i|^\nu] + \mathbb{E}[|\partial F_i(\mathbf{0})^\top\tilde{\mathbf{Z}}_i|^\nu]$ for $\nu\in(2,3]$ and

$$\partial F_i(\mathbf{0}) \;=\; 2\sum_{1\le j<i}\Lambda^K\tilde{\mathbf{V}}_j + 2\sum_{i<j\le n}\Lambda^K\tilde{\mathbf{Z}}_j + 2(n-1)\Lambda^K\tilde{\mu} \;\in\; \mathbb{R}^K \;.$$

We first focus on bounding the first expectation. By convexity of the function $x\mapsto|x|^\nu$, we can apply Jensen's inequality to bound

$$\begin{aligned}\mathbb{E}\big[|\partial F_i(\mathbf{0})^\top\tilde{\mathbf{V}}_i|^\nu\big] \;&=\; \mathbb{E}\Big[\Big|2\sum_{j<i}\tilde{\mathbf{V}}_j^\top\Lambda^K\tilde{\mathbf{V}}_i + 2\sum_{j>i}\tilde{\mathbf{Z}}_j^\top\Lambda^K\tilde{\mathbf{V}}_i + 2(n-1)\tilde{\mu}^\top\Lambda^K\tilde{\mathbf{V}}_i\Big|^\nu\Big]\\ &\le \tfrac{1}{3}\mathbb{E}\big[\big|6\sum_{j<i}\tilde{\mathbf{V}}_j^\top\Lambda^K\tilde{\mathbf{V}}_i\big|^\nu\big] + \tfrac{1}{3}\mathbb{E}\big[\big|6\sum_{j>i}\tilde{\mathbf{Z}}_j^\top\Lambda^K\tilde{\mathbf{V}}_i\big|^\nu\big] + \tfrac{1}{3}\mathbb{E}\big[\big|6(n-1)\tilde{\mu}^\top\Lambda^K\tilde{\mathbf{V}}_1\big|^\nu\big]\\ &\le 72\big(\mathbb{E}\big[\big|\sum_{j<i}\tilde{\mathbf{V}}_j^\top\Lambda^K\tilde{\mathbf{V}}_i\big|^\nu\big] + \mathbb{E}\big[\big|\sum_{j>i}\tilde{\mathbf{Z}}_j^\top\Lambda^K\tilde{\mathbf{V}}_i\big|^\nu\big] + \mathbb{E}\big[\big|(n-1)\tilde{\mu}^\top\Lambda^K\tilde{\mathbf{V}}_1\big|^\nu\big]\big) \;,\end{aligned}$$

where we have noted that $\nu\le 3$. Since $\tilde{\mathbf{V}}_i$'s are i.i.d., $\tilde{\mathbf{Z}}_i$'s are i.i.d. and all variables involved are zero-mean, $(\tilde{\mathbf{V}}_j^\top\Lambda^K\tilde{\mathbf{V}}_i)_{j=1}^{i-1}$ forms a martingale difference sequence with respect to the filtration $\sigma(\tilde{\mathbf{V}}_i,\tilde{\mathbf{V}}_1),\ldots,\sigma(\tilde{\mathbf{V}}_i,\tilde{\mathbf{V}}_1,\ldots,\tilde{\mathbf{V}}_{i-1})$, and so is $(\tilde{\mathbf{Z}}_j^\top\Lambda^K\tilde{\mathbf{V}}_i)_{j=i+1}^n$

with respect to the filtration $\sigma(\tilde{\mathbf{V}}_i, \tilde{\mathbf{Z}}_{i+1}), \ldots, \sigma(\tilde{\mathbf{V}}_i, \tilde{\mathbf{Z}}_{i+1}, \ldots, \tilde{\mathbf{Z}}_n)$. This allows the above two moments of sums to be bounded via the martigale moment inequality from Lemma A.4: There exists an absolute constant $C_0 > 0$ such that

$$\begin{aligned}
&\mathbb{E}\big[|\partial F_i(\mathbf{0})^\top \tilde{\mathbf{V}}_i|^\nu\big] \\
&\le C_0\Big((i-1)^{\nu/2-1}\textstyle\sum_{j=1}^{i-1}\mathbb{E}[|\tilde{\mathbf{V}}_j^\top \Lambda^K \tilde{\mathbf{V}}_i|^\nu] + (n-i)^{\nu/2-1}\sum_{j=i+1}^{n}\mathbb{E}[|\tilde{\mathbf{Z}}_j^\top \Lambda^K \tilde{\mathbf{V}}_i|^\nu] \\
&\qquad + (n-1)^\nu\,\mathbb{E}[|\tilde{\mu}^\top \Lambda^K \tilde{\mathbf{V}}_1|^\nu]\Big) \\
&\le C_0(n-1)^{\nu/2}\Big(\mathbb{E}[|\tilde{\mathbf{V}}_1^\top \Lambda^K \tilde{\mathbf{V}}_2|^\nu] + \mathbb{E}[|\tilde{\mathbf{Z}}_1^\top \Lambda^K \tilde{\mathbf{V}}_1|^\nu] + (n-1)^{\nu/2}\mathbb{E}[|\tilde{\mu}^\top \Lambda^K \tilde{\mathbf{V}}_1|^\nu]\Big)\,.
\end{aligned}$$

By the exact same argument, the other expectation we want to bound can also be controlled as

$$\begin{aligned}
&\mathbb{E}\big[|\partial F_i(\mathbf{0})^\top \tilde{\mathbf{Z}}_i|^\nu\big] \\
&\quad\le\; C_0(n-1)^{\nu/2}\Big(\mathbb{E}[|\tilde{\mathbf{Z}}_1^\top \Lambda^K \tilde{\mathbf{Z}}_2|^\nu] + \mathbb{E}[|\tilde{\mathbf{Z}}_1^\top \Lambda^K \tilde{\mathbf{V}}_1|^\nu] + (n-1)^{\nu/2}\mathbb{E}[|\tilde{\mu}^\top \Lambda^K \tilde{\mathbf{Z}}_1|^\nu]\Big)\,.
\end{aligned}$$

Finally, we relate these moments terms to moments of $u(\mathbf{X}_1, \mathbf{X}_2)$, up to error terms that vanish as $K \to \infty$: Denoting $\mu_k \coloneqq \mathbb{E}[\phi_k(\mathbf{X}_1)]$, we have that by Lemma A.7,

$$\begin{aligned}
\mathbb{E}[|\tilde{\mu}^\top \Lambda^K \tilde{\mathbf{V}}_1|^\nu] \;&=\; \frac{1}{\sigma^\nu n^{\nu/2}(n-1)^{\nu/2}}\mathbb{E}\Big[\Big|\textstyle\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1)-\mu_k)\mu_k\Big|^\nu\Big] \\
&\le \frac{4((M_{\mathrm{cond};\nu})^\nu + \varepsilon_{K;\nu}^\nu)}{\sigma^\nu\, n^{\nu/2}(n-1)^{\nu/2}}\,,
\end{aligned}$$

and for some absolute constant $C_1 > 0$,

$$\begin{aligned}
\mathbb{E}[|\tilde{\mathbf{V}}_1^\top \Lambda^K \tilde{\mathbf{V}}_2|^\nu] \;&=\; \frac{1}{\sigma^\nu n^{\nu/2}(n-1)^{\nu/2}}\mathbb{E}\Big[\Big|\textstyle\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1)-\mu_k)(\phi_k(\mathbf{X}_2)-\mu_k)\Big|^\nu\Big] \\
&\le \frac{4C_1(M_{\mathrm{full};\nu})^\nu - \frac{1}{2}(M_{\mathrm{cond};\nu})^\nu + (4C_1+2)\varepsilon_{K;\nu}^\nu}{\sigma^\nu n^{\nu/2}(n-1)^{\nu/2}} \\
&\le \frac{4C_1(M_{\mathrm{full};\nu})^\nu + (4C_1+2)\varepsilon_{K;\nu}^\nu}{\sigma^\nu n^{\nu/2}(n-1)^{\nu/2}}\,.
\end{aligned}$$

For the moment terms involving the Gaussians $\tilde{\mathbf{Z}}_1$ and $\tilde{\mathbf{Z}}_2$, we apply Lemma A.8 to show that

$$\begin{aligned}
\mathbb{E}[|\tilde{\mu}^\top \Lambda^K \tilde{\mathbf{Z}}_1|^\nu] \;&=\; \frac{\mathbb{E}[|(\mathbb{E}[\mathbf{V}_1])^\top \Lambda^K \mathbf{Z}_1|^\nu]}{\sigma^\nu n^{\nu/2}(n-1)^{\nu/2}} \;\le\; \frac{7(\sigma_{\mathrm{cond}}^\nu + 8\varepsilon_{K;2}^\nu)}{\sigma^\nu n^{\nu/2}(n-1)^{\nu/2}} \;\le\; \frac{7((M_{\mathrm{cond};\nu})^\nu + 8\varepsilon_{K;\nu}^\nu)}{\sigma^\nu n^{\nu/2}(n-1)^{\nu/2}}\,, \\
\mathbb{E}[|\tilde{\mathbf{Z}}_1^\top \Lambda^K \tilde{\mathbf{Z}}_2|^\nu] \;&=\; \frac{\mathbb{E}[|\mathbf{Z}_1^\top \Lambda^K \mathbf{Z}_2|^\nu]}{\sigma^\nu n^{\nu/2}(n-1)^{\nu/2}} \;\le\; \frac{6(\sigma_{\mathrm{full}}^\nu + \varepsilon_{K;2}^\nu)}{\sigma^\nu n^{\nu/2}(n-1)^{\nu/2}} \;\le\; \frac{6((M_{\mathrm{full};\nu})^\nu + \varepsilon_{K;\nu}^\nu)}{\sigma^\nu n^{\nu/2}(n-1)^{\nu/2}}\,.
\end{aligned}$$

In the last inequalities for both bounds, we have noted that L_2 norm is dominated by L_ν norm since $\nu > 2$. Meanwhile by Lemma A.8 again, there exists some absolute constant $C_2 > 0$ such that

$$\mathbb{E}[|\tilde{\mathbf{Z}}_1^\top \Lambda^K \tilde{\mathbf{V}}_1|^\nu] \;=\; \frac{\mathbb{E}[|(\mathbf{V}_1 - \mathbb{E}[\mathbf{V}_1])^\top \Lambda^K \mathbf{Z}_1|^\nu]}{\sigma^\nu n^{\nu/2}(n-1)^{\nu/2}} \;\le\; \frac{8C_2(M_{\mathrm{full};\nu})^\nu + (8C_2+4)\varepsilon_{K;\nu}^\nu}{\sigma^\nu n^{\nu/2}(n-1)^{\nu/2}}\,.$$

Substituting the five moment bounds into the earlier bounds on $\mathbb{E}[|\partial F_i(\mathbf{0})^\top \tilde{\mathbf{V}}_i|^\nu]$ and $\mathbb{E}[|\partial F_i(\mathbf{0})^\top \tilde{\mathbf{Z}}_i|^\nu]$ and combining the constant terms, we get that there exists an absolute

constant $C > 0$ such that

$$\mathbb{E}\big[|\partial F_i(\mathbf{0})^\top \tilde{\mathbf{V}}_i|^\nu\big] + \mathbb{E}\big[|\partial F_i(\mathbf{0})^\top \tilde{\mathbf{Z}}_i|^\nu\big] \;\le\; \frac{C}{n^{\nu/2}}\Big(\frac{(M_{\mathrm{full};\nu})^\nu + \varepsilon_{K;\nu}^\nu}{\sigma^\nu} + \frac{(M_{\mathrm{cond};\nu})^\nu + \varepsilon_{K;\nu}^\nu}{(n-1)^{-\nu/2}\,\sigma^\nu}\Big)\,.$$

□

A.3.5. Proof of Lemma A.16

Proof of Lemma A.16. For convenience, we write

$$U_0 \;:=\; \frac{1}{n(n-1)} \sum_{1\le i\ne j\le n} (\eta_i^K)^\top (\Sigma^K)^{1/2} \Lambda^K (\Sigma^K)^{1/2} \eta_j^K + \frac{2}{n}\sum_{i=1}^n (\mu^K)^\top \Lambda^K (\Sigma^K)^{1/2} \eta_i^K\,,$$

so that

$$\tilde{D}_Z^K \;=\; \frac{\sqrt{n(n-1)}}{\sigma} U_0 + \frac{\sqrt{n(n-1)}}{\sigma} (\mu^K)^\top \Lambda^K \mu^K\,,\quad \tilde{U}_n^K \;=\; \frac{\sqrt{n(n-1)}}{\sigma} U_0 + \frac{\sqrt{n(n-1)}}{\sigma} D\,.$$

To approximate the distribution of $\tilde{D}_Z^K$ by that of $\tilde{U}_n^K$, the proof boils down to replacing $(\mu^K)^\top \Lambda^K \mu^K$ by D. We use a Markov-type argument so that we obtain an error term that is separate from the distribution terms.

Recall that Lemma A.11 allows us to approximate the distribution of a sum of two random variables by a single one provided that the other is negligible. Writing

$$\tilde{D}_Z^K \;=\; \tilde{U}_n^K + (\tilde{D}_Z^K - \tilde{U}_n^K) \;=\; \tilde{U}_n^K + \frac{\sqrt{n(n-1)}}{\sigma}\big((\mu^K)^\top \Lambda^K \mu^K - D\big)\,,$$

we can apply Lemma A.11 to obtain that for any $a, b \in \mathbb{R}$ and $\epsilon > 0$,

$$\mathbb{P}(a \le \tilde{D}_Z^K \le b) \;\le \mathbb{P}\big(a - \epsilon \le \tilde{U}_n^K \le b + \epsilon\big) + \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma}\big|(\mu^K)^\top \Lambda^K \mu^K - D\big| \ge \epsilon\Big)\,,$$

$$\mathbb{P}(a \le \tilde{D}_Z^K \le b) \;\ge \mathbb{P}(a + \epsilon \le \tilde{U}_n^K \le b - \epsilon) - \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma}\big|(\mu^K)^\top \Lambda^K \mu^K - D\big| \ge \epsilon\Big)\,.$$

Note that $|(\mu^K)^\top \Lambda^K \mu^K - D|$ is deterministic. By a Markov inequality and the bound from Lemma A.7, we get that

$$\begin{aligned}\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma}\big|(\mu^K)^\top \Lambda^K \mu^K - D\big| \ge \epsilon\Big) \;&\le \frac{\sqrt{n(n-1)}}{\epsilon\sigma}\mathbb{E}\Big[\big|(\mu^K)^\top \Lambda^K \mu^K - D\big|\Big] \\ &= \frac{\big|\sum_{k=1}^K \lambda_K \mu_k^2 - D\big|}{\epsilon\, n^{-1/2}(n-1)^{-1/2}\sigma} \;\le\; \frac{\varepsilon_{K;1}}{\epsilon\, n^{-1/2}(n-1)^{-1/2}\sigma}\,.\end{aligned}$$

Combining the two results gives the desired bounds. □

A.3.6. Proof of Lemma A.17

Proof overview. The key ingredient of the proof is Theorem 8 of Carbery and Wright (2001), which gives an anti-concentration bound for the distribution of a polynomial of Gaussians in terms of its variance. In Lemma A.12, we have rewritten the result in the

special case of a degree-two polynomial, which allows us to control the distribution of $\tilde{U}_n^K$ in terms of its variance.

We introduce some matrix shorthands: For any $m \in \mathbb{N}$, denote O_m as the zero matrix in $\mathbb{R}^{m\times m}$, J_m as the all-one matrix in $\mathbb{R}^{m\times m}$ and I_m as the identity matrix in $\mathbb{R}^{m\times m}$. Define the $nK \times nK$ matrix M as

$$M \coloneqq \begin{pmatrix} O_K & \Lambda^K & \dots & \Lambda^K \\ \Lambda^K & O_K & \ddots & \vdots \\ \vdots & \ddots & \ddots & \Lambda^K \\ \Lambda^K & \dots & \Lambda^K & O_K \end{pmatrix} = \Lambda^K \otimes (J_n - I_n) \,,$$

as well as

$$\mu \coloneqq \big((\mu^K)^\top, \dots, (\mu^K)^\top\big)^\top \in \mathbb{R}^{nK} \,, \qquad \Sigma \coloneqq \Sigma^K \otimes I_n \in \mathbb{R}^{nK\times nK} \,,$$
$$\Lambda \coloneqq \Lambda^K \otimes I_n \in \mathbb{R}^{nK\times nK} \,.$$

We also consider the concatenated nK-dimensional standard Gaussian vector

$$\eta \coloneqq \big((\eta_1^K)^\top, \dots, (\eta_n^K)^\top\big)^\top \,.$$

Proof of Lemma A.17. The goal is to bound the distribution function between $a \le b \in \mathbb{R}$ of

$$\begin{aligned} \tilde{U}_n^K = \frac{\sqrt{n(n-1)}}{\sigma} U_n^K &= \frac{1}{\sigma\sqrt{n(n-1)}} \sum\nolimits_{1\le i\ne j\le n} (\eta_i^K)^\top (\Sigma^K)^{1/2} \Lambda^K (\Sigma^K)^{1/2} \eta_j^K \\ &\quad + \frac{2\sqrt{n-1}}{\sigma\sqrt{n}} \sum\nolimits_{i=1}^n (\mu^K)^\top \Lambda^K (\Sigma^K)^{1/2} \eta_i^K + \frac{\sqrt{n(n-1)}}{\sigma} D \\ &= \frac{1}{\sigma\sqrt{n(n-1)}} \eta^\top \Sigma^{1/2} M \Sigma^{1/2} \eta + \frac{2\sqrt{n-1}}{\sigma\sqrt{n}} \mu^\top \Lambda \Sigma^{1/2} \eta + \frac{\sqrt{n(n-1)}}{\sigma} D \,. \end{aligned}$$

For convenience, define

$$Q_1 \coloneqq \eta^\top \Sigma^{1/2} M \Sigma^{1/2} \eta \,, \quad Q_2 \coloneqq \mu^\top \Lambda \Sigma^{1/2} \eta \,, \quad \tilde{U}_0 \coloneqq \frac{1}{\sigma\sqrt{n(n-1)}} Q_1 + \frac{2\sqrt{n-1}}{\sigma\sqrt{n}} Q_2 \,.$$

Denote $\alpha \coloneqq \frac{b-a}{2}$ and $\beta \coloneqq \frac{a+b}{2}$. Rewriting the probability in terms of $\tilde{U}_0$, α and β, we get that

$$\begin{aligned} \mathbb{P}(a \le \tilde{U}_n^K \le b) &= \mathbb{P}\Big((\beta-\alpha) \le \tilde{U}_0 + \frac{\sqrt{n(n-1)}}{\sigma} D \le (\beta+\alpha)\Big) \\ &= \mathbb{P}\Big(\Big|\tilde{U}_0 + \frac{\sqrt{n(n-1)}}{\sigma} D - \beta\Big| \le \alpha\Big) \,. \end{aligned}$$

Since $\tilde{U}_0 + \frac{\sqrt{n(n-1)}}{\sigma} D - \beta$ is a degree-two polynomial of η, we can apply Lemma A.12

to bound the above probability: For an absolute constant C', we have

$$\mathbb{P}(a \le \tilde{U}_n^K \le b) \;\le C'\alpha^{1/2}\big(\mathrm{Var}[\tilde{U}_0]\big)^{-1/4}\,, \tag{A.11}$$

where the variance term can be expanded as

$$\mathrm{Var}\big[\tilde{U}_0\big] \;= \frac{1}{n(n-1)\sigma^2}\mathrm{Var}[Q_1] + \frac{4(n-1)}{n\sigma^2}\mathrm{Var}[Q_2] + \frac{4}{n\,\sigma^2}\,\mathrm{Cov}[Q_1, Q_2]\;.$$

We now provide bound the individual terms in the variance. By noting that each summand in Q_1 is zero-mean when $i \neq j$ and that each summand in Q_2 is zero-mean, the covariance term can be computed as

$$\begin{aligned}\mathrm{Cov}[Q_1, Q_2] \;&= \sum_{1\le i\neq j\le n}\sum_{l=1}^{n}\mathbb{E}\Big[(\eta_i^K)^\top(\Sigma^K)^{1/2}\Lambda^K(\Sigma^K)^{1/2}\eta_j^K \times (\mu^K)^\top\Lambda^K(\Sigma^K)^{1/2}\eta_l^K\Big]\\ &= \frac{1}{2}\mathbb{E}\Big[(\eta_1^K)^\top(\Sigma^K)^{1/2}\Lambda^K(\Sigma^K)^{1/2}\eta_1^K \times (\mu^K)^\top\Lambda^K(\Sigma^K)^{1/2}\eta_1^K\Big]\;.\end{aligned}$$

Denote ξ_k as the k-th coordinate of η_1^K. Then the above expectation is taken over a linear combination of terms of the form $\xi_{k_1}\xi_{k_2}\xi_{k_3}$. If any of k_1, k_2, k_3 is distinct from the other two indices, the expectation is zero; if $k_1 = k_2 = k_3$, the expectation is again zero by property of a standard Gauassian. Therefore, we have

$$\mathrm{Cov}[Q_1, Q_2] \;=\; 0\;.$$

On the other hand, the first variance can be computed by using the moment formula for a quadratic form of Gaussian from Lemma A.5 and the cyclic property of trace:

$$\begin{aligned}\mathrm{Var}[Q_1] \;&= 2\mathrm{Tr}\big((\Sigma^{1/2}M\Sigma^{1/2})^2\big) \;=\; 2\mathrm{Tr}\big((\Sigma M)^2\big)\\ &= 2\mathrm{Tr}\big((\Sigma^K\Lambda^K)^2 \otimes (J_n - I_n)^2\big)\\ &= 2\mathrm{Tr}\big((\Sigma^K\Lambda^K)^2 \otimes J_n^2\big) - 4\mathrm{Tr}\big((\Sigma^K\Lambda^K)^2 \otimes J_n\big) + 2\mathrm{Tr}\big((\Sigma^K\Lambda^K)^2 \otimes I_n\big)\\ &= \big(2n^2 - 4n + 2n\big)\mathrm{Tr}\big((\Sigma^K\Lambda^K)^2\big)\\ &= 2n(n-1)\mathrm{Tr}\big((\Lambda^K\Sigma^K)^2\big)\\ &\ge 2n(n-1)(\sigma_{\text{full}} - \varepsilon_{K;2})^2\;.\end{aligned}$$

In the last inequality, we have used the bound from Lemma A.8 on $\mathrm{Tr}\big((\Lambda^K\Sigma^K)^2\big)$. The second variance is on a Gaussian random variable and can be bounded by Lemma A.8 again as

$$\mathrm{Var}[Q_2] \;= \mu^\top\Lambda\Sigma\Lambda\mu \;=\; n(\mu^K)^\top\Lambda^K\Sigma^K\Lambda^K\mu^K \;\ge\; n(\sigma_{\text{cond}}^2 - 2\sigma_{\text{cond}}\varepsilon_{K;2} - 4\varepsilon_{K;2})\;.$$

This implies that

$$\mathrm{Var}\big[\tilde{U}_0\big] \;\ge \frac{2}{\sigma^2}(\sigma_{\text{full}} - \varepsilon_{K;2})^2 + \frac{4(n-1)}{\sigma^2}(\sigma_{\text{cond}}^2 - 2\sigma_{\text{cond}}\varepsilon_{K;2} - 4\varepsilon_{K;2})\;.$$

Substituting this into (A.11) and redefining the constants, we get that there exists an

absolute constant C such that

$$\mathbb{P}(a \le \tilde{U}_n^K \le b) \ \le C(b-a)^{1/2}\Big(\frac{1}{\sigma^2}(\sigma_{\text{full}} - \varepsilon_{K;2})^2 + \frac{n-1}{\sigma^2}(\sigma_{\text{cond}}^2 - 2\sigma_{\text{cond}}\varepsilon_{K;2} - 4\varepsilon_{K;2})\Big)^{-1/4}.$$

□

A.4 Proofs for the remaining results in Section 3.2

A.4.1. Proofs for variants and corollaries of the main result

The upper bound in Proposition 3.3 is a concentration inequality and is obtained by a standard argument via Chebyshev's inequality. The lower bound is a combination of the anti-concentration bound for a Gaussian quadratic form from Lemma A.17 and Theorem 3.1.

Proof of Proposition 3.3. Denote $\tilde{U}_n^K := \frac{\sqrt{n(n-1)}U_n^K}{\sigma_{\max}}$. In Lemma A.17, we have shown that for any $a, b \in \mathbb{R}$ with $a \le b$, there exists some absolute constant C' such that

$$\mathbb{P}(a \le \tilde{U}_n^K \le b) \le C'(b-a)^{1/2}\Big(\frac{1}{\sigma_{\max}^2}(\sigma_{\text{full}} - \varepsilon_{K;2})^2 + \frac{n-1}{\sigma_{\max}^2}(\sigma_{\text{cond}}^2 - 2\sigma_{\text{cond}}\varepsilon_{K;2} - 4\varepsilon_{K;2})\Big)^{-1/4}.$$

Take $K \to \infty$ and using Assumption 3.2 for $\nu \ge 2$, we get that $\varepsilon_{K;2} \to 0$. For a fixed $\epsilon > 0$, set $a = \frac{\sqrt{n(n-1)}}{\sigma_{\max}}D - \epsilon$ and $b = \frac{\sqrt{n(n-1)}}{\sigma_{\max}}D + \epsilon$, we get that

$$\begin{aligned}\lim_{K\to\infty} \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\max}}|U_n^K - D| \le \epsilon\Big) &\le \sqrt{2}\,C'\,\epsilon^{1/2}\Big(\frac{\sigma_{\text{full}}^2}{\sigma_{\max}^2} + \frac{(n-1)\sigma_{\text{cond}}^2}{\sigma_{\max}^2}\Big)^{-1/4} \\ &\le \sqrt{2}\,C'\,\epsilon^{1/2}.\end{aligned}$$

Now by Theorem 3.1, there exists an absolute constant C'' such that

$$\sup_{t\in\mathbb{R}}\Big|\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\max}}D_n > t\Big) - \lim_{K\to\infty}\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\max}}U_n^K > t\Big)\Big| \le C''\,n^{-\frac{\nu-2}{4\nu+2}}\Big(\frac{M_{\max;\nu}}{\sigma_{\max}}\Big)^{\frac{\nu}{2\nu+1}}.$$

By the triangle inequality, we get that

$$\begin{aligned}\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\max}}|D_n - D| > \epsilon\Big) &\ge \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\max}}|U_n^K - D| > \epsilon\Big) \\ &\quad - 2C''\,n^{-\frac{\nu-2}{4\nu+2}}\Big(\frac{M_{\max;\nu}}{\sigma_{\max}}\Big)^{\frac{\nu}{2\nu+1}} \\ &\ge 1 - \sqrt{2}\,C'\epsilon^{1/2} - 2C''\,n^{-\frac{\nu-2}{4\nu+2}}\Big(\frac{M_{\max;\nu}}{\sigma_{\max}}\Big)^{\frac{\nu}{2\nu+1}}.\end{aligned}$$

By replacing ϵ with $\frac{\sqrt{n(n-1)}}{\sigma_{\max}}\epsilon$ and redefining constants, we get the desired lower bound that there exists absolute constants $C_1, C_2 > 0$ such that

$$\mathbb{P}(|D_n - D| > \epsilon) \;\geq 1 - C_1\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\max}}\Big)^{1/2}\epsilon^{1/2} - C_2\, n^{-\frac{\nu-2}{4\nu+2}}\Big(\frac{M_{\max;\nu}}{\sigma_{\max}}\Big)^{\frac{\nu}{2\nu+1}} .$$

For the upper bound, we apply a Chebyshev inequality directly to D_n and bound the variance by Lemma A.6: There exists some absolute constant $C_3' > 0$ such that

$$\begin{aligned}\mathbb{P}(|D_n - D| > \epsilon) \;\leq\; \epsilon^{-2}\mathrm{Var}[D_n] \;&\leq C_3'\epsilon^{-2}\Big(\frac{\sigma_{\mathrm{cond}}^2}{n^{-1}(n-1)^2} + \frac{\sigma_{\mathrm{full}}^2}{(n-1)^2}\Big)\\ &\leq C_3'\epsilon^{-2}\Big(\frac{\sigma_{\max}}{n-1}\Big)^2 \;\leq\; C_3\epsilon^{-2}\Big(\frac{\sigma_{\max}}{\sqrt{n(n-1)}}\Big)^2 .\end{aligned}$$

In the last inequality, we have noted that $\frac{1}{n-1} \leq \frac{2}{n}$ for $n \geq 2$ and defined $C_3 = 2C_3'$. This finishes the proof. □

Theorem 3.1 provides an approximation of the distribution of D_n by that of a Gaussian quadratic form. Proposition 3.5 combines Theorem 3.1 with a Markov argument, which makes a further approximation of the Gaussian quadratic form by a weighted sum of chi-squares U_n^K. The approximation error introduced vanishes as n, d grow provided that $\rho_d = \omega(n^{1/2})$, i.e. $n^{-1/2}\sigma_{\mathrm{full}} = \omega(\sigma_{\mathrm{cond}})$.

Proof of Proposition 3.5. We first seek to compare W_n^K to the distribution of

$$\begin{aligned}U_n^K \;&= \frac{1}{n(n-1)}\sum\nolimits_{1\leq i\neq j\leq n}(\eta_i^K)^\top(\Sigma^K)^{1/2}\Lambda^K(\Sigma^K)^{1/2}\eta_j^K\\ &\quad + \frac{2}{n}\sum\nolimits_{i=1}^n(\mu^K)^\top\Lambda^K(\Sigma^K)^{1/2}\eta_i^K + D ,\end{aligned}$$

where $\{\eta_i^K\}_{i=1}^n$ are i.i.d. standard Gaussian vectors in $\mathbb{R}^K$. The first step is to write

$$U_n^K \;=\; \frac{\sqrt{n-1}}{\sqrt{n}}W_0 + D + \Big(1 - \frac{\sqrt{n-1}}{\sqrt{n}}\Big)W_0 + W_1 + W_2 ,$$

where we have defined the zero-mean random variables

$$\begin{aligned}W_0 \;&:= \frac{1}{n(n-1)}\Big(\sum\nolimits_{i,j=1}^n(\eta_i^K)^\top(\Sigma^K)^{1/2}\Lambda^K(\Sigma^K)^{1/2}\eta_j^K - n\mathrm{Tr}(\Sigma^K\Lambda^K)\Big) ,\\ W_1 \;&:= \frac{1}{n(n-1)}\Big(\sum\nolimits_{i=1}^n(\eta_i^K)^\top(\Sigma^K)^{1/2}\Lambda^K(\Sigma^K)^{1/2}\eta_i^K - n\mathrm{Tr}(\Sigma^K\Lambda^K)\Big) ,\\ W_2 \;&:= \frac{2}{n}\sum\nolimits_{i=1}^n(\mu^K)^\top\Lambda^K(\Sigma^K)^{1/2}\eta_i^K .\end{aligned}$$

Fix $\epsilon_0, \epsilon_1, \epsilon_2 > 0$. We first use the bound from Lemma A.11: For any $a, b \in \mathbb{R}$, we have

$$\begin{aligned}&\mathbb{P}\Big(a \leq \frac{\sqrt{n(n-1)}}{\sigma_{\mathrm{full}}}\Big(\frac{\sqrt{n-1}}{\sqrt{n}}W_0 + D\Big) \leq b\Big)\\ &\leq \mathbb{P}\Big(a - \epsilon_0 - \epsilon_1 - \epsilon_2 \leq \frac{\sqrt{n(n-1)}}{\sigma_{\mathrm{full}}}U_n^K \leq b + \epsilon_0 + \epsilon_1 + \epsilon_2\Big)\end{aligned}$$

$$+\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}\Big(1-\frac{\sqrt{n-1}}{\sqrt{n}}\Big)|W_0|\geq\epsilon_0\Big)+\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}|W_1|\geq\epsilon_1\Big)$$
$$+\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}|W_2|\geq\epsilon_2\Big)$$

and

$$\begin{aligned}&\mathbb{P}\Big(a\leq\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}\Big(\frac{\sqrt{n-1}}{\sqrt{n}}W_0+D\Big)\leq b\Big)\\ &\geq\mathbb{P}\Big(a+\epsilon_0+\epsilon_1+\epsilon_2\leq\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}U_n^K\leq b-\epsilon_0-\epsilon_1-\epsilon_2\Big)\\ &\quad-\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}\Big(1-\frac{\sqrt{n-1}}{\sqrt{n}}\Big)|W_0|\geq\epsilon_0\Big)-\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}|W_1|\geq\epsilon_1\Big)\\ &\quad-\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}|W_2|\geq\epsilon_2\Big)\,.\end{aligned}$$

We now bound the error terms. By the Chebyshev's inequality, the variance formula of a quadratic form of Gaussians from Lemma A.5 and the bound from Lemma A.8, we get that

$$\begin{aligned}\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}|W_1|\geq\epsilon_1\Big)\;\leq\;\epsilon_1^{-2}\text{Var}\Big[\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}W_1\Big]\;&=\frac{2}{\epsilon_1^2(n-1)\sigma_{\text{full}}^2}\text{Tr}\big((\Lambda^K\Sigma^K)^2\big)\\ &\leq\frac{2(\sigma_{\text{full}}+\varepsilon_{K;2})^2}{\epsilon_1^2(n-1)\sigma_{\text{full}}^2}\,.\end{aligned}$$

Similarly, by the Chebyshev's inequality, the variance formula of a Gaussian and the bound from Lemma A.8, we get that

$$\begin{aligned}\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}|W_2|\geq\epsilon\Big)\;\leq\;\epsilon_2^{-2}\text{Var}\Big[\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}W_2\Big]\;&=\frac{4(n-1)}{\epsilon_2^2\sigma_{\text{full}}^2}\mathbb{E}\big[(\mu^K)^\top\Lambda^K\Sigma^K\Lambda^K\mu^K\big]\\ &\leq\frac{4(n-1)(\sigma_{\text{cond}}+2\varepsilon_{K;2})^2}{\epsilon_2^2\sigma_{\text{full}}^2}\,.\end{aligned}$$

By Lemma A.9, we can replace W_0 by using the following equality in distribution:

$$\begin{aligned}\frac{\sqrt{n-1}}{\sqrt{n}}W_0\;&=\frac{1}{n^{3/2}(n-1)^{1/2}}\Big(\textstyle\sum_{i,j=1}^n(\eta_i^K)^\top(\Sigma^K)^{1/2}\Lambda^K(\Sigma^K)^{1/2}\eta_j^K-n\text{Tr}(\Sigma^K\Lambda^K)\Big)\\ &\stackrel{d}{=}W_n^K-D\,.\end{aligned}$$

Finally, using a Chebyshev's inequality together with the moment bound in Lemma A.9, we get that

$$\begin{aligned}\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}}\Big(1-\frac{\sqrt{n-1}}{\sqrt{n}}\Big)|W_0|\geq\epsilon_0\Big)\;&\leq\frac{n(n-1)}{\epsilon_0^2\sigma_{\text{full}}^2}\Big(1-\frac{\sqrt{n-1}}{\sqrt{n}}\Big)^2\text{Var}\big[W_0\big]\\ &=\frac{n^2}{\epsilon_0^2\sigma_{\text{full}}^2}\Big(1-\frac{\sqrt{n-1}}{\sqrt{n}}\Big)^2\text{Var}\big[W_n^K\big]\\ &\leq\frac{2n(\sigma_{\text{full}}+\varepsilon_{K;2})^2}{\epsilon_0^2(n-1)\sigma_{\text{full}}^2}\Big(1-\frac{\sqrt{n-1}}{\sqrt{n}}\Big)^2\\ &\leq\frac{2(\sigma_{\text{full}}+\varepsilon_{K;2})^2}{\epsilon_0^2(n-1)\sigma_{\text{full}}^2}\,.\end{aligned}$$

In the last inequality, we have noted that $\sqrt{n}-\sqrt{n-1}\leq 1$. Combining the above

bounds, we get that

$$\begin{aligned}
\mathbb{P}\Big(a \le \frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} W_n^K \le b\Big) &\le \mathbb{P}\Big(a - \epsilon_0 - \epsilon_1 - \epsilon_2 \le \frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} U_n^K \le b + \epsilon_0 + \epsilon_1 + \epsilon_2\Big) \\
&\quad + \frac{2(\sigma_{\text{full}} + \varepsilon_{K;2})^2}{(n-1)\sigma_{\text{full}}^2}\big(\epsilon_0^{-2} + \epsilon_1^{-2}\big) + \frac{4(n-1)(\sigma_{\text{cond}} + 2\varepsilon_{K;2})^2}{\epsilon_2^2 \sigma_{\text{full}}^2}\,, \\
\mathbb{P}\Big(a \le \frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} W_n^K \le b\Big) &\ge \mathbb{P}\Big(a + \epsilon_0 + \epsilon_1 + \epsilon_2 \le \frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} U_n^K \le b - \epsilon_0 - \epsilon_1 - \epsilon_2\big) \\
&\quad - \frac{2(\sigma_{\text{full}} + \varepsilon_{K;2})^2}{(n-1)\sigma_{\text{full}}^2}\big(\epsilon_0^{-2} + \epsilon_1^{-2}\big) - \frac{4(n-1)(\sigma_{\text{cond}} + 2\varepsilon_{K;2})^2}{\epsilon_2^2 \sigma_{\text{full}}^2}\,.
\end{aligned}$$

Taking $b \to \infty$ and $a \to t$ from the right, we get that

$$\begin{aligned}
&\Big|\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} W_n^K > t\Big) - \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} U_n^K > t\Big)\Big| \\
&\le \max\Big\{\mathbb{P}\Big(t - \epsilon_0 - \epsilon_1 - \epsilon_2 \le \frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} U_n^K \le t\Big)\,, \\
&\qquad \mathbb{P}\Big(t \le \frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} U_n^K \le \epsilon_0 + \epsilon_1 + \epsilon_2\Big)\Big\} \\
&\quad + \frac{2(\sigma_{\text{full}} + \varepsilon_{K;2})^2}{(n-1)\sigma_{\text{full}}^2}\big(\epsilon_0^{-2} + \epsilon_1^{-2}\big) + \frac{4(n-1)(\sigma_{\text{cond}} + 2\varepsilon_{K;2})^2}{\epsilon_2^2 \sigma_{\text{full}}^2}\,.
\end{aligned}$$

This allows us to follow a similar argument to the proof of Theorem 3.1 to approximate W_n^K by U_n^K. To bound the maxima, we apply Lemma A.17 with $\sigma = \sigma_{\text{full}}$: There exists some absolute constant C' such that for any $a \le b \in \mathbb{R}$,

$$\begin{aligned}
&\mathbb{P}\Big(a \le \frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} U_n^K \le b\Big) \\
&\quad \le C'(b-a)^{1/2}\Big(\frac{1}{\sigma_{\text{full}}^2}(\sigma_{\text{full}} - \varepsilon_{K;2})^2 + \frac{n-1}{\sigma_{\text{full}}^2}(\sigma_{\text{cond}}^2 - 2\sigma_{\text{cond}}\varepsilon_{K;2} - 4\varepsilon_{K;2})\Big)^{-1/4}\,.
\end{aligned}$$

By additionally noting that $(\epsilon_0 + \epsilon_1 + \epsilon_2)^{1/2} \le \sqrt{\epsilon_0} + \sqrt{\epsilon_1} + \sqrt{\epsilon_2}$, we get that

$$\begin{aligned}
&\Big|\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} W_n^K > t\Big) - \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} U_n^K > t\Big)\Big| \\
&\le C'(\sqrt{\epsilon_0} + \sqrt{\epsilon_1} + \sqrt{\epsilon_2})\Big(\frac{1}{\sigma_{\text{full}}^2}(\sigma_{\text{full}} - \varepsilon_{K;2})^2 \\
&\qquad\qquad + \frac{n-1}{\sigma_{\text{full}}^2}(\sigma_{\text{cond}}^2 - 2\sigma_{\text{cond}}\varepsilon_{K;2} - 4\varepsilon_{K;2})\Big)^{-1/4} \\
&\quad + \frac{2(\sigma_{\text{full}} + \varepsilon_{K;2})^2}{(n-1)\sigma_{\text{full}}^2}\big(\epsilon_0^{-2} + \epsilon_1^{-2}\big) + \frac{4(n-1)(\sigma_{\text{cond}} + 2\varepsilon_{K;2})^2}{\epsilon_2^2 \sigma_{\text{full}}^2}\,.
\end{aligned}$$

Taking $K \to \infty$ on both sides, the inequality becomes

$$\begin{aligned}
&\Big|\lim_{K\to\infty} \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} W_n^K > t\Big) - \lim_{K\to\infty} \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} U_n^K > t\Big)\Big| \\
&\quad \le C'(\sqrt{\epsilon_0} + \sqrt{\epsilon_1} + \sqrt{\epsilon_2})\Big(1 + \frac{(n-1)\sigma_{\text{cond}}^2}{\sigma_{\text{full}}^2}\Big)^{-1/4} + \frac{2}{n-1}\big(\epsilon_0^{-2} + \epsilon_1^{-2}\big) + \frac{4(n-1)\sigma_{\text{cond}}^2}{\epsilon_2^2 \sigma_{\text{full}}^2} \\
&\quad \le C'(\sqrt{\epsilon_0} + \sqrt{\epsilon_1} + \sqrt{\epsilon_2}) + \frac{2}{n-1}\big(\epsilon_0^{-2} + \epsilon_1^{-2}\big) + \frac{4(n-1)\sigma_{\text{cond}}^2}{\epsilon_2^2 \sigma_{\text{full}}^2}\,.
\end{aligned}$$

Choosing $\epsilon_0 = \epsilon_1 = (n-1)^{-2/5}$ and $\epsilon_2 = \big((n-1)\sigma_{\text{cond}}^2/\sigma_{\text{full}}^2\big)^{2/5}$, redefining constants

and taking a supremum over $t \in \mathbb{R}$, we get that there exists some absolute constant $C'' > 0$ such that

$$\sup_{t\in\mathbb{R}} \Big| \lim_{K\to\infty} \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} W_n^K > t\Big) - \lim_{K\to\infty} \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} U_n^K > t\Big)\Big| \\ \leq C''\Big(\frac{1}{(n-1)^{1/5}} + \Big(\frac{\sqrt{n-1}\,\sigma_{\text{cond}}}{\sigma_{\text{full}}}\Big)^{2/5}\Big) .$$

The final step is to relate this bound to D_n. Consider the last step (A.7) of the proof of Theorem 3.1 in Appendix A.3.2. If we set $\sigma = \sigma_{\text{full}}$ instead of $\sigma_{\max}$, we get that there exists some absolute constant $C''' > 0$ such that

$$\sup_{t\in\mathbb{R}} \Big|\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} D_n > t\Big) - \lim_{K\to\infty} \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} U_n^K > t\Big)\Big| \\ \leq C''' n^{-\frac{\nu-2}{4\nu+2}} \Big(\frac{(M_{\text{full};\nu})^\nu}{\sigma_{\text{full}}^\nu} + \frac{(M_{\text{cond};\nu})^\nu}{(n-1)^{-\nu/2}\,\sigma_{\text{full}}^\nu}\Big)^{\frac{1}{2\nu+1}} .$$

Setting $C = \max\{C'', C'''\}$ and using the triangle inequality, we get the desired bound that

$$\sup_{t\in\mathbb{R}} \Big|\mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} D_n > t\Big) - \lim_{K\to\infty} \mathbb{P}\Big(\frac{\sqrt{n(n-1)}}{\sigma_{\text{full}}} W_n^K > t\Big)\Big| \\ \leq C\Big(\frac{1}{(n-1)^{1/5}} + \Big(\frac{\sqrt{n-1}\,\sigma_{\text{cond}}}{\sigma_{\text{full}}}\Big)^{2/5} + n^{-\frac{\nu-2}{4\nu+2}} \Big(\frac{(M_{\text{full};\nu})^\nu}{\sigma_{\text{full}}^\nu} + \frac{(M_{\text{cond};\nu})^\nu}{(n-1)^{-\nu/2}\,\sigma_{\text{full}}^\nu}\Big)^{\frac{1}{2\nu+1}}\Big) .$$

□

A.4.2. Proofs for results on W_n

Proof of Proposition 3.6. To prove the existence of distribution, we seek to apply Lévy's continuity theorem. We first verify that there exists a sufficiently large K^* such that the sequence $(W_n^K)_{K\geq K^*}$ is tight. Since Assumption 3.2 holds for some $\nu \geq 2$, we get that as $K \to \infty$,

$$\varepsilon_{K;2} \coloneqq \mathbb{E}\big[\big|\textstyle\sum_{k=1}^K \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2) - u(\mathbf{X}_1,\mathbf{X}_2)\big|^2\big]^{1/2} \to 0 .$$

In particular, there exists some sufficiently large K^* such that $\varepsilon_{K;2} \leq 1$ for all $K \geq K^*$. By Lemma A.9, we have that for all $K \geq K^*$,

$$\operatorname{Var}[W_n^K] \leq \frac{2}{n(n-1)}(\sigma_{\text{full}} + \varepsilon_{K;2})^2 \leq \frac{2}{n(n-1)}(\sigma_{\text{full}} + 1)^2 .$$

Note that by assumption, we have $|D|, \sigma_{\text{full}} < \infty$. This implies that the sequence $(W_n^K)_{K\geq K^*}$ is tight by a Markov inequality:

$$\lim_{x\to\infty} \Big(\sup_{K\geq K^*} \mathbb{P}\big(\big|W_n^K\big| > x\big)\Big) \leq \lim_{x\to\infty} \Big(x^{-2} \sup_{K\geq K^*} \mathbb{E}[(W_n^K)^2]\Big) \\ \leq \lim_{x\to\infty} \frac{2n^{-1}(n-1)^{-1}(\sigma_{\text{full}}+1)^2 + D^2}{x^2} = 0 .$$

We defer to Lemma A.19 to show that the characteristic function of $(W_n^K - D)$ converges pointwise as $K \to \infty$. This allows us to apply Lévy's continuity theorem and obtain that W_n exists. □

Proof of Lemma 3.7. The result holds by noting that for all $k > K^*$, $W_n^K = W_n^{K^*}$ almost surely, and the latter random variable does not depend on K. □

Lemma A.19. *The characteristic function of* $(W_n^K - D)$ *converges pointwise as* $K \to \infty$.

Proof of Lemma A.19. Define $a_k \coloneqq \frac{1}{\sqrt{n(n-1)}}\tau_{k;d}$ and $T_k \coloneqq a_k(\xi_k^2 - 1)$, which allows us to write

$$W_n^K \;=\; \frac{1}{\sqrt{n(n-1)}} \sum_{k=1}^K \tau_{k;d}(\xi_k^2 - 1) + D \;=\; \sum_{k=1}^K T_k + D \;.$$

Denote $i = \sqrt{-1}$ as the imaginary unit and Y as a chi-squared random variable with degree 1. Since each T_k is a scaled and shifted chi-squared random variable with degree 1, it has the characteristic function

$$\psi_{T_k}(t) \;=\; \mathbb{E}[\exp(it\,T_k)] \;=\; \mathbb{E}[\exp(ia_kYt)]\exp(-ia_kt) \;= (1 - 2ia_kt)^{-1/2}\exp(-ia_kt) \;.$$

Since T_k's are independent, by the convolution theorem, the characteristic function of $W_n^K - D$ is given by

$$\psi_{W_n^K - D}(t) \;=\; \exp\Big(-i\sum_{k=1}^K a_kt\Big)\prod_{k=1}^K (1 - 2ia_kt)^{-1/2} \;.$$

We want to prove that for every $t \in \mathbb{R}$, $\psi_{W_n^K - D}(t)$ converges to some function as limit $K \to \infty$. By taking the principal-valued complex logarithm (i.e. discontinuity along negative real axis), we get that

$$\log\psi_{W_n^K - D}(t) \;=\; \sum_{k=1}^K \Big(-ia_kt - \frac{1}{2}\log(1 - 2ia_kt)\Big) + 2im_K\pi \;=:\; S_K + 2im_K\pi \;, \tag{A.12}$$

for some $m_K \in \mathbb{N}$ for each K that adjusts for values at discontinuity. Now consider the real part of the logarithm:

$$\begin{aligned}\mathrm{Re}\big(\log\psi_{W_n^K - D}(t)\big) \;&= \mathrm{Re}(S_K) \;=\; -\frac{1}{2}\sum_{k=1}^K \log|1 - 2ia_kt| \\ &= -\frac{1}{2}\sum_{k=1}^K \log\sqrt{1 + 4a_k^2t^2} \;=\; -\frac{1}{4}\sum_{k=1}^K \log(1 + 4a_k^2t^2) \;.\end{aligned}$$

Recall by Lemma A.8 that

$$\sum_{k=1}^K a_k^2 \;=\; \frac{1}{n(n-1)}\sum_{k=1}^K \tau_{k;d}^2 \;=\; \mathrm{Tr}((\Sigma^K\Lambda^K)^2) \xrightarrow{K\to\infty} \sigma_{\text{full}}^2 \;. \tag{A.13}$$

Fix $\epsilon > 0$. The above implies that there exists a sufficiently large K^* such that for all $K_1, K_2 \geq K^*$, $\sum_{k=K_1}^{K_2} a_k^2 < \epsilon$. Then for all $K_1, K_2 \geq K^*$, we have

$$0 \;\leq\; \sum_{k=K_1}^{K_2} \log(1 + 4a_k^2 t^2) \;\leq\; 4t^2 \sum_{k=K_1}^{K_2} a_k^2 \;\leq\; 4t^2\epsilon \,.$$

This implies that $(\mathrm{Re}(S_K))_{K\in\mathbb{N}}$ is the Cauchy sequence and therefore converges. Now we handle the imaginary part. First let $m'_K \in \mathbb{Z}$ be such that

$$\mathrm{Im}\Big(\sum_{k=1}^{K} \log(1 - 2ia_k t)\Big) \;=\; \sum_{k=1}^{K} \arctan(-2a_k t) + 2m'_K\pi \,.$$

Then we have

$$\mathrm{Im}(S_K) \;=\; \sum_{k=1}^{K} \Big(-a_k t + \frac{1}{2}\arctan(2a_k t)\Big) - m'_K\pi \;=:\; I_K - m'_K\pi \,. \tag{A.14}$$

To show that I_K converges, we first note that by a third-order Taylor expansion, we have that $\arctan(x) = x + \frac{6(x_*)^2-2}{6(x_*^2+1)^3}x^3$ for some $x_* \in [0, x]$ (we use this to denote $[0, x]$ for $x \geq 0$ as well as $[x, 0]$ for $x < 0$, with an abuse of notation). This implies that for all $K_1, K_2 \geq K^*$, where K^* is defined as before,

$$\begin{aligned}
\Big|\sum_{k=K_1}^{K_2} \Big(-a_k t + \frac{1}{2}\arctan(2a_k t)\Big)\Big| \;&=\; \Big|\sum_{k=K_1}^{K_2} \Big(-a_k t + \frac{1}{2}\arctan(2a_k t)\Big)\Big| \\
&\leq \sum_{k=K_1}^{K_2} \sup_{b_k\in[0,a_k]} \Big|\frac{1}{2}\frac{24b_k^2t^2 - 2}{6(4b_k^2t^2+1)^3} 8a_k^3t^3\Big| \\
&= 4t^3 \sum_{k=K_1}^{K_2} |a_k|^3 \Big(\sup_{b_k\in[0,a_k]} \Big|\frac{24b_k^2t^2 + 6 - 8}{6(4b_k^2t^2+1)^3}\Big|\Big) \\
&= 4t^3 \sum_{k=K_1}^{K_2} |a_k|^3 \Big(\sup_{b_k\in[0,a_k]} \Big|\frac{1}{(4b_k^2t^2+1)^2} - \frac{4}{3(4b_k^2t^2+1)^2}\Big|\Big) \\
&\leq 20t^3 \sum_{k=K_1}^{K_2} |a_k|^3 \;\leq\; 20t^3\Big(\sum_{k=K_1}^{K_2} (a_k)^2\Big)^{3/2} \;\leq\; 20t^3\epsilon^{3/2} \,,
\end{aligned}$$

where, in the last line, we have used the relative sizes of l_p norms. This implies that I_K converges. To show that Equation (A.14) converges, we need to show that m_K in Equation (A.14) is eventually constant. By using Equation (A.14) and the triangle inequality, we have that

$$\begin{aligned}
\pi|m'_{K+1} - m'_K| \;&\leq |I_{K+1} - I_K| + \Big|\mathrm{Im}(S_{K+1}) - \mathrm{Im}(S_K)\Big| \\
&= |I_{K+1} - I_K| + \Big|a_{K+1}t + \frac{1}{2}\log(1 - 2ia_{K+1}t)\Big| \,.
\end{aligned}$$

The first term converges to zero, since we have shown that I_K converges. Since $a_K \to 0$ by Equation (A.13) and the complex logarithm we use is continuous outside $\{z : \mathrm{Re}(z) > 0\}$, the second term above also converges to zero. Therefore $|m'_{K+1} - m'_K| \to 0$, and since $(m'_K)_{K\in\mathbb{N}}$ is an integer sequence, $(m'_K)_{K\in\mathbb{N}}$ converges. By Equation (A.14), this implies that $\mathrm{Im}(S_K)$ converges, and since we have shown $\mathrm{Re}(S_K)$ converges, we get that S_K converges. Finally, to show that $\psi_{W_n^K - D}(t)$ converges, since $\mathrm{Re}(S_K) = \mathrm{Re}\big(\psi_{W_n^K - D}(t)\big)$, we only need to show that $\mathrm{Im}\big(\psi_{W_n^K - D}(t)\big)$ converges. By Equation (A.12), this again

reduces to showing that m_K is eventually constant. As before, by the triangle inequality,

$$\begin{aligned} 2\pi|m_{K+1} - m_K| \;&\leq |\mathrm{Im}(S_{K+1}) - \mathrm{Im}(S_K)| \\ &\quad + \left|\mathrm{Im}(\log \psi_{W_n^{K+1}-D}(t)) - \mathrm{Im}(\log \psi_{W_n^{K}-D}(t))\right| \\ &= |\mathrm{Im}(S_{K+1}) - \mathrm{Im}(S_K)| + \left|a_{K+1}t + \frac{1}{2}\log(1 - 2ia_{K+1}t)\right| \;\xrightarrow{K\to\infty}\; 0\,, \end{aligned}$$

where the convergence of both terms has been shown earlier. This proves that the characteristic function $\psi_{W_n^K - D}(t)$ converges for every $t \in \mathbb{R}$. □

A.5 Proofs for results in Section 3.4

A.5.1. Proofs for the general results

Proof of Lemma 3.10. To prove the first result, note that since κ is a kernel, there exists a RKHS $\mathcal{H}$ and a map $\Phi : \mathbb{R}^d \to \mathcal{H}$ such that we can write

$$\begin{aligned} u^{\mathrm{MMD}}\big((\mathbf{x},\mathbf{y}),(\mathbf{x}',\mathbf{y}')\big) \;&= \langle\Phi(\mathbf{x}),\Phi(\mathbf{y}')\rangle_{\mathcal{H}} + \langle\Phi(\mathbf{y}),\Phi(\mathbf{y}')\rangle_{\mathcal{H}} - \langle\Phi(\mathbf{x}),\Phi(\mathbf{y}')\rangle_{\mathcal{H}} \\ &\quad - \langle\Phi(\mathbf{x}'),\Phi(\mathbf{y})\rangle_{\mathcal{H}} \\ &= \langle\Phi(\mathbf{x}) - \Phi(\mathbf{y}),\Phi(\mathbf{x}') - \Phi(\mathbf{y}')\rangle_{\mathcal{H}}\,. \end{aligned}$$

Defining $\Phi_*\big((\mathbf{x},\mathbf{y})\big) \coloneqq \Phi(\mathbf{x}) - \Phi(\mathbf{y})$ proves that u^{MMD} is a kernel. To prove the second result, note that by the definition of a weak Mercer representation, we have that almost surely

$$\left|\sum_{k=1}^K \lambda_k\phi_k(\mathbf{Z}_1)\phi_k(\mathbf{Z}_2) - u^{\mathrm{MMD}}(\mathbf{Z}_1,\mathbf{Z}_2)\right| \;\xrightarrow{K\to\infty}\; 0\,,$$

which in particular implies convergence in probability. The argument uses the Vitali convergence theorem. By Assumption 3.3, there exists some $\nu^* > \nu$ such that

$$\sup_{K\geq 1}\mathbb{E}\left[\left|\sum_{k=1}^K \lambda_k\phi_k(\mathbf{Z}_1)\phi_k(\mathbf{Z}_2)\right|^{\nu^*}\right] < \infty \quad \text{and} \quad \mathbb{E}[|u^{\mathrm{MMD}}(\mathbf{Z}_1,\mathbf{Z}_2)|^{\nu^*}] < \infty\,.$$

By the triangle inequality and the Jensen's inequality, we have

$$\begin{aligned} &\sup_{K\geq 1}\mathbb{E}\left[\left|\sum_{k=1}^K \lambda_k\phi_k(\mathbf{Z}_1)\phi_k(\mathbf{Z}_2) - u^{\mathrm{MMD}}(\mathbf{Z}_1,\mathbf{Z}_2)\right|^{\nu^*}\right] \\ \leq\; &\sup_{K\geq 1}\mathbb{E}\left[\left|\left|\sum_{k=1}^K \lambda_k\phi_k(\mathbf{Z}_1)\phi_k(\mathbf{Z}_2)\right| + \left|u^{\mathrm{MMD}}(\mathbf{Z}_1,\mathbf{Z}_2)\right|\right|^{\nu^*}\right] \\ \leq\; &2^{\nu^*-1}\sup_{K\geq 1}\mathbb{E}\left[\left|\sum_{k=1}^K \lambda_k\phi_k(\mathbf{Z}_1)\phi_k(\mathbf{Z}_2)\right|^{\nu^*}\right] + 2^{\nu^*-1}\mathbb{E}\left[\left|u^{\mathrm{MMD}}(\mathbf{Z}_1,\mathbf{Z}_2)\right|^{\nu^*}\right] \;<\; \infty\,. \end{aligned}$$

This implies for any $\nu \in (2,\nu^*)$, the sequence

$$\Big(\Big(\sum_{k=1}^K \lambda_k\phi_k(\mathbf{Z}_1)\phi_k(\mathbf{Z}_2) - u^{\mathrm{MMD}}(\mathbf{Z}_1,\mathbf{Z}_2)\Big)^{\nu}\Big)_{K\in\mathbb{N}}$$

is uniformly integrable, and therefore converges to zero in $L_1(\mathbb{R}^{2d}, P \otimes Q)$ by the Vitali convergence theorem. Since convergence in L_ν implies convergence in $L_{\min\{\nu,3\}}$, we get that Assumption 3.2 holds for $\min\{\nu, 3\}$. □

Before we prove the next result, recall that $\{\lambda_k\}_{k=1}^\infty$ and $\{\phi_k\}_{k=1}^\infty$ are defined as the weak Mercer representation for the kernel κ under Q, and we have assumed that ϕ_k's are differentiable. We have also defined the sequence of values $\{\alpha_k\}_{k=1}^\infty$ and the sequence of functions $\{\psi_k\}_{k=1}^\infty$ in (3.10) as

$$\alpha_{(k'-1)d+l} \coloneqq \lambda_{k'} \qquad \text{and} \qquad \psi_{(k'-1)d+l}(\mathbf{x}) \coloneqq (\partial_{x_l} \log p(\mathbf{x}))\phi_{k'}(\mathbf{x}) + \partial_{x_l}\phi_{k'}(\mathbf{x}) \,,$$

for $1 \le l \le d$ and $k' \in \mathbb{N}$. For convenience, we denote $\psi_{k';l} \coloneqq \psi_{(k'-1)d+l}$ in the proof below.

Proof of Lemma 3.11. Recall that $\psi_{k';l}(\mathbf{x}) \coloneqq (\partial_{x_l} \log p(\mathbf{x}))\phi_{k'}(\mathbf{x}) + \partial_{x_l}\phi_{k'}(\mathbf{x})$. Write $\tilde{\psi}_{k'}(\mathbf{x}) \coloneqq (\psi_{k';1}(\mathbf{x}), \ldots, \psi_{k';n}(\mathbf{x}))^\top$. We first consider the error term with dK' summands for some $K' \in \mathbb{N}$:

$$\begin{aligned}
&\mathbb{E}\Big[\Big|\textstyle\sum_{k=1}^{dK'} \alpha_k \psi_k(\mathbf{X}_1)\psi_k(\mathbf{X}_2) - u_P^{\mathrm{KSD}}(\mathbf{X}_1, \mathbf{X}_2)\Big|^\nu\Big] \\
&= \mathbb{E}\Big[\Big|\textstyle\sum_{l=1}^{d}\sum_{k'=1}^{K'} \lambda_{k'} \psi_{k';l}(\mathbf{X}_1)\psi_{k';l}(\mathbf{X}_2) - u_P^{\mathrm{KSD}}(\mathbf{X}_1, \mathbf{X}_2)\Big|^\nu\Big] \\
&= \mathbb{E}\Big[\Big|\textstyle\sum_{k'=1}^{K'} \lambda_{k'} \big(\tilde{\psi}_{k'}(\mathbf{X}_1)\big)^\top \big(\tilde{\psi}_{k'}(\mathbf{X}_2)\big) - u_P^{\mathrm{KSD}}(\mathbf{X}_1, \mathbf{X}_2)\Big|^\nu\Big] \\
&= \mathbb{E}\big[\big|T_1 + T_2 + T_3 + T_4 - u_P^{\mathrm{KSD}}(\mathbf{X}_1, \mathbf{X}_2)\big|^\nu\big] \,,
\end{aligned}$$

where the random quantities are defined in terms of $\mathbf{X}_1, \mathbf{X}_2 \overset{i.i.d.}{\sim} Q$:

$$\begin{aligned}
T_1 &\coloneqq \big(\nabla \log p(\mathbf{X}_1)\big)^\top \big(\nabla \log p(\mathbf{X}_2)\big) \textstyle\sum_{k'=1}^{K'} \lambda_{k'} \phi_{k'}(\mathbf{X}_1)\phi_{k'}(\mathbf{X}_2) \,, \\
T_2 &\coloneqq \big(\nabla \log p(\mathbf{X}_1)\big)^\top \Big(\textstyle\sum_{k'=1}^{K'} \lambda_{k'} \big(\nabla \phi_{k'}(\mathbf{X}_2)\big)\phi_{k'}(\mathbf{X}_1)\Big) \,, \\
T_3 &\coloneqq \big(\nabla \log p(\mathbf{X}_2)\big)^\top \Big(\textstyle\sum_{k'=1}^{K'} \lambda_{k'} \big(\nabla \phi_{k'}(\mathbf{X}_1)\big)\phi_{k'}(\mathbf{X}_2)\Big) \,, \\
T_4 &\coloneqq \textstyle\sum_{k'=1}^{K'} \lambda_{k'} \big(\nabla \phi_{k'}(\mathbf{X}_1)\big)^\top \big(\nabla \phi_{k'}(\mathbf{X}_2)\big) \,.
\end{aligned}$$

Recall that by Assumption 3.3, there exists some $\nu^* > \nu$ such theta

$$\sup_{K \ge 1} \mathbb{E}\Big[\Big|\textstyle\sum_{k=1}^{K} \lambda_k \phi_k(\mathbf{Z}_1)\phi_k(\mathbf{Z}_2)\Big|^{\nu^*}\Big] < \infty \qquad \text{and} \qquad \mathbb{E}[|u^{\mathrm{MMD}}(\mathbf{Z}_1, \mathbf{Z}_2)|^{\nu^*}] < \infty \,.$$

By using the proof of the second part of Lemma 3.10 above, for $\nu^\Delta \coloneqq \frac{\nu+\nu^*}{2} \in (\nu, \nu^*)$, we have

$$\mathbb{E}\Big[\big|\textstyle\sum_{k'=1}^{K'} \lambda_{k'} \phi_{k'}(\mathbf{X}_1)\phi_{k'}(\mathbf{X}_2) - u(\mathbf{X}_1, \mathbf{X}_2)\big|^{\nu^\Delta}\Big] \xrightarrow{K' \to \infty} 0 \,.$$

Meanwhile by Assumption 3.4, $\big\|\|\nabla \log p(\mathbf{X}_1)\|_2\big\|_{L_{2\nu^{**}}} < \infty$, where

$$\nu^{**} \;=\; \frac{\nu(\nu+\nu^*)}{\nu^*-\nu} \;=\; \Big(\frac{1}{\nu}-\frac{2}{\nu+\nu^*}\Big)^{-1} \;=\; \Big(\frac{1}{\nu}-\frac{1}{\nu^{\Delta}}\Big)^{-1} \;>\; \nu\,.$$

By the Cauchy-Schwarz inequality and the Hölder's inequality, we have that

$$\big\|\big(\nabla \log p(\mathbf{X}_1)\big)^{\top}\big(\nabla \log p(\mathbf{X}_2)\big)\big\|_{L_{\nu^{**}}} \;\le\; \big\|\,\|\nabla \log p(\mathbf{X}_1)\|_2\,\big\|_{L_{2\nu^{**}}} < \infty\,.$$

Now by the Hölder's inequality and noting that $(\nu^{**})^{-1}+(\nu^{\Delta})^{-1}=\nu^{-1}$, we can now bound the error of using T_1 to approximate the first term of u_P^{KSD} as

$$\begin{aligned}
&\mathbb{E}[|E_1|^{\nu}] \;\coloneqq\; \mathbb{E}\big[\big|T_1-\big(\nabla \log p(\mathbf{X}_1)\big)^{\top}\big(\nabla \log p(\mathbf{X}_2)\big)\,u(\mathbf{X}_1,\mathbf{X}_2)\big|^{\nu}\big]\\
&=\; \big\|T_1-\big(\nabla \log p(\mathbf{X}_1)\big)^{\top}\big(\nabla \log p(\mathbf{X}_2)\big)\,u(\mathbf{X}_1,\mathbf{X}_2)\big\|_{L_\nu}^{\nu}\\
&\le\; \big\|\big(\nabla \log p(\mathbf{X}_1)\big)^{\top}\big(\nabla \log p(\mathbf{X}_2)\big)\big\|_{L_{\nu^{**}}}^{\nu}\Big\|\sum_{k'=1}^{K'}\lambda_{k'}\phi_{k'}(\mathbf{X}_1)\phi_{k'}(\mathbf{X}_2)-u(\mathbf{X}_1,\mathbf{X}_2)\Big\|_{L_{\nu^{\Delta}}}^{\nu}\\
&\xrightarrow{K'\to\infty}\; 0\,.
\end{aligned}$$

For T_2, we consider a similar approximation error quantity and apply the Cauchy-Schwarz inequality:

$$\begin{aligned}
&\mathbb{E}[\,|E_2|^{\nu}] \;\coloneqq\; \mathbb{E}\big[\big|T_2-\big(\nabla \log p(\mathbf{X}_1)\big)^{\top}\nabla_2\kappa(\mathbf{X}_1,\mathbf{X}_2)\big|^{\nu}\big]\\
&\quad=\; \mathbb{E}\Big[\Big|\big(\nabla \log p(\mathbf{X}_1)\big)^{\top}\Big(\textstyle\sum_{k'=1}^{K'}\lambda_{k'}\big(\nabla\phi_{k'}(\mathbf{X}_2)\big)\phi_{k'}(\mathbf{X}_1)-\nabla_2\kappa(\mathbf{X}_1,\mathbf{X}_2)\Big)\Big|^{\nu}\Big]\\
&\quad\le\; \|\,\|\nabla \log p(\mathbf{X}_1)\|_2\,\|_{L_{2\nu}}^{\nu}\Big\|\Big\|\textstyle\sum_{k'=1}^{K'}\lambda_{k'}\big(\nabla\phi_{k'}(\mathbf{X}_2)\big)\phi_{k'}(\mathbf{X}_1)-\nabla_2\kappa(\mathbf{X}_1,\mathbf{X}_2)\Big\|_2\Big\|_{L_{2\nu}}^{\nu}\\
&\quad\xrightarrow{K'\to\infty}\; 0\,,
\end{aligned}$$

where we have noted that the first term is bounded since $2\nu<2\nu^{**}$ and used Assumption 3.4(iv). By symmetry of κ and the fact that $\mathbf{X}_1$ and $\mathbf{X}_2$ are exchangeable, we have the same result for T_3:

$$\mathbb{E}[|E_3|^{\nu}] \;\coloneqq \mathbb{E}\big[\big|T_3-\big(\nabla \log p(\mathbf{X}_2)\big)^{\top}\nabla_1\kappa(\mathbf{X}_1,\mathbf{X}_2)\big|^{\nu}\big] \;\xrightarrow{K'\to\infty}\; 0\,.$$

Meanwhile, the second condition of Assumption 3.4(iv) directly says that

$$\mathbb{E}[|E_4|^{\nu}] \;\coloneqq \mathbb{E}\big[\big|T_4-\mathrm{Tr}\big(\nabla_1\nabla_2\kappa(\mathbf{X}_1,\mathbf{X}_2)\big)\big|^{\nu}\big] \;\xrightarrow{K'\to\infty}\; 0\,.$$

Combining the results and applying the Jensen's inequality to the convex function $x\mapsto |x|^{\nu}$, we have

$$\begin{aligned}
\mathbb{E}\Big[\Big|\textstyle\sum_{k=1}^{dK'}\alpha_k\psi_k(\mathbf{X}_1)\psi_k(\mathbf{X}_2)-u_P^{\mathrm{KSD}}(\mathbf{X}_1,\mathbf{X}_2)\Big|^{\nu}\Big] \;&=\; \mathbb{E}\big[\big|E_1+E_2+E_3+E_4\big|^{\nu}\big]\\
&\le \mathbb{E}\big[\big|\tfrac{1}{4}(4E_1)+\tfrac{1}{4}(4E_2)+\tfrac{1}{4}(4E_3)+\tfrac{1}{4}(4E_4)\big|^{\nu}\big]\\
&\le 4^{\nu-1}\big(\mathbb{E}[|E_1|^{\nu}]+\mathbb{E}[|E_2|^{\nu}]+\mathbb{E}[|E_3|^{\nu}]+\mathbb{E}[|E_4|^{\nu}]\big) \;\xrightarrow{K'\to\infty}\; 0\,.
\end{aligned}$$

Now consider $K\in\mathbb{N}$ that is not necessarily divisible by d, and let K' be the greatest

integer such that $K \geq dK'$. Then by the triangle inequality and the Jensen's inequality similarly as above, we get

$$\begin{aligned}
&\mathbb{E}\Big[\Big|\sum_{k=1}^{K} \alpha_k \psi_k(\mathbf{X}_1)\psi_k(\mathbf{X}_2) - u_P^{\mathrm{KSD}}(\mathbf{X}_1, \mathbf{X}_2)\Big|^{\nu}\Big] \\
&\qquad \leq 2^{\nu-1}\mathbb{E}\Big[\Big|\sum_{k=1}^{dK'} \alpha_k \psi_k(\mathbf{X}_1)\psi_k(\mathbf{X}_2) - u_P^{\mathrm{KSD}}(\mathbf{X}_1, \mathbf{X}_2)\Big|^{\nu}\Big] \\
&\qquad\quad + 2^{\nu-1}\mathbb{E}\Big[\Big|\sum_{k=dK'+1}^{K} \alpha_k \psi_k(\mathbf{X}_1)\psi_k(\mathbf{X}_2)\Big|^{\nu}\Big] \,. \qquad \text{(A.15)}
\end{aligned}$$

The first term is $o(1)$ as $K \to \infty$ by the previous argument, so we only need to focus on the second term. The expectation can be bounded by noting that $\alpha_k = \lambda_{K'+1} \geq 0$ for all $dK' + 1 \leq k \leq K$ and using the triangle inequality followed by the Jensen's inequality:

$$\begin{aligned}
&\mathbb{E}\Big[\Big|\sum_{k=dK'+1}^{K} \alpha_k \psi_k(\mathbf{X}_1)\psi_k(\mathbf{X}_2)\Big|^{\nu}\Big] \\
&\leq (\lambda_{K'+1})^{\nu}\mathbb{E}\Big[\Big(\frac{1}{K-dK'}\sum_{k=dK'+1}^{K}(K-dK')|\psi_k(\mathbf{X}_1)\psi_k(\mathbf{X}_2)|\Big)^{\nu}\Big] \\
&\leq (\lambda_{K'+1})^{\nu}(K-dK')^{\nu-1}\sum_{k=dK'+1}^{K}\mathbb{E}[|\psi_k(\mathbf{X}_1)\psi_k(\mathbf{X}_2)|^{\nu}] \\
&\leq (\lambda_{K'+1})^{\nu} d^{\nu} \sup_{k\in\{dK'+1,\ldots,dK'+d\}}\mathbb{E}[|\psi_k(\mathbf{X}_1)\psi_k(\mathbf{X}_2)|^{\nu}] \\
&= (\lambda_{K'+1})^{\nu} d^{\nu} \sup_{1\leq l\leq d}\mathbb{E}[|\psi_{dK'+l}(\mathbf{X}_1)|^{\nu}]^2 \,.
\end{aligned}$$

In the last equality, we have noted that $\mathbf{X}_1$ and $\mathbf{X}_2$ are identically distributed. Now by the definition of ψ_k, the Jensen's inequality on $x \mapsto |x|^{\nu}$ and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\mathbb{E}[|\psi_{dK'+l}(\mathbf{X}_1)|^{\nu}] \;&=\; \mathbb{E}[|(\partial_{x_l}\log p(\mathbf{X}_1))\phi_{K'+1}(\mathbf{X}_1) + \partial_{x_l}\phi_{K'+1}(\mathbf{X}_1)|^{\nu}] \\
&\leq 2^{\nu-1}\mathbb{E}[|(\partial_{x_l}\log p(\mathbf{X}_1))\phi_{K'+1}(\mathbf{X}_1)|^{\nu}] + 2^{\nu-1}\mathbb{E}[|\partial_{x_l}\phi_{K'+1}(\mathbf{X}_1)|^{\nu}] \\
&\leq 2^{\nu-1}\mathbb{E}[|\partial_{x_l}\log p(\mathbf{X}_1)|^{2\nu}]^{1/2}\,\mathbb{E}[|\phi_{K'+1}(\mathbf{X}_1)|^{2\nu}]^{1/2} + 2^{\nu-1}\mathbb{E}[|\partial_{x_l}\phi_{K'+1}(\mathbf{X}_1)|^{\nu}] \\
&\leq 2^{\nu-1}\mathbb{E}[\|\nabla\log p(\mathbf{X}_1)\|_2^{2\nu}]^{1/2}\,\mathbb{E}[|\phi_{K'+1}(\mathbf{X}_1)|^{2\nu}]^{1/2} + 2^{\nu-1}\mathbb{E}[\|\nabla\phi_{K'+1}(\mathbf{X}_1)\|_2^{\nu}] \\
&= 2^{\nu-1}\|\|\nabla\log p(\mathbf{X}_1)\|_2\|_{L_{2\nu}}^{\nu}\,\|\phi_{K'+1}(\mathbf{X}_1)\|_{L_{2\nu}}^{\nu} + 2^{\nu-1}\|\|\nabla\phi_{K'+1}(\mathbf{X}_1)\|_2\|_{L_{\nu}}^{\nu} \,.
\end{aligned}$$

By Assumption 3.4(i), (ii) and (iii), all three norms are bounded, so $\mathbb{E}[|\psi_{dK'+l}(\mathbf{X}_1)|^{\nu}] < \infty$. By the definition of λ_k from the weak Mercer representation, as $K \to \infty$ and therefore $K' \to \infty$, $\lambda_{K'+1} \to 0$, which implies

$$\mathbb{E}\Big[\Big|\sum_{k=dK'+1}^{K} \alpha_k \psi_k(\mathbf{X}_1)\psi_k(\mathbf{X}_2)\Big|^{\nu}\Big] \;=\; o(1) \,.$$

This means that both terms in (A.15) converge to 0 as $K \to \infty$. In other words,

$$\mathbb{E}\Big[\Big|\sum_{k=1}^{K} \alpha_k \psi_k(\mathbf{X}_1)\psi_k(\mathbf{X}_2) - u_P^{\mathrm{KSD}}(\mathbf{X}_1, \mathbf{X}_2)\Big|^{\nu}\Big] \;\xrightarrow{K\to\infty}\; 0 \,.$$

Since L_{ν}-convergence implies $L_{\min\{\nu,3\}}$-convergence and we have assumed that $\nu > 2$, we get that Assumption 3.2 holds for $\min\{\nu, 3\}$ with respect to the u_P^{KSD}, α_k and ψ_k. $\square$

A.6 Proofs for Appendix A.1

A.6.1. Proofs for RBF decomposition and verifying Assumption 3.2

In this section, we prove Lemma A.1, Lemma A.2 and Lemma A.3.

Proof of Lemma A.1. We first focus on the one-dimensional RBF kernel, denoted as κ_1, which can be expressed for $x, x' \in \mathbb{R}$ as

$$|\kappa_1(x,x')| \;=\; \left|\exp(-(x-x')^2/(2\gamma))\right| \;=\; \left|\exp\left(\frac{xx'}{\gamma}\right)e^{-x^2/(2\gamma)}e^{-(x')^2/(2\gamma)}\right| .$$

By applying a Taylor expansion around 0 to the infinitely differentiable function $z \mapsto \exp(\frac{z}{\gamma})$ for $z \in \mathbb{R}$, we obtain that for any $K \in \mathbb{N}$ and every $x, x' \in \mathbb{R}$.

$$\begin{aligned}&\left|\kappa_1(x,x') - \sum\nolimits_{k=0}^{K} \frac{1}{k!}\left(\frac{xx'}{\gamma}\right)^k e^{-x^2/(2\gamma)}e^{-(x')^2/(2\gamma)}\right| \\ &\qquad\leq \sup\nolimits_{z\in[0,xx']}\left|\frac{1}{(K+1)!}\left(\frac{xx'}{\gamma}\right)^{K+1}e^{z/\gamma}\right| e^{-x^2/(2\gamma)}e^{-(x')^2/(2\gamma)} .\end{aligned}$$

Fix $\nu \in (2,4]$. Consider two independent normal random variables $U \sim \mathcal{N}(b_1,1)$ and $V \sim \mathcal{N}(b_2,1)$ for some $b_1, b_2 \in \mathbb{R}$, and recall that $\phi_k^*(x) := x^k e^{-x^2/(2\gamma)}$ and $\lambda_k^* := \frac{1}{k!\,\gamma^k}$. The above then implies that

$$\begin{aligned}&\mathbb{E}\Big[\Big|\kappa_1(U,V) - \sum\nolimits_{k=0}^{K}\lambda_k^*\phi_k^*(U)\phi_k^*(V)\Big|^{\nu}\Big] \\ &\qquad\leq \mathbb{E}\Big[\sup\nolimits_{z\in[0,UV]}\left|\frac{1}{(K+1)!}\left(\frac{UV}{\gamma}\right)^{K+1}e^{z/\gamma}\right|^{\nu} e^{-\nu U^2/(2\gamma)}e^{-\nu V^2/(2\gamma)}\Big] \\ &\qquad= \frac{1}{((K+1)!\,\gamma^{K+1})^{\nu}}\,\mathbb{E}\big[|UV|^{\nu(K+1)}e^{-\nu U^2/(2\gamma)-\nu V^2/(2\gamma)+\sup_{z\in[0,UV]}\nu z/\gamma}\big] \\ &\qquad\leq \frac{1}{((K+1)!\,\gamma^{K+1})^{\nu}}\,\mathbb{E}\big[|UV|^{\nu(K+1)}e^{-\nu(|U|-|V|)^2/(2\gamma)}\big] \\ &\qquad\leq \frac{1}{((K+1)!\,\gamma^{K+1})^{\nu}}\,\mathbb{E}\big[|U|^{\nu(K+1)}\big]\,\mathbb{E}\big[|V|^{\nu(K+1)}\big] .\end{aligned}$$

In the last inequality, we have noted that U and V are independent and bounded the exponential term from above by 1. By the formula of absolute moments of a Gaussian, we get that

$$\mathbb{E}\big[|U-b_1|^{\nu(K+1)}\big] \;=\; \mathbb{E}\big[|V-b_2|^{\nu(K+1)}\big] \;=\; \frac{2^{(\nu K)/2}}{\sqrt{\pi}}\Gamma\Big(\frac{\nu K+1}{2}\Big) .$$

By the Jensen's inequality applied to the convex function $x \mapsto |x|^{\nu(K+1)}$, we get that

$$\begin{aligned}&\mathbb{E}\big[|U|^{\nu(K+1)}\big] \;=\; \mathbb{E}\big[|b_1+(U-b_1)|^{\nu(K+1)}\big] \;=\; \mathbb{E}\Big[\Big|\frac{1}{2}(2b_1)+\frac{1}{2}(2(U-b_1))\Big|^{\nu(K+1)}\Big] \\ &\leq 2^{\nu(K+1)-1}\big(b^{\nu(K+1)}+\mathbb{E}[|U-b_1|^{\nu(K+1)}]\big) \;=\; \frac{(2b_1)^{\nu(K+1)}}{2}+\frac{2^{\frac{3}{2}\nu(K+1)}}{2\sqrt{\pi}}\Gamma\Big(\frac{\nu(K+1)+1}{2}\Big) .\end{aligned}$$

Similarly, we get that

$$\mathbb{E}\big[|V|^{\nu(K+1)}\big] \;\leq\; \frac{(2b_2)^{\nu(K+1)}}{2}+\frac{2^{\frac{3}{2}\nu(K+1)}}{2\sqrt{\pi}}\Gamma\Big(\frac{\nu(K+1)+1}{2}\Big) . \tag{A.16}$$

Substituting these moment bounds and noting that $(K+1)! = \Gamma(K+2)$, we get that

$$\begin{aligned}
\mathbb{E}\Big[\Big|\kappa_1(U,V) - \textstyle\sum_{k=0}^{K} \lambda_k^* \phi_k^*(U)\phi_k^*(V)\Big|^{\nu}\Big] \\
&\leq \frac{1}{\gamma^{\nu(K+1)}\big(\Gamma(K+2)\big)^{\nu}}\Big(\frac{(2b_1)^{\nu(K+1)}}{2} + \frac{2^{\frac{3}{2}\nu(K+1)}}{2\sqrt{\pi}}\Gamma\Big(\frac{\nu(K+1)+1}{2}\Big)\Big) \\
&\quad \times \Big(\frac{(2b_2)^{\nu(K+1)}}{2} + \frac{2^{\frac{3}{2}\nu(K+1)}}{2\sqrt{\pi}}\Gamma\Big(\frac{\nu(K+1)+1}{2}\Big)\Big) \\
&=: T\,(A_1+B)(A_2+B)\;.
\end{aligned}$$

As K grows, the dominating terms are the Gamma functions, so we only need to control their ratios. By Stirling's formula for the gamma function, we have

$$\Gamma(x) \;=\; \sqrt{2\pi}\,x^{x-1/2}e^{-x}\big(1+O(x^{-1})\big)$$

for $x > 0$. This immediately implies that

$$TA_1A_2 \;=\; \Theta\Big(\frac{(4b_1b_2/\gamma)^{\nu(K+1)}}{(K+2)^{\nu(K+3/2)}e^{-\nu(K+2)}}\Big) \;=\; o(1)$$

as $K \to \infty$. Meanwhile,

$$\frac{\Gamma\big(\frac{\nu(K+1)+1}{2}\big)}{\big(\Gamma(K+2)\big)^{\nu}} \;= \Theta\Big(\frac{K^{\nu K/2}}{K^{\nu K}}\Big) \;=\; \Theta\Big(K^{-\nu K/2}\Big)\;,$$

which implies that

$$TA_1B \;=\; \Theta\Big((4\sqrt{2}b_1/\gamma)^{\nu K}K^{-\nu K/2}\Big) \;=\; o(1)\;,$$

since the dominating term is $K^{-\nu K/2}$. Similarly, $TA_2B \;=\; o(1)$. On the other hand, another application of Stirling's formula gives that

$$\begin{aligned}
\frac{\big(\Gamma\big(\frac{\nu(K+1)+1}{2}\big)\big)^2}{\big(\Gamma(K+2)\big)^{\nu}} \;&= (2\pi)^{-(\nu-2)/2}\,\frac{\big(\frac{\nu(K+1)+1}{2}\big)^{\nu(K+1)}}{(K+2)^{\nu(K+3/2)}}\,e^{-\nu(K+1)-1+\nu(K+2)}\,\frac{\big(1+O(K^{-1})\big)^2}{\big(1+O(K^{-1})\big)^{\nu}} \\
&= \Theta\Big(\frac{(\nu/2)^{\nu K}K^{\nu K}}{K^{\nu(K+3/2)}}\Big) \;=\; \Theta\big((\nu/2)^{\nu K}K^{-3\nu/2}\big)\;.
\end{aligned}$$

This implies that

$$TB^2 \;=\; \Theta\big((8/\gamma)^{\nu K}(\nu/2)^{\nu K}K^{-3\nu/2}\big) \;=\; \Theta\big((2\nu/\gamma)^{\nu K}K^{-3\nu/2}\big) \;=\; o(1)\;,$$

where we have recalled that $\nu \leq 4$ and used the assumption that $\gamma > 8$. In summary, we have proved that for $\nu \in (2,4]$ and *any* fixed $b_1, b_2 \in \mathbb{R}$,

$$\mathbb{E}\Big[\Big|\kappa_1(U,V) - \textstyle\sum_{k=0}^{K} \lambda_k^* \phi_k^*(U)\phi_k^*(V)\Big|^{\nu}\Big] \;\leq\; T(A_1+B)(A_2+B) \;\xrightarrow{K\to\infty}\; 0\;.$$

To extend this to multiple dimensions, we note that for the vectors $\mathbf{x} = (x_1,\ldots,x_d) \in \mathbb{R}^d$ and $\mathbf{x}' = (x_1,\ldots,x_d) \in \mathbb{R}^d$, the multi-dimensional RBF kernel can then be expressed as

$$\begin{aligned}
\kappa(\mathbf{x},\mathbf{x}') \;=\; \exp\big(-\|\mathbf{x}-\mathbf{x}'\|_2^2/(2\gamma)\big) \;&=\; \textstyle\prod_{l=1}^{d} \exp\big(-(x_l-x_l')^2/(2\gamma)\big) \\
&= \textstyle\prod_{l=1}^{d} \kappa_1(x_l,x_l')\;.
\end{aligned}$$

Recall that we have defined the independent normal vectors $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, I_d)$ and $\mathbf{V} \sim \mathcal{N}(\mu, I_d)$. Let $U_1, \ldots, U_d$ be the coordinates of $\mathbf{U}$ and $V_1, \ldots, V_d$ be those of $\mathbf{V}$, which are all independent since the covariance matrices are I_d. For $0 \le l \le d$ and $K \in \mathbb{N}$, define the random quantities

$$S_{j;K} := \sum_{k=0}^{K} \lambda_k^* \phi_k^*(U_j)\phi_k^*(V_j) \quad \text{and} \quad W_{l;K} := \Big(\prod_{j=1}^{l} \kappa_1(U_j, V_j)\Big)\Big(\prod_{j=l+1}^{d} S_{j;K}\Big) .$$

In particular $\kappa(\mathbf{U}, \mathbf{V}) = W_{d;K}$. Now by expanding a telescoping sum and applying the triangle inequality followed by the Jensen's inequality, we have

$$\begin{aligned}
&\mathbb{E}\big[|\kappa(\mathbf{U}, \mathbf{V}) - W_{0;K}|^\nu\big] \;=\; \mathbb{E}\Big[\Big|\sum_{l=1}^{d}(W_{l;K} - W_{l-1;K})\Big|^\nu\Big] \\
&\le \mathbb{E}\Big[\Big(\sum_{l=1}^{d}|W_{l;K} - W_{l-1;K}|\Big)^\nu\Big] \\
&\le d^{\nu-1}\sum_{l=1}^{d}\mathbb{E}[|W_{l;K} - W_{l-1;K}|^\nu] \\
&= d^{\nu-1}\sum_{l=1}^{d}\Big(\prod_{j=1}^{l-1}\mathbb{E}[|\kappa_1(U_j, V_j)|^\nu]\Big)\mathbb{E}[|\kappa_1(U_l, V_l) - S_{l;K}|^\nu]\Big(\prod_{j=l+1}^{d}\mathbb{E}[|S_{j;K}|^\nu]\Big) .
\end{aligned}$$

In the last equality, we have used the independence of U_j's and V_j's. To bound the summands, we first note that κ_1 is uniformly bounded in norm by 1, which implies that $\mathbb{E}[|\kappa_1(U_j, V_j)|^\nu] \le 1$. By the previous result, $\mathbb{E}[|\kappa_1(U_l, V_l) - S_{l;K}|^\nu] = o(1)$ as $K \to \infty$. By the triangle inequality and the Jensen's inequality, we have that

$$\begin{aligned}
\mathbb{E}[|S_{j;K}|^\nu] \;&\le \mathbb{E}\big[\big||\kappa_1(U_j, V_j)| + |S_{j;K} - \kappa_1(U_j, V_j)|\big|^\nu\big] \\
&\le 2^{\nu-1}\mathbb{E}\big[|\kappa_1(U_j, V_j)|^\nu\big] + 2^{\nu-1}\mathbb{E}\big[|S_{j;K} - \kappa_1(U_j, V_j)|^\nu\big] \;\le\; 2^{\nu-1} + o(1) .
\end{aligned}$$

This implies that each summand satisfies

$$\Big(\prod_{j=1}^{l-1}\mathbb{E}[|\kappa_1(U_j, V_j)|^\nu]\Big)\mathbb{E}[|\kappa_1(U_l, V_l) - S_{l;K}|^\nu]\Big(\prod_{j=l+1}^{d}\mathbb{E}[|S_{j;K}|^\nu]\Big) \;=\; o(1)$$

as $K \to \infty$. Since d is not affected by K, we have shown the desired result

$$\begin{aligned}
\mathbb{E}\Big[\Big|\kappa(\mathbf{U}, \mathbf{V}) - \prod_{j=1}^{d}\Big(\sum_{k=0}^{K}\lambda_k^*\phi_k^*(U_j)\phi_k^*(V_j)\Big)\Big|^\nu\Big] \;&= \mathbb{E}\big[|\kappa(\mathbf{U}, \mathbf{V}) - W_{0;K}|^\nu\big] \\
&\xrightarrow{K\to\infty} 0 .
\end{aligned}$$

□

Proof of Lemma A.2. We first rewrite u_P^{KSD} as

$$\begin{aligned}
u_P^{\mathrm{KSD}}(\mathbf{x}, \mathbf{x}') \;&= e^{-\|\mathbf{x}-\mathbf{x}'\|_2^2/(2\gamma)}\Big(\mathbf{x}^\top\mathbf{x}' - \frac{\gamma+1}{\gamma^2}\|\mathbf{x} - \mathbf{x}'\|_2^2 + \frac{d}{\gamma}\Big) \\
&= e^{-\|\mathbf{x}-\mathbf{x}'\|_2^2/(2\gamma)}\Big(-\frac{\gamma+1}{\gamma^2}(\|\mathbf{x}\|_2^2 + \|\mathbf{x}'\|_2^2) + \frac{\gamma^2+2\gamma+2}{\gamma^2}\mathbf{x}^\top\mathbf{x}' + \frac{d}{\gamma}\Big) \\
&= e^{-\|\mathbf{x}-\mathbf{x}'\|_2^2/(2\gamma)}\Big(-\frac{\gamma+1}{\gamma^2}(\|\mathbf{x}\|_2^2 + 1)(\|\mathbf{x}'\|_2^2 + 1) + \frac{\gamma+1}{\gamma^2}\|\mathbf{x}\|_2^2\|\mathbf{x}'\|_2^2 \\
&\qquad + \frac{\gamma^2+2\gamma+2}{\gamma^2}\sum_{l=1}^{d}x_l x_l' + \Big(\frac{d}{\gamma} + \frac{\gamma+1}{\gamma^2}\Big)\Big) .
\end{aligned}$$

For $K' \in \mathbb{N}$, write $S_{K'} \coloneqq \sum_{k'=1}^{K'} \alpha_{k'} \psi_{k'}(\mathbf{X}_1)\psi_{k'}(\mathbf{X}_2)$, and define the following random variables comparing each set of eigenvalue and eigenfunction to the corresponding term in u_P^{KSD}:

$$\begin{aligned}
T_{K';1} &= \sum_{k'=1}^{K'} \lambda_{(k'-1)(d+3)+1}\, \phi_{(k'-1)(d+3)+1}(\mathbf{X}_1)\, \phi_{(k'-1)(d+3)+1}(\mathbf{X}_1) \\
&\quad - e^{-\|\mathbf{X}_1-\mathbf{X}_2\|_2^2/(2\gamma)}\Big(- \frac{\gamma+1}{\gamma^2}(\|\mathbf{X}_1\|_2^2+1)(\|\mathbf{X}_2\|_2^2+1)\Big) \\
&= - \frac{\gamma+1}{\gamma^2}(\|\mathbf{X}_1\|_2^2+1)(\|\mathbf{X}_2\|_2^2+1) S_{K'} \,, \\
T_{K';2} &= \sum_{k'=1}^{K'} \lambda_{(k'-1)(d+3)+2}\, \phi_{(k'-1)(d+3)+2}(\mathbf{X}_1)\, \phi_{(k'-1)(d+3)+2}(\mathbf{X}_1) \\
&\quad - e^{-\|\mathbf{X}_1-\mathbf{X}_2\|_2^2/(2\gamma)}\Big(\frac{\gamma+1}{\gamma^2}\|\mathbf{X}_1\|_2^2\|\mathbf{X}_2\|_2^2\Big) \\
&= \frac{\gamma+1}{\gamma^2}\|\mathbf{X}_1\|_2^2\|\mathbf{X}_2\|_2^2\, S_{K'} \,,
\end{aligned}$$

$$\begin{aligned}
T_{K';3} &= \sum_{k'=1}^{K'} \lambda_{(k'-1)(d+3)+3}\, \phi_{(k'-1)(d+3)+3}(\mathbf{X}_1)\, \phi_{(k'-1)(d+3)+3}(\mathbf{X}_1) \\
&\quad - e^{-\|\mathbf{X}_1-\mathbf{X}_2\|_2^2/(2\gamma)}\Big(\frac{d}{\gamma} + \frac{\gamma+1}{\gamma^2}\Big) \\
&= \Big(\frac{d}{\gamma} + \frac{\gamma+1}{\gamma^2}\Big)\, S_{K'} \,, \\
T_{K';3+l} &= \sum_{k'=1}^{K'} \lambda_{(k'-1)(d+3)+3+l}\, \phi_{(k'-1)(d+3)+3+l}(\mathbf{X}_1)\, \phi_{(k'-1)(d+3)+3+l}(\mathbf{X}_1) \\
&\quad - e^{-\|\mathbf{X}_1-\mathbf{X}_2\|_2^2/(2\gamma)}\Big(\frac{\gamma^2+2\gamma+2}{\gamma^2}(\mathbf{X}_1)_l(\mathbf{X}_2)_l\Big) \\
&= \Big(\frac{\gamma^2+2\gamma+2}{\gamma^2}(\mathbf{X}_1)_l(\mathbf{X}_2)_l\Big)\, S_{K'}
\end{aligned}$$

for $l = 1, \ldots, d$, where we have denoted the l-th coordinates of $\mathbf{X}_1$ and $\mathbf{X}_2$ by $(\mathbf{X}_1)_l$ and $(\mathbf{X}_2)_l$ respectively. We now bound the approximation error with $(d+3)K'$ summands for $K' \in \mathbb{N}$ and $\nu \in (2, 3]$. Fix some $\nu_1 \in (\nu, 4]$ and let $\nu_2 = 1/(\nu^{-1} - \nu_1^{-1})$. By using the quantites defined above, the Jensen's inequality to the convex function $x \mapsto |x|^\nu$ and the Hölder's inequality to each $\mathbb{E}[|T_{K';l}|^\nu]$, we have

$$\begin{aligned}
&\mathbb{E}\Big[\Big| \sum_{k=1}^{(d+3)K'} \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2) - u_P^{\mathrm{KSD}}(\mathbf{X}_1, \mathbf{X}_2)\Big|^\nu\Big] \\
&= \mathbb{E}\Big[\Big| \sum_{l=1}^{d+3} T_{K';l}\Big|^\nu\Big] \\
&\le (d+3)^{\nu-1} \sum_{l=1}^{d+3} \mathbb{E}[|T_{K';l}|^\nu] \\
&\le (d+3)^{\nu-1} \mathbb{E}[|S_{K'}|^{\nu_1}]^{\nu/\nu_1} \Big(\Big(\frac{\gamma+1}{\gamma^2}\Big)^\nu \mathbb{E}[(\|\mathbf{X}_1\|_2^2+1)^{\nu_2}]^{\nu/\nu_2}\, \mathbb{E}[(\|\mathbf{X}_2\|_2^2+1)^{\nu_2}]^{\nu/\nu_2} \\
&\qquad + \Big(\frac{\gamma+1}{\gamma^2}\Big)^\nu \mathbb{E}\Big[\|\mathbf{X}_1\|_2^{2\nu_2}\Big]^{\nu/\nu_2} \mathbb{E}\Big[\|\mathbf{X}_2\|_2^{2\nu_2}\Big]^{\nu/\nu_2} + \Big(\frac{d}{\gamma} + \frac{\gamma+1}{\gamma^2}\Big)^\nu \\
&\qquad + \sum_{l=1}^{d} \Big(\frac{\gamma^2+2\gamma+2}{\gamma^2}\Big)^\nu \mathbb{E}\Big[|(\mathbf{X}_1)_l|^{\nu_2}\Big]^{\nu/\nu_2} \mathbb{E}\Big[|(\mathbf{X}_2)_l|^{\nu_2}\Big]^{\nu/\nu_2}\Big) \,.
\end{aligned}$$

The only K'-dependence above comes from $\mathbb{E}[|S_{K'}|^{\nu_1}]^{\nu/\nu_1} = \|S_{K'}\|_{L_{\nu_1}}^\nu$, which converges to 0 as K' grows by Lemma A.1. Therefore

$$\mathbb{E}\Big[\Big| \sum_{k=1}^{(d+3)K'} \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2) - u_P^{\mathrm{KSD}}(\mathbf{X}_1, \mathbf{X}_2)\Big|^\nu\Big] \xrightarrow{K' \to \infty} 0 \,.$$

Now for $K \in \mathbb{N}$ not necessarily divisible by $d+3$, we let K' be the largest integer such that $dK' \leq K$. By the triangle inequality and the Jensen's inequality, we have

$$\begin{aligned}
&\mathbb{E}\big[\big|\sum_{k=1}^{K} \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2) - u_P^{\mathrm{KSD}}(\mathbf{X}_1, \mathbf{X}_2)\big|^\nu\big] \\
&\quad \leq \mathbb{E}\Big[\Big(\big|\sum_{k=1}^{(d+3)K'} \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2) - u_P^{\mathrm{KSD}}(\mathbf{X}_1, \mathbf{X}_2)\big| \\
&\qquad\qquad + \big|\sum_{k=(d+3)K'+1}^{K} \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\big|\Big)^\nu\Big] \\
&\quad \leq 2^{\nu-1}\mathbb{E}\big[\big|\sum_{k=1}^{(d+3)K'} \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2) - u_P^{\mathrm{KSD}}(\mathbf{X}_1, \mathbf{X}_2)\big|^\nu\big] \\
&\qquad\qquad + 2^{\nu-1}\mathbb{E}\big[\big|\sum_{k=(d+3)K'+1}^{K} \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\big|^\nu\big] \,.
\end{aligned}$$

The goal is to show that the bound converges to 0 as K grows. We have already shown that the first term is $o(1)$, so we focus on the second term. The expectation in the second term can be bounded using the Jensen's inequality as

$$\begin{aligned}
\mathbb{E}\big[\big|\sum_{k=(d+3)K'+1}^{K} \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\big|^\nu\big] \;&\leq\; \mathbb{E}\big[\big(\sum_{k=(d+3)K'+1}^{K} |\lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)|\big)^\nu\big] \\
&\leq (K-(d+3)K')^{\nu-1} \sum_{k=(d+3)K'+1}^{K} \mathbb{E}\big[\big(\lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\big)^\nu\big] \\
&\leq d^\nu \sup_{k \in \{(d+3)K'+1, \ldots, (d+3)K'+(d+3)\}} \mathbb{E}\big[\big(\lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\big)^\nu\big] \\
&= d^\nu \sup_{1 \leq l \leq d+3} \mathbb{E}\big[\big(\lambda_{(d+3)K'+l} \phi_{(d+3)K'+l}(\mathbf{X}_1)\phi_{(d+3)K'+l}(\mathbf{X}_2)\big)^\nu\big] \,.
\end{aligned}$$

By observing the formula for λ_k and ϕ_k, we see that there exists some K-independent constant $C_{d,\gamma}$ such that for $1 \leq l \leq d+3$,

$$|\lambda_{(d+3)K'+l}| \;\leq\; C_{d,\gamma}\alpha_{K'+1} \quad \text{and} \quad |\phi_{(d+3)K'+l}| \;\leq\; C_{d,\gamma}\psi_{K'+1}(\mathbf{x})(\|\mathbf{x}\|_2^2 + \|\mathbf{x}\|_2 + 1) \,.$$

This allows us to obtain the bound

$$\begin{aligned}
&\mathbb{E}\big[\big|\sum_{k=(d+3)K'+1}^{K} \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\big|^\nu\big] \\
&\leq d^\nu C_{d,\gamma}^2 \alpha_{K'+1}^\nu \\
&\qquad \times \mathbb{E}[(\psi_{K'+1}(\mathbf{X}_1)\psi_{K'+1}(\mathbf{X}_2))^\nu \, (\|\mathbf{X}_1\|_2^2 + \|\mathbf{X}_1\|_2 + 1)^\nu (\|\mathbf{X}_2\|_2^2 + \|\mathbf{X}_2\|_2 + 1)^\nu] \\
&\overset{(a)}{=} d^\nu C'_{d,\gamma} \Big(\prod_{l=1}^{d} \lambda^*_{[g_d(K'+1)]_l}\Big)^\nu \mathbb{E}\Big[\Big(\prod_{l=1}^{d} \phi^*_{[g_d(K'+1)]_l}\big((\mathbf{X}_1)_l\big)\, \phi^*_{[g_d(K'+1)]_l}\big((\mathbf{X}_2)_l\big)\Big)^\nu \\
&\qquad\qquad \times (\|\mathbf{X}_1\|_2^2 + \|\mathbf{X}_1\|_2 + 1)^\nu (\|\mathbf{X}_2\|_2^2 + \|\mathbf{X}_2\|_2 + 1)^\nu\Big] \\
&\overset{(b)}{\leq} d^\nu C'_{d,\gamma} \mathbb{E}\big[(\|\mathbf{X}_1\|_2^2 + \|\mathbf{X}_1\|_2 + 1)^{2\nu} (\|\mathbf{X}_2\|_2^2 + \|\mathbf{X}_2\|_2 + 1)^{2\nu}\big]^{1/2} \\
&\qquad \times \Big(\prod_{l=1}^{d} \lambda^*_{[g_d(K'+1)]_l}\Big)^\nu \mathbb{E}\Big[\Big(\prod_{l=1}^{d} \phi^*_{[g_d(K'+1)]_l}\big((\mathbf{X}_1)_l\big)\, \phi^*_{[g_d(K'+1)]_l}\big((\mathbf{X}_2)_l\big)\Big)^{2\nu}\Big]^{1/2} \\
&\overset{(c)}{=} d^\nu C'_{d,\gamma} \mathbb{E}\big[(\|\mathbf{X}_1\|_2^2 + \|\mathbf{X}_1\|_2 + 1)^{2\nu} (\|\mathbf{X}_2\|_2^2 + \|\mathbf{X}_2\|_2 + 1)^{2\nu}\big]^{1/2} \\
&\qquad \times \prod_{l=1}^{d} \big(\lambda^*_{[g_d(K'+1)]_l}\big)^\nu \\
&\qquad\qquad \Big(\prod_{l=1}^{d} \mathbb{E}\big[\big(\phi^*_{[g_d(K'+1)]_l}\big((\mathbf{X}_1)_l\big)\big)^{2\nu}\big] \mathbb{E}\big[\big(\phi^*_{[g_d(K'+1)]_l}\big((\mathbf{X}_2)_l\big)\big)^{2\nu}\big]\Big)^{1/2} \\
&\overset{(d)}{=} d^\nu C'_{d,\gamma} \mathbb{E}\big[(\|\mathbf{X}_1\|_2^2 + \|\mathbf{X}_1\|_2 + 1)^{2\nu} (\|\mathbf{X}_2\|_2^2 + \|\mathbf{X}_2\|_2 + 1)^{2\nu}\big]^{1/2}
\end{aligned}$$

$$\times \prod_{l=1}^{d} \Big(\big(\lambda^*_{[g_d(K'+1)]_l}\big)^{\nu} \, \mathbb{E}\big[\big(\phi^*_{[g_d(K'+1)]_l}((\mathbf{X}_1)_l)\big)^{2\nu} \big] \Big) \, ,$$

where we have used the definitions of α_k and ψ_k from (A.1) in (a), the Cauchy-Schwarz inequality in (b), the independence of $(\mathbf{X}_1)_l$ and $(\mathbf{X}_2)_l$ for $1 \leq l \leq d$ due to the identity covariance matrix in (c) and finally the fact that $\mathbf{X}_1$ and $\mathbf{X}_2$ are identically distributed in (d). The only quantity that depends on K' now is

$$\big(\lambda^*_{[g_d(K'+1)]_l}\big)^{\nu} \, \mathbb{E}\big[\big(\phi^*_{[g_d(K'+1)]_l}((\mathbf{X}_1)_l)\big)^{2\nu} \big]$$

for $1 \leq l \leq d$. We now seek to bound this quantity. Recall from Lemma A.1 that $\lambda^*_k := \frac{1}{k! \, \gamma^k}$, and for $V \sim \mathcal{N}(b, 1)$, we have

$$\mathbb{E}[(\phi^*_k(U))^{2\nu}] \;=\; \mathbb{E}\big[|U|^{2\nu k} e^{-\nu U^2/\gamma}\big] \;\leq\; \mathbb{E}\big[|U|^{2\nu k}\big] \;\leq\; \frac{(2b)^{2\nu k}}{2} + \frac{2^{3\nu k}}{2\sqrt{\pi}} \, \Gamma\Big(\frac{2\nu k + 1}{2}\Big) \, .$$

where we have used a bound similar to (A.16) in the proof of Lemma A.1. By Stirling's formula for the gamma function, we have $\Gamma(x) = \sqrt{2\pi} \, x^{x-1/2} e^{-x} \big(1 + O(x^{-1})\big)$ for $x > 0$, which implies

$$\begin{aligned}
(\lambda^*_k)^{\nu} \, \mathbb{E}[(\phi^*_k(U))^{2\nu}] \;&\leq \frac{1}{(k!)^{\nu} \, \gamma^{\nu k}} \Big(\frac{(2b)^{2\nu k}}{2} + \frac{2^{3\nu k}}{2\sqrt{\pi}} \, \Gamma\Big(\frac{2\nu k + 1}{2}\Big) \Big) \\
&= O\Big(\Big(\frac{8}{\gamma}\Big)^{\nu k} \frac{(\nu k)^{\nu k} e^{-\nu k}}{(k+1)^{\nu(k+1/2)} e^{-\nu(k+1)}} \Big) \\
&= O\Big(\Big(\frac{8\nu}{\gamma}\Big)^{\nu k} \Big) \;=\; O\Big(\Big(\frac{24}{\gamma}\Big)^{\nu k} \Big) \;=\; o(1)
\end{aligned}$$

as $k \to \infty$, where we have used the assumption that $\gamma > 24$. By construction of g_d in (A.1), as $K' \to \infty$, $\min_{1 \leq l \leq d} [g_d(K'+1)]_l \to \infty$, which implies that

$$\big(\lambda^*_{[g_d(K'+1)]_l}\big)^{\nu} \, \mathbb{E}\big[\big(\phi^*_{[g_d(K'+1)]_l}((\mathbf{X}_1)_l)\big)^{2\nu} \big] \;\xrightarrow{K' \to \infty}\; 0 \, .$$

Therefore

$$\mathbb{E}\big[\big| \textstyle\sum_{k=(d+3)K'+1}^{K} \lambda_k \phi_k(\mathbf{X}_1) \phi_k(\mathbf{X}_2) \big|^{\nu} \big] \;\xrightarrow{K \to \infty}\; 0 \, ,$$

which finishes the proof that

$$\mathbb{E}\big[\big| \textstyle\sum_{k=1}^{K} \lambda_k \phi_k(\mathbf{X}_1) \phi_k(\mathbf{X}_2) - u_P^{\mathrm{KSD}}(\mathbf{X}_1, \mathbf{X}_2) \big|^{\nu} \big] \;\xrightarrow{K \to \infty}\; 0 \, .$$

In other words, Assumption 3.2 holds. □

Proof of Lemma A.3. Fix $\nu \in (2, 3]$. Consider the independent Gaussian vectors $\mathbf{X}_1, \mathbf{X}_2 \overset{\text{i.i.d.}}{\sim} P \equiv \mathcal{N}(\mathbf{0}, I_d)$ and $\mathbf{Y}_1, \mathbf{Y}_2 \overset{i.i.d.}{\sim} Q \equiv \mathcal{N}(\mu, I_d)$. Write $\mathbf{Z}_1 = (\mathbf{X}_1, \mathbf{Y}_1)$, $\mathbf{Z}_2 = (\mathbf{X}_2, \mathbf{Y}_2)$ and

$$T_K(\mathbf{x}, \mathbf{x}') \;:=\; e^{-\|\mathbf{x} - \mathbf{x}'\|_2^2/(2\gamma)} - \textstyle\sum_{k=1}^{K} \alpha_k \psi_k(\mathbf{x}) \psi_k(\mathbf{x}')$$

for $K \in \mathbb{N}$, and recall that

$$u^{\mathrm{MMD}}(\mathbf{Z}_1, \mathbf{Z}_2) = e^{-\|\mathbf{X}_1-\mathbf{X}_2\|_2^2/(2\gamma)} - e^{-\|\mathbf{X}_1-\mathbf{Y}_2\|_2^2/(2\gamma)} - e^{-\|\mathbf{X}_2-\mathbf{Y}_1\|_2^2/(2\gamma)} + e^{-\|\mathbf{Y}_1-\mathbf{Y}_2\|_2^2/(2\gamma)} .$$

Then by the triangle inequality and Jensen's inequality, we get that

$$\begin{aligned}
&\mathbb{E}\Big[\Big|u^{\mathrm{MMD}}(\mathbf{Z}_1, \mathbf{Z}_2) - \textstyle\sum_{k=1}^K \lambda_k \phi_k(\mathbf{Z}_1)\phi_k(\mathbf{Z}_2)\Big|^\nu\Big] \\
&= \mathbb{E}\Big[\Big|u^{\mathrm{MMD}}(\mathbf{Z}_1, \mathbf{Z}_2) - \textstyle\sum_{k=1}^K \alpha_k\big(\psi_k(\mathbf{X}_1) - \psi_k(\mathbf{Y}_1)\big)\big(\psi_k(\mathbf{X}_2) - \psi_k(\mathbf{Y}_2)\big)\Big|^\nu\Big] \\
&= \mathbb{E}\big[\big|T_K(\mathbf{X}_1, \mathbf{X}_2) - T_K(\mathbf{X}_1, \mathbf{Y}_2) - T_K(\mathbf{X}_2, \mathbf{Y}_1) + T_K(\mathbf{X}_2, \mathbf{Y}_2)\big|^\nu\big] \\
&\le 4^{\nu-1}\big(\mathbb{E}[|T_K(\mathbf{X}_1, \mathbf{X}_2)|^\nu] + \mathbb{E}[|T_K(\mathbf{X}_1, \mathbf{Y}_2)|^\nu] \\
&\qquad + \mathbb{E}[|T_K(\mathbf{X}_2, \mathbf{Y}_1)|^\nu] + \mathbb{E}[|T_K(\mathbf{Y}_1, \mathbf{Y}_2)|^\nu]\big) .
\end{aligned}$$

Since each expectation is taken with respect to a product of two Gaussian distributions with identity covariance matrices, by Lemma A.1 and (A.2), they all decay to 0 as $K \to \infty$. This proves that

$$\mathbb{E}\Big[\Big|u^{\mathrm{MMD}}(\mathbf{Z}_1, \mathbf{Z}_2) - \textstyle\sum_{k=1}^K \lambda_k \phi_k(\mathbf{Z}_1)\phi_k(\mathbf{Z}_2)\Big|^\nu\Big] \xrightarrow{K\to\infty} 0 ,$$

and therefore Assumption 3.2 holds. □

A.7 Proofs for Appendix A.2

A.7.1. Proofs for Appendix A.2.1

The proof of Lemma A.4 combines the following two results:

Lemma A.20 (Theorem 2, von Bahr and Esseen (1965))**.** *Fix $\nu \in [1, 2]$. For a martingale difference sequence $Y_1, \ldots, Y_n$ taking values in $\mathbb{R}$,*

$$\mathbb{E}\big[\big|\textstyle\sum_{i=1}^n Y_i\big|^\nu\big] \le 2\textstyle\sum_{i=1}^n \mathbb{E}[|Y_i|^\nu] .$$

Lemma A.21 (Dharmadhikari et al. (1968b))**.** *Fix $\nu \ge 2$. For a martingale difference sequence $Y_1, \ldots, Y_n$ taking values in $\mathbb{R}$,*

$$\mathbb{E}\big[\big|\textstyle\sum_{i=1}^n Y_i\big|^\nu\big] \le C_\nu n^{\nu/2-1}\textstyle\sum_{i=1}^n \mathbb{E}[|Y_i|^\nu] ,$$

where $C_\nu = (8(\nu-1)\max\{1, 2^{\nu-3}\})^\nu$.

Proof of Lemma A.4. Since the second line directly follows from Burkholder (1966)'s original result, it suffices to prove the first line. We first consider the upper bound. For $\nu \in [1, 2]$, the result follows directly from the Von Bahn-Esseen inequality as stated

below in Lemma A.20, and for $\nu > 1$, the result follows directly from Lemma A.21. As for the lower bound, by Theorem 9 of Burkholder (1966), there exists an absolute constant $c_\nu > 0$ depending only on ν such that

$$\mathbb{E}\big[\big|\sum_{i=1}^n Y_i\big|^\nu\big] \;\geq\; c_\nu\, \mathbb{E}\big[\big(\sum_{i=1}^n Y_i^2\big)^{\nu/2}\big]\,.$$

For $\nu \in [1,2]$, by applying Jensen's inequality on the concave function $x \mapsto x^{\nu/2}$, we get that

$$\mathbb{E}\big[\big|\sum_{i=1}^n Y_i\big|^\nu\big] \;\geq c_\nu\, \mathbb{E}\big[\big(\tfrac{1}{n}\sum_{i=1}^n nY_i^2\big)^{\nu/2}\big] \;\geq\; c_\nu\, n^{\nu/2-1} \sum_{i=1}^n \mathbb{E}[|Y_i|^\nu]\,.$$

For $\nu > 2$, by noting that $(a+b)^{\nu/2} \geq a^{\nu/2} + b^{\nu/2}$ for $a, b \geq 0$, we get that

$$\mathbb{E}\big[\big|\sum_{i=1}^n Y_i\big|^\nu\big] \;\geq c_\nu\, \mathbb{E}\big[\sum_{i=1}^n \big(Y_i^2\big)^{\nu/2}\big] \;\geq\; c_\nu \sum_{i=1}^n \mathbb{E}[|Y_i|^\nu]\,.$$

Combining the two results above give the desired bound. □

A.7.2. Proofs for Appendix A.2.2

Proof of Lemma A.6. Consider the sequence of sigma algebras with $\mathcal{F}_0$ being the trivial sigma algebra and $\mathcal{F}_i := \sigma(X_1, \ldots, X_i)$ for $i = 1, \ldots, n$. This allows us to define a martingale difference sequence: For $i = 1, \ldots, n$, let

$$Y_i \;:=\; \mathbb{E}[D_n|\mathcal{F}_i] - \mathbb{E}[D_n|\mathcal{F}_{i-1}]\,.$$

This implies that $\mathbb{E}[|D_n - \mathbb{E}D_n|^\nu] = \mathbb{E}\big[\big|\sum_{i=1}^n Y_i\big|^\nu\big]$. By Lemma A.4, we get that for some universal constants c'_ν, C'_ν,

$$c'_\nu \sum_{i=1}^n \mathbb{E}[|Y_i|^\nu] \;\leq\; \mathbb{E}[|D_n - \mathbb{E}D_n|^\nu] \;\leq\; C'_\nu\, n^{\nu/2-1} \sum_{i=1}^n \mathbb{E}[|Y_i|^\nu]\,. \tag{A.17}$$

To compute the ν-th moment of Y_i, recall that $D_n = \frac{1}{n(n-1)} \sum_{j,l\in[n],j\neq l} u(\mathbf{X}_j, \mathbf{X}_l)$, which implies

$$\begin{aligned}
\mathbb{E}[|Y_i|^\nu] &= \mathbb{E}\big[\big|\mathbb{E}[D_n|\mathcal{F}_i] - \mathbb{E}[D_n|\mathcal{F}_{i-1}]\big|^\nu\big] \\
&= \frac{1}{n^\nu(n-1)^\nu}\mathbb{E}\Big[\Big|\sum\nolimits_{j,l\in[n],j\neq l}\big(\mathbb{E}[u(\mathbf{X}_j,\mathbf{X}_l)|\mathcal{F}_i] - \mathbb{E}[u(\mathbf{X}_j,\mathbf{X}_l)|\mathcal{F}_{i-1}]\big)\Big|^\nu\Big] \\
&\overset{(a)}{=} \frac{2}{n^\nu(n-1)^\nu}\mathbb{E}\Big[\Big|\sum\nolimits_{j\in[n],j\neq i}\big(\mathbb{E}[u(\mathbf{X}_i,\mathbf{X}_j)|\mathcal{F}_i] - \mathbb{E}[u(\mathbf{X}_i,\mathbf{X}_j)|\mathcal{F}_{i-1}]\big)\Big|^\nu\Big] \\
&=: \frac{2}{n^\nu(n-1)^\nu}\,\mathbb{E}[|S_i|^\nu]\,.
\end{aligned}$$

In (a), we have used that each summand is zero if both j and l do not equal i, and that u is symmetric. In the case $j < i$, we can compute each summand of S_i as

$$\begin{aligned}
\mathbb{E}[u(\mathbf{X}_i,\mathbf{X}_j)|\mathcal{F}_i] - \mathbb{E}[u(\mathbf{X}_i,\mathbf{X}_j)|\mathcal{F}_{i-1}] &= u(\mathbf{X}_i,\mathbf{X}_j) - \mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_j)|\mathbf{X}_j] \\
&= A_{ij} - B_j + B_i\,,
\end{aligned}$$

where $A_{ij} := u(\mathbf{X}_i, \mathbf{X}_j) - \mathbb{E}[u(\mathbf{X}_i, \mathbf{X}_1)|\mathbf{X}_i]$ and

$$B_i \;:=\; \mathbb{E}[u(\mathbf{X}_i, \mathbf{X}_1)|\mathbf{X}_i] - \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)] \;=\; \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_i)|\mathbf{X}_i] - \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)]$$

by symmetry of u. In the case $j > i$, we can compute each summand as

$$\mathbb{E}[u(\mathbf{X}_i, \mathbf{X}_j)|\mathcal{F}_i] - \mathbb{E}[u(\mathbf{X}_i, \mathbf{X}_j)|\mathcal{F}_{i-1}] \;=\; \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_i)|\mathbf{X}_i] - \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)] \;=\; B_i \,.$$

Therefore

$$S_i \;=\; \textstyle\sum_{j<i}(A_{ij} - B_j) + nB_i \,.$$

Consider $R_1 := nB_i$ and $R_2 := \sum_{j<i}(A_{ij} - B_j)$, which forms a two-element martingale difference sequence with respect to the filtration $\sigma(\mathbf{X}_i) \subseteq \sigma(\mathbf{X}_i, \mathbf{X}_1 \ldots, \mathbf{X}_{i-1})$. By Lemma A.4 again, there exist constants c_ν^* and C_ν^* depending only on ν such that

$$\begin{aligned}
\mathbb{E}[|S_i|^\nu] \;=\; \mathbb{E}\big[\big|\textstyle\sum_{l=1}^2 R_l\big|^\nu\big] \;&\le\; C_\nu^*\Big(\mathbb{E}[|nB_i|^\nu] + \mathbb{E}\big[\big|\textstyle\sum_{j<i}(A_{ij} - B_j)\big|^\nu\big]\Big) \\
&=\; C_\nu^*\big(n^\nu M_{\mathrm{cond};\nu}^\nu + \mathbb{E}\big[\big|\textstyle\sum_{j<i}(A_{ij} - B_j)\big|^\nu\big]\big) \,, \\
\mathbb{E}[|S_i|^\nu] \;=\; \mathbb{E}\big[\big|\textstyle\sum_{l=1}^2 R_l\big|^\nu\big] \;&\ge\; c_\nu^*\Big(\mathbb{E}[|nB_i|^\nu] + \mathbb{E}\big[\big|\textstyle\sum_{j<i}(A_{ij} - B_j)\big|^\nu\big]\Big) \\
&=\; c_\nu^*\big(n^\nu M_{\mathrm{cond};\nu}^\nu + \mathbb{E}\big[\big|\textstyle\sum_{j<i}(A_{ij} - B_j)\big|^\nu\big]\big) \,.
\end{aligned}$$

Now consider $T_j := A_{ij} - B_j$ for $j = 1, \ldots, i-1$, which again forms a martingale difference sequence with respect to $\sigma(\mathbf{X}_i, \mathbf{X}_1), \ldots, \sigma(\mathbf{X}_i, \mathbf{X}_1 \ldots, \mathbf{X}_{i-1})$. Then by Lemma A.4 again, there exist constants c_ν^Δ and C_ν^Δ depending only on ν such that

$$\begin{aligned}
\mathbb{E}\big[\big|\textstyle\sum_{j<i}(A_{ij} - B_j)\big|^\nu\big] \;&\le C_\nu^\Delta\,(i-1)^{\nu/2-1}\textstyle\sum_{j=1}^{i-1}\mathbb{E}[|A_{ij} - B_j|^\nu] \\
&= C_\nu^\Delta\,(i-1)^{\nu/2} M_{\mathrm{full};\nu}^\nu \,, \\
\mathbb{E}\big[\big|\textstyle\sum_{j<i}(A_{ij} - B_j)\big|^\nu\big] \;&\ge\; c_\nu^\Delta\,\textstyle\sum_{j=1}^{i-1}\mathbb{E}[|A_{ij} - B_j|^\nu] \;=\; c_\nu^\Delta\,(i-1) M_{\mathrm{full};\nu}^\nu \,.
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathbb{E}[|S_i|^\nu] \;&\le C_\nu^* n^\nu M_{\mathrm{cond};\nu}^\nu + C_\nu^* C_\nu^\Delta\,(i-1)^{\nu/2} M_{\mathrm{full};\nu}^\nu \,, \\
\mathbb{E}[|S_i|^\nu] \;&\ge c_\nu^* n^\nu M_{\mathrm{cond};\nu}^\nu + c_\nu^* c_\nu^\Delta\,(i-1) M_{\mathrm{full};\nu}^\nu \,,
\end{aligned}$$

which yield the following bounds on the ν-th moment of Y_i:

$$\begin{aligned}
\mathbb{E}[|Y_i|^\nu] \;&\le\; 2C_\nu^*\big((n-1)^{-\nu} M_{\mathrm{cond};\nu}^\nu + C_\nu^\Delta n^{-\nu}(n-1)^{-\nu}\,(i-1)^{\nu/2} M_{\mathrm{full};\nu}^\nu\big) \,, \\
\mathbb{E}[|Y_i|^\nu] \;&\ge\; 2c_\nu^*\big((n-1)^{-\nu} M_{\mathrm{cond};\nu}^\nu + c_\nu^\Delta n^{-\nu}(n-1)^{-\nu}\,(i-1) M_{\mathrm{full};\nu}^\nu\big) \,,
\end{aligned}$$

To sum these terms over $i = 1, \ldots, n$, we note that since $\nu/2 > 0$,

$$\textstyle\sum_{i=1}^n (i-1)^{\nu/2} \;\le\; \int_0^n x^{\nu/2} dx \;=\; \frac{n^{1+\nu/2}}{1+\nu/2} \,, \qquad \sum_{i=1}^n (i-1) \;=\; \frac{n(n-1)}{2} \,.$$

Define $C_\nu := \frac{2C'_\nu C^*_\nu \max\{1, C^\Delta_\nu\}}{1+\nu/2}$ and $c_\nu := c'_\nu c^*_\nu \min\{1, c^\Delta_\nu\}$. By summing the bounds on $\mathbb{E}[|Y_i|^\nu]$ and substituting into (A.17), we get the desired bounds

$$\begin{aligned}
\mathbb{E}[|D_n - \mathbb{E}D_n|^\nu] &\le C_\nu\, n^{\nu/2-1}\left(n(n-1)^{-\nu} M^\nu_{\text{cond};\nu} + n^{-\nu}(n-1)^{-\nu} n^{1+\nu/2} M^\nu_{\text{full};\nu}\right) \\
&= C_\nu\, n^{\nu/2}(n-1)^{-\nu} M^\nu_{\text{cond};\nu} + C_\nu\,(n-1)^{-\nu} M^\nu_{\text{full};\nu}\ , \\
\mathbb{E}[|D_n - \mathbb{E}D_n|^\nu] &\ge c_\nu n(n-1)^{-\nu} M^\nu_{\text{cond};\nu} + c_\nu n^{-(\nu-1)}(n-1)^{-(\nu-1)} M^\nu_{\text{full};\nu}\ .
\end{aligned}$$

□

Proof of Lemma A.7. The first result is directly obtained from linearity of expectation and Jensen's inequality:

$$\begin{aligned}
\left|D - \textstyle\sum_{k=1}^K \lambda_k \mu_k^2\right| &= \left|\mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)] - \textstyle\sum_{k=1}^K \lambda_k \mathbb{E}[\phi_k(\mathbf{X}_1)]\mathbb{E}[\phi_k(\mathbf{X}_2)]\right| \\
&= \left|\mathbb{E}\left[u(\mathbf{X}_1, \mathbf{X}_2) - \textstyle\sum_{k=1}^K \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\right]\right| \\
&\le \mathbb{E}\left|u(\mathbf{X}_1, \mathbf{X}_2) - \textstyle\sum_{k=1}^K \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\right| \;=\; \varepsilon_{K;1}\ .
\end{aligned}$$

To prove the next few bounds, we first derive a useful inequality: For $a, b \in \mathbb{R}$ and $\nu \ge 1$, by Jensen's inequality, we have

$$|a+b|^\nu \;=\; \left|\tfrac{1}{2}(2a) + \tfrac{1}{2}(2b)\right|^\nu \;\le\; \tfrac{1}{2}|2a|^\nu + \tfrac{1}{2}|2b|^\nu \;=\; 2^{\nu-1}(|a|^\nu + |b|^\nu)\ .$$

By the triangle inequality followed by applying the above inequality again with a replaced by $|a| - |b|$ and b replaced by $|b|$, we have

$$|a+b|^\nu \;\ge\; \big||a| - |b|\big|^\nu \;\ge\; 2^{-(\nu-1)}|a|^\nu - |b|^\nu\ .$$

Since $\nu \in [1, 3]$, we have $2^{\nu-1} \in [1, 4]$. Therefore

$$\tfrac{1}{4}|a|^\nu - |b|^\nu \;\le\; |a+b|^\nu \;\le\; 4(|a|^\nu + |b|^\nu)\ . \tag{A.18}$$

Now to prove the conditional bound, we make use of the fact that $\mathbf{X}_1, \mathbf{X}_2$ are i.i.d. to see that

$$\begin{aligned}
&\mathbb{E}\left[\left|\textstyle\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1) - \mu_k)\mu_k\right|^\nu\right] \\
&= \mathbb{E}\left[\left|\textstyle\sum_{k=1}^K \lambda_k\big(\mathbb{E}[\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)|\mathbf{X}_1] - \mathbb{E}[\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)]\big)\right|^\nu\right] \\
&= \mathbb{E}\left[\left|\mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{X}_1] - \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)] + \Delta_{K;1} - \Delta_{K;2}\right|^\nu\right]\ ,
\end{aligned} \tag{A.19}$$

where

$$\begin{aligned}
\Delta_{K;1} &:= \textstyle\sum_{k=1}^K \lambda_k \mathbb{E}[\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)|\mathbf{X}_1] - \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{X}_1]\ , \\
\Delta_{K;2} &:= \textstyle\sum_{k=1}^K \lambda_k \mathbb{E}[\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)] - \mathbb{E}[u(\mathbf{X}_1, \mathbf{X}_2)]\ .
\end{aligned}$$

Moments of the two error terms can be bounded by Jensen's inequality applied to $x \mapsto |x|^\nu$ with respect to the conditional expectation $\mathbb{E}[\,\bullet\,|\mathbf{X}_2]$ and the expectation $\mathbb{E}[\,\bullet\,]$:

$$\begin{aligned}\mathbb{E}[|\Delta_{K;1}|^\nu], \mathbb{E}[|\Delta_{K;2}|^\nu] &\le \mathbb{E}\Big[\Big|u(\mathbf{X}_1,\mathbf{X}_2) - \textstyle\sum_{k=1}^K \lambda_k\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\Big|^\nu\Big] \\ &= \Big\|u(\mathbf{X}_1,\mathbf{X}_2) - \textstyle\sum_{k=1}^K \lambda_k\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\Big\|_{L_\nu}^\nu \;=\; \varepsilon_{K;\nu}^\nu\,.\end{aligned}$$

On the other hand,

$$(M_{\mathrm{cond};\nu})^\nu \;=\; \mathbb{E}\big[\big|\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{X}_1] - \mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)]\big|^\nu\big]\,.$$

Therefore applying (A.18) gives

$$\tfrac{1}{4}(M_{\mathrm{cond};\nu})^\nu - \varepsilon_{K;\nu}^\nu \;\le\; \mathbb{E}\Big[\Big|\textstyle\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1)-\mu_k)\mu_k\Big|^\nu\Big] \;\le\; 4((M_{\mathrm{cond};\nu})^\nu + \varepsilon_{K;\nu}^\nu)$$

For the last bound, we start by considering the following quantity, which can be thought of as the truncated version of $M_{\mathrm{full};\nu}^\nu$:

$$\begin{aligned}m_K &:= \mathbb{E}\Big[\Big|\textstyle\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2) - \mu_k^2)\Big|^\nu\Big] \\ &= \mathbb{E}\Big[\Big|\textstyle\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1)-\mu_k)\phi_k(\mathbf{X}_2) + \sum_{k=1}^K \lambda_k\mu_k(\phi_k(\mathbf{X}_2)-\mu_k)\Big|^\nu\Big] \\ &=: \mathbb{E}[|T_2+T_1|^\nu]\,.\end{aligned}$$

Since $\{T_1, T_2\}$ forms a two-element martingale difference sequence with respect to $\sigma(\mathbf{X}_2) \subseteq \sigma(\mathbf{X}_1,\mathbf{X}_2)$, by Lemma A.4, there exists absolute constants $c_\nu', C_\nu' > 0$ depending only on ν such that

$$c_\nu'\big(\mathbb{E}[|T_1|^\nu] + \mathbb{E}[|T_2|^\nu]\big) \;\le\; m_K \;\le\; C_\nu'\big(\mathbb{E}[|T_1|^\nu] + \mathbb{E}[|T_2|^\nu]\big)\,.$$

Similarly, by writing

$$\begin{aligned}\mathbb{E}[|T_2|^\nu] &= \mathbb{E}\Big[\Big|\textstyle\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1)-\mu_k)\phi_k(\mathbf{X}_2)\Big|^\nu\Big] \\ &= \mathbb{E}\Big[\Big|\textstyle\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1)-\mu_k)(\phi_k(\mathbf{X}_2)-\mu_k) + \sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1)-\mu_k)\mu_k\Big|^\nu\Big] \\ &= \mathbb{E}[|R_2+R_1|^\nu]\,,\end{aligned}$$

and noting that $\{R_1, R_2\}$ forms a two-element martingale difference sequence with respect to $\sigma(\mathbf{X}_1) \subseteq \sigma(\mathbf{X}_1,\mathbf{X}_2)$, by Lemma A.4, there exists absolute constants $c_\nu'', C_\nu'' > 0$ depending only on ν such that

$$c_\nu''\big(\mathbb{E}[|R_1|^\nu] + \mathbb{E}[|R_2|^\nu]\big) \;\le\; \mathbb{E}[|T_2|^\nu] \;\le\; C_\nu''\big(\mathbb{E}[|R_1|^\nu] + \mathbb{E}[|R_2|^\nu]\big)\,.$$

Combining the results and setting

$$A \;=\; \sup_{\nu\in[1,3]} C_\nu' \max\{C_\nu'', 1\} \qquad \text{and} \qquad a \;=\; \inf_{\nu\in[1,3]} c_\nu' \min\{c_\nu'', 1\}\,,$$

we have shown that

$$a\big(\mathbb{E}[|T_1|^\nu] + \mathbb{E}[|R_1|^\nu] + \mathbb{E}[|R_2|^\nu]\big) \;\le\; m_K \;\le A\big(\mathbb{E}[|T_1|^\nu] + \mathbb{E}[|R_1|^\nu] + \mathbb{E}[|R_2|^\nu]\big)\,.$$

Notice that the quantity we would like to control is exactly

$$\mathbb{E}[|R_2|^\nu] \;= \mathbb{E}\Big[\Big|\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1)-\mu_k)(\phi_k(\mathbf{X}_2)-\mu_k)\Big|^\nu\Big]\;,$$

and that $\mathbb{E}[|T_1|^\nu] = \mathbb{E}[|R_1|^\nu]$. By setting $c = A^{-1}$ and $C = a^{-1}$, this allows us to obtain a bound about $\mathbb{E}[|R_2|^\nu]$ as

$$cm_K - 2\mathbb{E}[|T_1|^\nu] \le \mathbb{E}\Big[\Big|\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1)-\mu_k)(\phi_k(\mathbf{X}_2)-\mu_k)\Big|^\nu\Big] \le Cm_K - 2\mathbb{E}[|T_1|^\nu]\;.$$

Now notice that

$$\mathbb{E}[|T_1|^\nu] \;=\; \mathbb{E}\Big[\Big|\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1)-\mu_k)\mu_k\Big|^\nu\Big]\;,$$

which has already been controlled by the second result of the lemma as

$$\frac{1}{4}(M_{\text{cond};\nu})^\nu - \varepsilon_{K;\nu}^\nu \;\le\; \mathbb{E}[|T_1|^\nu] \;\le\; 4((M_{\text{cond};\nu})^\nu + \varepsilon_{K;\nu}^\nu)\;.$$

On the other hand, we can use an exactly analogous argument by using (A.18) and applying Jensen's inequality to control the errors to show that

$$\frac{1}{4}(M_{\text{full};\nu})^\nu - \varepsilon_{K;\nu}^\nu \;\le\; m_K \;\le\; 4((M_{\text{full};\nu})^\nu + \varepsilon_{K;\nu}^\nu)\;.$$

Applying these two results to the previous bound gives the desired bounds:

$$\begin{aligned}
&\mathbb{E}\Big[\Big|\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1)-\mu_k)(\phi_k(\mathbf{X}_2)-\mu_k)\Big|^\nu\Big] \\
&\qquad\le\; 4C(M_{\text{full};\nu})^\nu - \frac{1}{2}(M_{\text{cond};\nu})^\nu + (4C+2)\varepsilon_{K;\nu}^\nu\;, \\
&\mathbb{E}\Big[\Big|\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1)-\mu_k)(\phi_k(\mathbf{X}_2)-\mu_k)\Big|^\nu\Big] \\
&\qquad\ge\; \frac{c}{4}(M_{\text{full};\nu})^\nu - 8(M_{\text{cond};\nu})^\nu - (c+8)\varepsilon_{K;\nu}^\nu\;.
\end{aligned}$$

□

Proof of Lemma A.8. To compute the first bound, we rewrite the expression of interest as a quantity that we have already considered in the proof of Lemma A.7:

$$\begin{aligned}
(\mu^K)^\top\Lambda^K\Sigma^K\Lambda^K(\mu^K) \;&= (\mu^K)^\top\Lambda^K\mathbb{E}\Big[\big(\phi^K(\mathbf{X}_1)-\mu^K\big)\big(\phi^K(\mathbf{X}_1)-\mu^K\big)^\top\Big]\Lambda^K(\mu^K) \\
&= \mathbb{E}\Big[\Big(\big(\phi^K(\mathbf{X}_1)-\mu^K\big)^\top\Lambda^K\mu^K\Big)^2\Big] \\
&= \mathbb{E}\Big[\Big(\sum_{k=1}^K \lambda_k(\phi_k(\mathbf{X}_1)-\mu_k)\mu_k\Big)^2\Big] \\
&= \mathbb{E}\Big[\Big(\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{X}_2] - \mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)] + \Delta_{K;1} - \Delta_{K;2}\Big)^2\Big]\;,
\end{aligned}$$

where we have used the calculation in (A.19) with $\nu = 2$ and defined the same error terms

$$\Delta_{K;1} \;\coloneqq\; \sum_{k=1}^K \lambda_k\mathbb{E}[\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)|\mathbf{X}_2] - \mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{X}_2]\;,$$

$$\Delta_{K;2} \;:=\; \sum_{k=1}^{K} \lambda_k \mathbb{E}[\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)] - \mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)] \;.$$

Since we are dealing with the second moment, we can provide a finer bound by expanding the square explicitly:

$$\begin{aligned}(\mu^K)^\top \Lambda^K \Sigma^K \Lambda^K (\mu^K) \;&= \mathrm{Var}\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{X}_2] + \mathbb{E}[(\Delta_{K;1}-\Delta_{K;2})^2] \\ &\quad + 2\mathbb{E}\big[\big(\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{X}_2] - \mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)]\big)(\Delta_{K;1}-\Delta_{K;2})^2\big] \;.\end{aligned}$$

Then by the Cauchy-Schwartz inequality, we get that

$$\begin{aligned}&\Big|(\mu^K)^\top \Lambda^K \Sigma^K \Lambda^K (\mu^K) - \mathrm{Var}\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{X}_2]\Big| \\ &= \; 2\Big|\mathbb{E}\big[\big(\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{X}_2] - \mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)]\big)(\Delta_{K;1}-\Delta_{K;2})^2\big]\Big| + \mathbb{E}[(\Delta_{K;1}-\Delta_{K;2})^2] \\ &\leq \; 2\sqrt{\mathrm{Var}\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{X}_2]}\sqrt{\mathbb{E}[(\Delta_{K;1}-\Delta_{K;2})^2]} + \mathbb{E}[(\Delta_{K;1}-\Delta_{K;2})^2] \;.\end{aligned}$$

The variance term is exactly σ^2_{cond}. Since the individual error terms have already been bounded in the proof of Lemma A.7 as $\mathbb{E}[\Delta^2_{K;1}], \mathbb{E}[\Delta^2_{K;2}] \leq \varepsilon^2_{K;2}$, by the triangle inequality and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}|\mathbb{E}[(\Delta_{K;1}-\Delta_{K;2})^2]| \;&= |\mathbb{E}[\Delta^2_{K;1}] - 2\mathbb{E}[\Delta_{K;1}\Delta_{K;2}] + \mathbb{E}[\Delta^2_{K;2}]| \\ &\leq |\mathbb{E}[\Delta^2_{K;1}]| + 2\sqrt{|\mathbb{E}[\Delta^2_{K;1}]||\mathbb{E}[\Delta^2_{K;2}]|} + |\mathbb{E}[\Delta^2_{K;2}]| \;\leq\; 4\varepsilon^2_{K;2} \;.\end{aligned}$$

Combining the bounds gives

$$\Big|(\mu^K)^\top \Lambda^K \Sigma^K \Lambda^K (\mu^K) - (\sigma_{\mathrm{cond}})^2\Big| \;\leq\; 4\varepsilon^2_{K;2} + 4\sigma_{\mathrm{cond}}\varepsilon_{K;2} \;,$$

which rearranges to give

$$\begin{aligned}\sigma^2_{\mathrm{cond}} - 4\sigma_{\mathrm{cond}}\varepsilon_{K;2} - 4\varepsilon^2_{K;2} \;\leq\; (\mu^K)^\top \Lambda^K \Sigma^K \Lambda^K (\mu^K) \;&\leq \sigma^2_{\mathrm{cond}} + 4\sigma_{\mathrm{cond}}\varepsilon_{K;2} + 4\varepsilon^2_{K;2} \\ &\leq (\sigma_{\mathrm{cond}} + 2\varepsilon_{K;2})^2 \;.\end{aligned}$$

The second bound is obtained similarly by giving a finer control than the bound in Lemma A.7. We first rewrite the expression of interest by using linearity of expectation and the cyclic property of trace:

$$\begin{aligned}\mathrm{Tr}((\Lambda^K\Sigma^K)^2) \;&= \mathrm{Tr}\big(\Lambda^K \mathbb{E}\big[\phi^K(\mathbf{X}_1)\phi^K(\mathbf{X}_1)^\top\big]\Lambda^K \mathbb{E}\big[\phi^K(\mathbf{X}_2)\phi^K(\mathbf{X}_2)^\top\big]\big) \\ &= \mathbb{E}\Big[\big(\phi^K(\mathbf{X}_1)^\top \Lambda^K \phi^K(\mathbf{X}_2)\big)^2\Big] \;=\; \mathbb{E}\Big[\Big(\sum_{k=1}^K \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\Big)^2\Big] \;.\end{aligned}$$

Again by expanding the square explicitly, we get that

$$\begin{aligned}\mathrm{Tr}((\Lambda^K\Sigma^K)^2) \;&= \mathbb{E}\Big[\Big(\sum_{k=1}^K \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2) - \bar{u}(\mathbf{X}_1,\mathbf{X}_2) + \bar{u}(\mathbf{X}_1,\mathbf{X}_2)\Big)^2\Big] \\ &= \mathbb{E}\Big[\Big(\sum_{k=1}^K \lambda_k \phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2) - \bar{u}(\mathbf{X}_1,\mathbf{X}_2)\Big)^2\Big] + \mathbb{E}\big[\bar{u}(\mathbf{X}_1,\mathbf{X}_2)^2\big] \\ &\quad + 2\Delta_{K;3} \\ &= \varepsilon^2_{K;2} + \sigma^2_{\mathrm{full}} + 2\Delta_{K;3} \;,\end{aligned}$$

where we have defined the additional error term as

$$\Delta_{K;3} \;\coloneqq\; \mathbb{E}\Big[\Big(\sum_{k=1}^{K}\lambda_k\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2) - \bar{u}(\mathbf{X}_1,\mathbf{X}_2)\Big)\bar{u}(\mathbf{X}_1,\mathbf{X}_2)\Big]\;.$$

By the Cauchy-Schwarz inequality, we get that

$$\begin{aligned}\big|\mathrm{Tr}((\Lambda^K\Sigma^K)^2) - \sigma^2_{\mathrm{full}} - \varepsilon^2_{K;2}\big| \;&=\; 2|\Delta_{K;3}| \\ \leq 2\sqrt{\mathbb{E}\big[\big(\textstyle\sum_{k=1}^{K}\lambda_k\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2) - \bar{u}(\mathbf{X}_1,\mathbf{X}_2)\big)^2\big]}\sqrt{\mathbb{E}\big[\bar{u}(\mathbf{X}_1,\mathbf{X}_2)^2\big]} \;&=\; 2\varepsilon_{K;2}\sigma_{\mathrm{full}}\;.\end{aligned}$$

Combining the above two bounds yields the desired inequality that

$$(\sigma_{\mathrm{full}} - \varepsilon_{K;2})^2 \;\leq\; \mathrm{Tr}((\Lambda^K\Sigma^K)^2) \;\leq\; (\sigma_{\mathrm{full}} + \varepsilon_{K;2})^2\;.$$

To prove the third bound, note that $(\mu^K)^\top\Lambda^K\mathbf{Z}_1$ is a zero-mean normal random variable with variance given by $(\mu^K)^\top\Lambda^K\Sigma^K\mu^K$, which is already bounded above. By applying the formula of the ν-th absolute moment of a normal distribution and noting that $\nu \leq 3$, we obtain

$$\begin{aligned}\mathbb{E}[|(\mu^K)^\top\Lambda^K\mathbf{Z}_1|^\nu] \;&=\; \frac{2^{\nu/2}}{\sqrt{\pi}}\Gamma\Big(\frac{\nu+1}{2}\Big)\big((\mu^K)^\top\Lambda^K\Sigma^K\Lambda^K\mu^K\big)^{\nu/2} \\ \leq\; \frac{2^{\nu/2}}{\sqrt{\pi}}(\sigma_{\mathrm{full}} + 2\varepsilon_{K;2})^\nu \;&\overset{(a)}{\leq}\; \frac{2^{\nu/2}}{\sqrt{\pi}}\max\{1, 2^{\nu-1}\}\big(\sigma^\nu_{\mathrm{cond}} + 2^\nu\varepsilon^\nu_{K;2}\big) \;\overset{(b)}{\leq}\; 7(\sigma^\nu_{\mathrm{cond}} + 8\varepsilon^\nu_{K;2})\;.\end{aligned}$$

In (a), we have noted that given $a, b > 0$, for $\nu/2 \in (0, 1]$, $(a+b)^{\nu/2} \leq a^{\nu/2} + b^{\nu/2}$ and for $\nu/2 > 1$, the bound follows from Jensen's inequality. In (b), we have noted that $\nu \leq 3$. This finishes the proof for the third bound.

To prove the fourth bound, we can first condition on $\mathbf{Z}_2$:

$$\mathbb{E}[|\mathbf{Z}_1^\top\Lambda^K\mathbf{Z}_2|^\nu] \;=\; \mathbb{E}\big[\,\mathbb{E}[|\mathbf{Z}_1^\top\Lambda^K\mathbf{Z}_2|^\nu\,|\,\mathbf{Z}_2]\,\big]\;.$$

The inner expectation is again the ν-th absolute moment of a conditionally Gaussian random variable with variance $\mathbf{Z}_2^\top\Lambda^K\Sigma^K\Lambda^K\mathbf{Z}_2$, so again by the formula of the ν-th absolute moment of a normal distribution, we get that

$$\mathbb{E}[|\mathbf{Z}_1^\top\Lambda^K\mathbf{Z}_2|^\nu] \;\leq\; \frac{2^{\nu/2}}{\sqrt{\pi}}\,\mathbb{E}\Big[\big(\mathbf{Z}_2^\top\Lambda^K\Sigma^K\Lambda^K\mathbf{Z}_2\big)^{\nu/2}\Big] \;\leq\; \frac{2^{\nu/2}}{\sqrt{\pi}}\,\mathbb{E}\Big[\big(\mathbf{Z}_2^\top\Lambda^K\Sigma^K\Lambda^K\mathbf{Z}_2\big)^2\Big]^{\nu/4}\;.$$

We have noted that $\nu \leq 3$ and used the Hölder's inequality. The remaining expectation is taken over a quadratic form of normal variables. Writing $\Sigma_* = (\Sigma^K)^{1/2}\Lambda^K(\Sigma^K)^{1/2}$ for short, the second moment can be computed by the formula from Lemma A.5 as

$$\mathbb{E}\Big[\big(\mathbf{Z}_2^\top\Lambda^K\Sigma^K\Lambda^K\mathbf{Z}_2\big)^2\Big] \;=\mathrm{Tr}(\Sigma_*^2)^2 + 2\mathrm{Tr}\big(\Sigma_*^4\big) \;\overset{(a)}{\leq}\; 3\mathrm{Tr}(\Sigma_*^2)^2 \;=\; 3\mathrm{Tr}\big((\Lambda^K\Sigma^K)^2\big)^2\;.$$

Note that in (a), we have used the fact that the square of a symmetric matrix, Σ_*^2, has non-negative eigenvalues, and therefore $\mathrm{Tr}(\Sigma_*^4) \leq \mathrm{Tr}(\Sigma_*^2)^2$. Since we have already bounded

$\mathrm{Tr}((\Lambda^K\Sigma^K)^2)$ earlier, substituting the above result into the previous bound, we get that

$$\begin{aligned}\mathbb{E}[|\mathbf{Z}_1^\top\Lambda^K\mathbf{Z}_2|^\nu] &\le \frac{2^{\nu/2}}{\sqrt{\pi}}\,\mathbb{E}\Big[\big(\mathbf{Z}_2^\top\Lambda^K\Sigma^K\Lambda^K\mathbf{Z}_2\big)^2\Big]^{\nu/4} \le \frac{2^{\nu/2}3^{\nu/4}}{\sqrt{\pi}}\,\mathrm{Tr}\big((\Lambda^K\Sigma^K)^2\big)^{\nu/2}\\ &\le \frac{2^{\nu/2}3^{\nu/4}}{\sqrt{\pi}}(\sigma_{\text{full}}^2+\varepsilon_{K;2}^2)^{\nu/2}\\ &\le \frac{2^{\nu/2}3^{\nu/4}}{\sqrt{\pi}}\max\{1,2^{\nu/2-1}\}\big(\sigma_{\text{full}}^\nu+\varepsilon_{K;2}^\nu\big) \;\le\; 6\big(\sigma_{\text{full}}^\nu+\varepsilon_{K;2}^\nu\big)\,.\end{aligned}$$

In the last two inequalities, we have used the same argument as in the proof for the third bound to expand the term with ν-th power. This gives the desired bound.

To prove the final bound, we first condition on $\mathbf{X}_1$:

$$\mathbb{E}\big[\big|(\phi^K(\mathbf{X}_1)-\mu^K)^\top\Lambda^K\mathbf{Z}_1\big|^\nu\big] \;=\; \mathbb{E}\big[\mathbb{E}\big[\big|(\phi^K(\mathbf{X}_1)-\mu^K)^\top\Lambda^K\mathbf{Z}_1\big|^\nu\big|\mathbf{X}_1\big]\big]\,.$$

The inner expectation is the ν-th absolute moment of a conditionally Gaussian random variable with variance $(\phi^K(\mathbf{X}_1)-\mu^K)^\top\Lambda^K\Sigma^K\Lambda^K(\phi^K(\mathbf{X}_1)-\mu^K)$, so by the formula of the ν-th absolute moment of a normal distribution with $\nu\le 3$, we get that

$$\begin{aligned}&\mathbb{E}[\,|(\phi^K(\mathbf{X}_1)-\mu^K)^\top\Lambda^K\mathbf{Z}_2|^\nu]\\ &\le \frac{2^{\nu/2}}{\sqrt{\pi}}\,\mathbb{E}\Big[\big((\phi^K(\mathbf{X}_1)-\mu^K)^\top\Lambda^K\Sigma^K\Lambda^K(\phi^K(\mathbf{X}_1)-\mu^K)\big)^{\nu/2}\Big]\\ &= \frac{2^{\nu/2}}{\sqrt{\pi}}\,\mathbb{E}\Big[\big((\phi^K(\mathbf{X}_1)-\mu^K)^\top\Lambda^K\mathbb{E}\big[(\phi^K(\mathbf{X}_2)-\mu^K)(\phi^K(\mathbf{X}_2)-\mu^K)^\top\big]\\ &\qquad\qquad \Lambda^K(\phi^K(\mathbf{X}_1)-\mu^K)\big)^{\nu/2}\Big]\\ &\overset{(a)}{\le} \frac{2^{\nu/2}}{\sqrt{\pi}}\,\mathbb{E}\Big[\big|(\phi^K(\mathbf{X}_1)-\mu^K)^\top\Lambda^K(\phi^K(\mathbf{X}_2)-\mu^K)\big|^\nu\Big]\\ &= \frac{2^{\nu/2}}{\sqrt{\pi}}\,\mathbb{E}\Big[\big|\textstyle\sum_{k=1}^K\lambda_k\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\big|^\nu\Big]\\ &\overset{(b)}{\le} 8C(M_{\text{full};\nu})^\nu-(M_{\text{cond};\nu})^\nu+(8C+4)\varepsilon_{K;\nu}^\nu\,.\end{aligned}$$

In (a), we have applied Jensen's inequality to the convex function $x\mapsto|x|^{\nu/2}$ to move the inner expectation outside the norm. In (b), we have applied the bound in Lemma A.7 and noted that $\frac{2^{\nu/2}}{\sqrt{\pi}}<2$ for $\nu\in[1,3]$. This gives the desired result. □

Proof of Lemma A.9. For the first equality in distribution, we recall that $\{\tau_{k;d}\}_{k=1}^K$ are the eigenvalues of $(\Sigma^K)^{1/2}\Lambda^K(\Sigma^K)^{1/2}$ and $\{\xi_k\}_{k=1}^K$ are a sequence of i.i.d. standard Gaussian variables. Let $\{\eta_{ik}\}_{i\in[n],k\in[K]}$ be a set of i.i.d. standard Gaussian variables. Since Gaussianity is preserved under orthogonal transformation, we have

$$\begin{aligned}&\frac{1}{n^{3/2}(n-1)^{1/2}}\Big(\textstyle\sum_{i,j=1}^n(\eta_i^K)^\top(\Sigma^K)^{1/2}\Lambda^K(\Sigma^K)^{1/2}\eta_j^K-n\mathrm{Tr}(\Sigma^K\Lambda^K)\Big)\\ \overset{d}{=}\;&\frac{1}{n^{3/2}(n-1)^{1/2}}\Big(\textstyle\sum_{k=1}^K\sum_{i,j=1}^n\tau_{k;d}\eta_{ik}\eta_{jk}-n\mathrm{Tr}((\Sigma^K)^{1/2}\Lambda^K(\Sigma^K)^{1/2})\Big)\end{aligned}$$

$$= \frac{1}{n^{3/2}(n-1)^{1/2}} \sum_{k=1}^{K} \tau_{k;d} \Big(\big(\sum_{i=1}^{n} \eta_{ik} \big) \big(\sum_{j=1}^{n} \eta_{jk} \big) - n \Big)$$
$$\stackrel{d}{=} \frac{1}{n^{1/2}(n-1)^{1/2}} \sum_{k=1}^{K} \tau_{k;d} (\xi_k^2 - 1) \;=\; W_n^K - D \;,$$

which proves the desired statement.

We now use the expression above for moment computation. The expectation is given by $\mathbb{E}[W_n^K] = D$ for every $K \in \mathbb{N}$. The variance can be computed by noting that the quantity is a quadratic form in Gaussian, applying Lemma A.5 and using the cyclic property of trace:

$$\begin{aligned} \mathrm{Var}[W_n^K] \;&= \frac{1}{n(n-1)} \mathrm{Var}\big[(\eta_1^K)^\top (\Sigma^K)^{1/2} \Lambda^K (\Sigma^K)^{1/2} \eta_1^K \big] \\ &= \frac{2}{n(n-1)} \mathrm{Tr}\big((\Lambda^K \Sigma^K)^2 \big) \;. \end{aligned}$$

By Lemma A.8, we get the desired bound that

$$\frac{2}{n(n-1)} (\sigma_{\text{full}} - \varepsilon_{K;2})^2 \;\le\; \mathrm{Var}[W_n^K] \;\le\; \frac{2}{n(n-1)} (\sigma_{\text{full}} + \varepsilon_{K;2})^2 \;.$$

The third central moment can be expanded using a binomial expansion and noting that each summand is zero-mean:

$$\begin{aligned} \mathbb{E}\big[(W_n^K - D)^3 \big] \;&= \frac{1}{n^{3/2}(n-1)^{3/2}} \mathbb{E}\Big[\Big(\sum_{k=1}^{K} \tau_{k;d} (\xi_k^2 - 1) \Big)^3 \Big] \\ &= \frac{1}{n^{3/2}(n-1)^{3/2}} \mathbb{E}\big[\sum_{k=1}^{K} \tau_{k;d}^3 (\xi_k^2 - 1)^3 \big] \\ &= \frac{8}{n^{3/2}(n-1)^{3/2}} \sum_{k=1}^{K} \tau_{k;d}^3 \;. \end{aligned}$$

Meanwhile, the sum can be further expressed as

$$\begin{aligned} &\sum_{k=1}^{K} \tau_{k;d}^3 \\ &= \mathrm{Tr}\Big(\big((\Sigma^K)^{1/2} \Lambda^K (\Sigma^K)^{1/2} \big)^3 \Big) \;=\; \mathrm{Tr}\Big(\big(\Sigma^K \Lambda^K \big)^3 \Big) \\ &= \mathrm{Tr}\Big(\big(\mathbb{E}\big[\phi^K(\mathbf{X}_1) (\phi^K(\mathbf{X}_1))^\top \big] \Lambda^K \big)^3 \Big) \\ &= \mathbb{E}\Big[(\phi^K(\mathbf{X}_1))^\top \Lambda^K \phi^K(\mathbf{X}_2) \, (\phi^K(\mathbf{X}_2))^\top \Lambda^K \phi^K(\mathbf{X}_3) \, (\phi^K(\mathbf{X}_3))^\top \Lambda^K \phi^K(\mathbf{X}_1) \Big] \\ &= \mathbb{E}\Big[\big(\sum_{k=1}^{K} \lambda_k \phi_k(\mathbf{X}_1) \phi_k(\mathbf{X}_2) \big) \big(\sum_{k=1}^{K} \lambda_k \phi_k(\mathbf{X}_2) \phi_k(\mathbf{X}_3) \big) \big(\sum_{k=1}^{K} \lambda_k \phi_k(\mathbf{X}_3) \phi_k(\mathbf{X}_1) \big) \Big] \\ &=: \; \mathbb{E}[S_{12} S_{23} S_{31}] \;. \end{aligned}$$

We now approximate each S_{ij} term by $u(\mathbf{X}_i, \mathbf{X}_j)$. For convenience, denote $U_{ij} = u(\mathbf{X}_i, \mathbf{X}_j)$ and $\Delta_{ij} = S_{ij} - U_{ij}$. Then

$$\begin{aligned} \sum_{k=1}^{K} \tau_{k;d}^3 \;&= \mathbb{E}\big[(U_{12} + \Delta_{12})(U_{23} + \Delta_{23})(U_{31} + \Delta_{31}) \big] \\ &= \mathbb{E}[U_{12} U_{23} U_{31}] + \mathbb{E}[U_{12} U_{23} \Delta_{31}] + \mathbb{E}[U_{12} \Delta_{23} U_{31}] + \mathbb{E}[U_{12} \Delta_{23} \Delta_{31}] \\ &\quad + \mathbb{E}[\Delta_{12} U_{23} U_{31}] + \mathbb{E}[\Delta_{12} U_{23} \Delta_{31}] + \mathbb{E}[\Delta_{12} \Delta_{23} U_{31}] + \mathbb{E}[\Delta_{12} \Delta_{23} \Delta_{31}] \;. \end{aligned}$$

Recall that $\varepsilon_{K;3} = \mathbb{E}[|\Delta_{ij}|^3]^{1/3}$ for $i \neq j$ by definition. Then by the triangle inequality followed by the Hölder's inequality, we get that

$$\begin{aligned}
&\Big|\sum_{k=1}^K \tau_{k;d}^3 - \mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_1)]\Big| \\
&\leq \big|\mathbb{E}[U_{12}U_{23}\Delta_{31}]\big| + \big|\mathbb{E}[U_{12}\Delta_{23}U_{31}]\big| + \big|\mathbb{E}[U_{12}\Delta_{23}\Delta_{31}]\big| \\
&\quad + \big|\mathbb{E}[\Delta_{12}U_{23}U_{31}]\big| + \big|\mathbb{E}[\Delta_{12}U_{23}\Delta_{31}]\big| + \big|\mathbb{E}[\Delta_{12}\Delta_{23}U_{31}]\big| + \big|\mathbb{E}[\Delta_{12}\Delta_{23}\Delta_{31}]\big| \\
&\leq 3\mathbb{E}[|u(\mathbf{X}_1,\mathbf{X}_2)|^3]^{2/3}\varepsilon_{K;3} + 3\mathbb{E}[|u(\mathbf{X}_1,\mathbf{X}_2)|^3]^{1/3}\varepsilon_{K;3}^2 + \varepsilon_{K;3}^3 \\
&= 3M_{\text{full};3}^2\varepsilon_{K;3} + 3M_{\text{full};3}\varepsilon_{K;3}^2 + \varepsilon_{K;3}^3 \,.
\end{aligned}$$

This implies that

$$\begin{aligned}
\textstyle\sum_{k=1}^K \tau_{k;d}^3 &\leq \mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_1)] - M_{\text{full};3}^3 + (M_{\text{full};3} + \varepsilon_{K;3})^3 \,, \\
\textstyle\sum_{k=1}^K \tau_{k;d}^3 &\geq \mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_1)] + M_{\text{full};3}^3 - (M_{\text{full};3} + \varepsilon_{K;3})^3 \,,
\end{aligned}$$

which gives the desired bounds:

$$\begin{aligned}
\mathbb{E}\big[(W_n^K - D)^3\big] &\leq \frac{8\big(\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_1)] - M_{\text{full};3}^3 + (M_{\text{full};3} + \varepsilon_{K;3})^3\big)}{n^{3/2}(n-1)^{3/2}} \,, \\
\mathbb{E}\big[(W_n^K - D)^3\big] &\geq \frac{8\big(\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_1)] + M_{\text{full};3}^3 - (M_{\text{full};3} + \varepsilon_{K;3})^3\big)}{n^{3/2}(n-1)^{3/2}} \,.
\end{aligned}$$

The fourth central moment can again be expanded using a binomial expansion and noting that each summand is zero-mean:

$$\begin{aligned}
&\mathbb{E}\big[(W_n^K - D)^4\big] \\
&= \frac{1}{n^2(n-1)^2}\mathbb{E}\Big[\Big(\sum_{k=1}^K \tau_{k;d}(\xi_k^2 - 1)\Big)^4\Big] \\
&= \frac{1}{n^2(n-1)^2}\Big(\mathbb{E}\Big[\sum_{k=1}^K \tau_{k;d}^4(\xi_k^2-1)^4\Big] + 3\mathbb{E}\Big[\sum\nolimits_{1\leq k\neq k'\leq K} \tau_{k;d}^2(\xi_k^2-1)^2\tau_{k';d}^2(\xi_{k'}^2-1)^2\Big]\Big) \\
&= \frac{1}{n^2(n-1)^2}\Big(60\sum_{k=1}^K \tau_{k;d}^4 + 12\sum\nolimits_{1\leq k\neq k'\leq K} \tau_{k;d}^2\tau_{k';d}^2\Big) \\
&= \frac{1}{n^2(n-1)^2}\Big(48\sum_{k=1}^K \tau_{k;d}^4 + 12\sum\nolimits_{1\leq k,k'\leq K} \tau_{k;d}^2\tau_{k';d}^2\Big) \\
&= \frac{12}{n^2(n-1)^2}\Big(4\sum_{k=1}^K \tau_{k;d}^4 + \big(\sum_{k=1}^K \tau_{k;d}^2\big)^2\Big) \,. \qquad \text{(A.20)}
\end{aligned}$$

Since we have already controlled $\sum_{k=1}^K \tau_{k;d}^2 = \mathrm{Tr}\big((\Sigma^K\Lambda^K)^2\big)$, we focus on bounding the first sum. Using notations from the third moment, we can express the sum as

$$\begin{aligned}
\textstyle\sum_{k=1}^K \tau_{k;d}^4 &= \mathbb{E}\Big[\big(\sum_{k=1}^K \lambda_k\phi_k(\mathbf{X}_1)\phi_k(\mathbf{X}_2)\big)\big(\sum_{k=1}^K \lambda_k\phi_k(\mathbf{X}_2)\phi_k(\mathbf{X}_3)\big) \\
&\qquad \big(\sum_{k=1}^K \lambda_k\phi_k(\mathbf{X}_3)\phi_k(\mathbf{X}_4)\big)\big(\sum_{k=1}^K \lambda_k\phi_k(\mathbf{X}_4)\phi_k(\mathbf{X}_1)\big)\Big] \\
&= \mathbb{E}[S_{12}S_{23}S_{34}S_{41}] \\
&= \mathbb{E}\big[(U_{12}+\Delta_{12})(U_{23}+\Delta_{23})(U_{34}+\Delta_{34})(U_{41}+\Delta_{41})\big] \,.
\end{aligned}$$

A similar argument as before shows that

$$\Big|\sum_{k=1}^K \tau_{k;d}^4 - \mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_4)u(\mathbf{X}_4,\mathbf{X}_1)]\Big| \\ \leq 4M_{\text{full};4}^3\varepsilon_{K;4} + 6M_{\text{full};4}^2\varepsilon_{K;4}^2 + 4M_{\text{full};4}\varepsilon_{K;4}^3 + \varepsilon_{K;4}^4 \,.$$

This implies that

$$\sum_{k=1}^K \tau_{k;d}^4 \leq \mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_4)u(\mathbf{X}_4,\mathbf{X}_1)] - M_{\text{full};4}^4 + (M_{\text{full};4}+\varepsilon_{K;4})^4 \,,$$
$$\sum_{k=1}^K \tau_{k;d}^4 \geq \mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_4)u(\mathbf{X}_4,\mathbf{X}_1)] + M_{\text{full};4}^4 - (M_{\text{full};4}+\varepsilon_{K;4})^4 \,.$$

On the other hand, by Lemma A.8, we have

$$(\sigma_{\text{full}} - \varepsilon_{K;2})^2 \leq \sum_{k=1}^K \tau_{k;d}^2 = \text{Tr}((\Lambda^K\Sigma^K)^2) \leq (\sigma_{\text{full}} + \varepsilon_{K;2})^2 \,.$$

Plugging these calculations into (A.20) give the desired bounds:

$$\mathbb{E}\big[(W_n^K - D)^4\big] \leq \frac{12}{n^2(n-1)^2}\Big(4\,\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_4)u(\mathbf{X}_4,\mathbf{X}_1)] \\ - 4M_{\text{full};4}^4 + 4(M_{\text{full};4}+\varepsilon_{K;4})^4 + (\sigma_{\text{full}}+\varepsilon_{K;2})^4\Big) \,,$$
$$\mathbb{E}\big[(W_n^K - D)^4\big] \geq \frac{12}{n^2(n-1)^2}\Big(4\,\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_4)u(\mathbf{X}_4,\mathbf{X}_1)] \\ + 4M_{\text{full};4}^4 - 4(M_{\text{full};4}+\varepsilon_{K;4})^4 + (\sigma_{\text{full}}-\varepsilon_{K;2})^4\Big) \,.$$

For the generic moment bound, we first use the Jensen's inequality to get that

$$\mathbb{E}\big[(W_n^K)^{2m}\big] = \mathbb{E}\Big[\Big(\frac{1}{n^{1/2}(n-1)^{1/2}}\sum_{k=1}^K \tau_{k;d}(\xi_k^2-1) + D\Big)^{2m}\Big] \\ \leq \frac{2^{2m-1}}{n^m(n-1)^m}\mathbb{E}\Big[\Big(\sum_{k=1}^K \tau_{k;d}(\xi_k^2-1)\Big)^{2m}\Big] + 2^{2m-1}\,D^{2m} \,.$$

Denote the set of all possible orderings of a length-$2m$ sequence consisting of elements from $[K]$ by $\mathcal{P}(K,2m)$ and denote its elements by p. Consider the subset

$$\mathcal{P}'(K,2m) \coloneqq \{p \in \mathcal{P}(K,2m) : \text{ every element in } p \text{ appears at least twice }\} \,.$$

By noting that $\xi_k - 1$ is zero-mean and $\{\xi_k\}_{k=1}^K$ are independent, we can re-express the sum first as a sum over $\mathcal{P}(K,2m)$ and then as a sum over $\mathcal{P}'(K,2m)$:

$$\mathbb{E}\Big[\Big(\sum_{k=1}^K \tau_{k;d}\,(\xi_k^2-1)\Big)^{2m}\Big] = \sum_{p\in\mathcal{P}(K,2m)}\big(\prod_{k\in p}\tau_{k;d}\big)\mathbb{E}\big[\prod_{k\in p}(\xi_k^2-1)\big] \\ = \sum_{p\in\mathcal{P}'(K,2m)}\big(\prod_{k\in p}\tau_{k;d}\big)\mathbb{E}\big[\prod_{k\in p}(\xi_k^2-1)\big] \\ + \sum_{p\in\left(\mathcal{P}(K,2m)\setminus\mathcal{P}'(K,2m)\right)}\big(\prod_{k\in p}\tau_{k;d}\big)\mathbb{E}\big[\prod_{k\in p}(\xi_k^2-1)\big] \\ = \sum_{p\in\mathcal{P}'(K,2m)}\big(\prod_{k\in p}\tau_{k;d}\big)\mathbb{E}\big[\prod_{k\in p}(\xi_k^2-1)\big] \,.$$

Write C_m' as the $2m$-th central moment of a chi-squared random variable with degree 1, which depends only on m and not on K or $\tau_{k;d}$. By the Hölder's inequality and the bound

from Lemma A.8, we get that

$$\begin{aligned}\mathbb{E}\Big[\Big(\sum_{k=1}^K \tau_{k;d}\,(\xi_k^2-1)\Big)^{2m}\Big] \;&\le\; C'_m \sum_{p\in\mathcal{P}'(K,2m)} \Big(\prod_{k\in p}\tau_{k;d}\Big)\\ &\le C'_m \binom{2m}{m}\Big(\sum_{k=1}^K \tau_{k;d}^2\Big)^m\\ &= C'_m \binom{2m}{m}\operatorname{Tr}\big((\Lambda^K\Sigma^K)^2\big)^m \;\le\; C'_m\binom{2m}{m}\big(\sigma_{\text{full}}+\varepsilon_{K;2}\big)^{2m}\,.\end{aligned}$$

Writing $C_m := 2^{2m-1}\max\{1, C'_m\binom{2m}{m}\}$, we get the desired bound that

$$\mathbb{E}\big[(W_n^K)^{2m}\big] \;\le\; \frac{C_m}{n^m(n-1)^m}\big(\sigma_{\text{full}}+\varepsilon_{K;2}\big)^{2m} + C_m\,D^{2m}\,.$$

Finally, if Assumption 3.2 is true for some $\nu \ge 2$, we have $\varepsilon_{K;2}\to 0$ as K grows. Taking $K\to\infty$ in the bound for second moment gives

$$\lim_{K\to\infty}\operatorname{Var}[W_n^K] \;=\; \frac{2}{n(n-1)}\sigma_{\text{full}}^2\,.$$

If Assumption 3.2 holds for $\nu\ge 3$, similarly we have

$$\lim_{K\to\infty}\mathbb{E}\big[(W_n^K-D)^3\big] \;=\; \frac{8\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_1)]}{n^{3/2}(n-1)^{3/2}}\,.$$

If Assumption 3.2 holds for $\nu\ge 4$, we have

$$\lim_{K\to\infty}\mathbb{E}\big[(W_n^K-D)^4\big] \;=\; \frac{12(4\mathbb{E}[u(\mathbf{X}_1,\mathbf{X}_2)u(\mathbf{X}_2,\mathbf{X}_3)u(\mathbf{X}_3,\mathbf{X}_4)u(\mathbf{X}_4,\mathbf{X}_1)]+\sigma_{\text{full}}^4)}{n^2(n-1)^2}\,.$$

□

A.7.3. Proofs for Appendix A.2.3

Proof of Lemma A.10. Write $\delta' := \delta/(m+1)$ for convenience. Define the m-times differentiable function

$$h_{m;\tau;\delta}(x) \;:=\; (\delta')^{-(m+1)}\int_x^{x+\delta'}\int_{y_1}^{y_1+\delta'}\cdots\int_{y_{m-1}}^{y_{m-1}+\delta'}\int_{y_m}^{y_m+\delta'}\mathbb{I}_{\{y>\tau\}}\,dy\,dy_m\ldots dy_1\,.$$

In the case $m=0$, the function is $h_{0;\tau;\delta}(x) \;:=\; \delta^{-1}\int_x^{x+\delta}\mathbb{I}_{\{y>\tau\}}\,dy$. By construction, $h_{m;\tau;\delta}(x)=0$ for $x\le\tau-\delta$, $h_{m;\tau;\delta}(x)\in[0,1]$ for $x\in(\tau-\delta,\tau]$ and $h_{m;\tau;\delta}(x)=1$ for $x>\tau$. This implies $\mathbb{I}_{\{x>\tau\}}\le h_{m;\tau;\delta}(x)\le\mathbb{I}_{\{x>\tau-\delta\}}$ and therefore the desired inequality

$$h_{m;\tau+\delta;\delta}(x) \;\le\; \mathbb{I}_{\{x>\tau\}} \;\le\; h_{m;\tau;\delta}(x)\,.$$

Next, we prove the properties of the derivatives of $h_{m;\tau;\delta}$. Denote recursively

$$J_{m+1}(x) \;:=\; \int_x^{x+\delta'}\mathbb{I}_{\{y>\tau\}}dy\,,\qquad J_r(x) \;:=\; \int_x^{x+\delta'}J_{r+1}(y)\,dy \quad\text{for}\;\; 0\le r\le m\,.$$

Since $h_{m;\tau;\delta}(x) = (\delta')^{-(m+1)} J_0(x)$ and $\frac{\partial}{\partial x} J_i(x) = J_{i+1}(x+\delta') - J_{i+1}(x)$ for $0 \le i \le m$, by induction, we have that for $0 \le r \le m$,

$$h^{(r)}_{m;\tau;\delta}(x) \;= (\delta')^{-(m+1)} \frac{\partial^r}{\partial x^r} J_0(x) \;=\; (\delta')^{-(m+1)} \sum_{i=0}^{r} \binom{r}{i} (-1)^i \, J_{r+1}\big(x + (r-i)\delta'\big) \,. \tag{A.21}$$

Note that J_{m+1} is continuous, uniformly bounded above by δ', and satisfies that $J_{m+1}(x) = 0$ for x outside $[\tau - \delta', \tau]$. By induction, we get that for $0 \le r \le m$, J_{r+1} is continuous, bounded above by $(\delta')^{m+1-r}$ and satisfies that $J_{r+1}(x) = 0$ for x outside $[\tau - (m+1-r)\delta', \tau]$. This shows that $h^{(r)}_{m;\tau;\delta}$ is continuous and $h^{(r)}_{m;\tau;\delta}(x) = 0$ for x outside $[\tau - \delta, \tau]$, and the uniform bound

$$\big|h^{(r)}_{m;\tau;\delta}(x)\big| \;\le\; (\delta')^{-r} \sum_{i=0}^{r} \binom{r}{i} \;=\; \big(\frac{2}{m+1}\big)^r \delta^{-r} \;\le\; \delta^{-r} \,.$$

Finally to prove the Hölder property of $h^{(m)}_{m;\tau;\delta}(x)$, we first note that J_{m+1} is constant outside $x \in [\tau - \delta', \tau]$ and linear within the interval with Lipschitz constant 1. The formula in (A.21) suggests that $h^{(m)}_{m;\tau;\delta}(x)$ is piecewise linear and the Lipschitz constant in the interval $[\tau - (m-i+1)\delta', \tau - (m-i)\delta']$ is given by the Lipschitz constant of the i-th summand. Therefore, $h^{(m)}_{m;\tau;\delta}$ is also Lipschitz with Lipschitz constant

$$L_m \;\coloneqq\; (\delta')^{-(m+1)} \max\nolimits_{0 \le i \le m} \binom{m}{i} \;=\; (\delta')^{-(m+1)} \binom{m}{\lfloor m/2 \rfloor} \,.$$

For $x, y \in [\tau - \delta, \tau]$, we then have

$$\begin{aligned} |h^{(m)}_{m;\tau;\delta}(x) - h^{(m)}_{m;\tau;\delta}(y)| \;\le\; L_m |x-y| \;&= L_m \delta \, \big|\tfrac{x-y}{\delta}\big| \\ &\le L_m \delta \, \big|\tfrac{x-y}{\delta}\big|^{\epsilon} \;=\; L_m \delta^{1-\epsilon} |x-y|^{\epsilon} \,, \end{aligned} \tag{A.22}$$

where we have noted that $\big|\frac{x-y}{\delta}\big| \le 1$ and $\epsilon \in [0,1]$. (A.22) is trivially true for x, y both outside $[\tau - \delta, \tau]$ since $h^{(m)}_{m;\tau;\delta}$ evaluates to zero. Now consider $x \in [\tau - \delta, \tau]$ and $y < \tau - \delta$. We have that

$$\begin{aligned} |h^{(m)}_{m;\tau;\delta}(x) - h^{(m)}_{m;\tau;\delta}(y)| \;=\; |h^{(m)}_{m;\tau;\delta}(x) - h^{(m)}_{m;\tau;\delta}(\tau-\delta)| \;&\overset{\text{(A.22)}}{\le}\; L_m \delta^{1-\epsilon} (x - \tau + \delta)^{\epsilon} \\ &\le L_m \delta^{1-\epsilon} |x-y|^{\epsilon} \,. \end{aligned}$$

Similarly for $x \in [\tau - \delta, \tau]$ and $y > \tau$, we have that

$$\begin{aligned} |h^{(m)}_{m;\tau;\delta}(x) - h^{(m)}_{m;\tau;\delta}(y)| \;&= |h^{(m)}_{m;\tau;\delta}(x) - h^{(m)}_{m;\tau;\delta}(\tau)| \;\overset{\text{(A.22)}}{\le}\; L_m \delta^{1-\epsilon} (\tau - x)^{\epsilon} \\ &\le L_m \delta^{1-\epsilon} |x-y|^{\epsilon} \,. \end{aligned}$$

Therefore (A.22) holds for all x, y. The proof for the derivative bound is complete by computing the constant explicitly as

$$L_m \delta^{1-\epsilon} \;=\; (\delta')^{-(m+\epsilon)} \binom{m}{\lfloor m/2 \rfloor} \;=\; \delta^{-(m+\epsilon)} \binom{m}{\lfloor m/2 \rfloor} (m+1)^{m+\epsilon} \,,$$

and therefore

$$|h^{(m)}_{m;\tau;\delta}(x) - h^{(m)}_{m;\tau;\delta}(y)| \;\leq\; C_{m,\epsilon}\, \delta^{-(m+\epsilon)}\, |x-y|^{\epsilon}\,, \tag{A.23}$$

with respect to the constant $C_{m,\epsilon} = \binom{m}{\lfloor m/2 \rfloor}(m+1)^{m+\epsilon}$. □

Proof of Lemma A.11. By conditioning on the size of Y, we have that for any $a, b \in \mathbb{R}$ and $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}(a \leq X+Y \leq b) \;&= \mathbb{P}(a \leq X+Y \leq b\,,\, |Y| \leq \epsilon) + \mathbb{P}(a \leq X \leq b\,,\, |Y| \geq \epsilon) \\ &\leq \mathbb{P}(a-\epsilon \leq X \leq b+\epsilon) + \mathbb{P}(|Y| \geq \epsilon)\,,\end{aligned}$$

and by using the order of inclusion of events, we have the lower bound

$$\begin{aligned}\mathbb{P}(a \leq X+Y \leq b) \;&\geq \mathbb{P}(a+\epsilon \leq X \leq b-\epsilon\,,\, |Y| \leq \epsilon) \\ &= \mathbb{P}(a+\epsilon \leq X \leq b-\epsilon) - \mathbb{P}(|Y| \geq \epsilon)\,.\end{aligned}$$

□

A.7.4. Proof for Appendix A.2.4

Proof of Lemma A.13. By Lemma 2.3 of Steinwart and Scovel (2012), the assumption that κ^* is measurable and $\mathbb{E}[\kappa^*(\mathbf{V}_1, \mathbf{V}_1)] < \infty$ implies the RKHS $\mathcal{H}$ associated with κ^* is compactly embedded into $L_2(\mathbb{R}^d, R)$. By Lemma 2.12 and Corollary 3.2 of Steinwart and Scovel (2012), for some index set $\mathcal{I} \subseteq \mathbb{N}$, there exists a sequence of non-negative, bounded values $\{\lambda_k\}_{k\in\mathcal{I}}$ that converges to 0 and a sequence of functions $\{\phi_k\}_{k\in\mathcal{I}}$ that form an orthonormal basis of $L_2(\mathbb{R}^d, R)$ such that

$$\textstyle\sum_{k\in\mathcal{I}} \lambda_k \psi_k(\mathbf{V}_1)\psi_k(\mathbf{V}_2) \;=\; \kappa^*(\mathbf{V}_1, \mathbf{V}_2)\,,$$

where the equality holds almost surely when $\mathcal{I}$ is finite and the convergence holds almost surely when $\mathcal{I}$ is infinite. We can extend $\mathcal{I}$ to $\mathbb{N}$ by adding zero values of λ_k and ϕ_k whenever necessary and drop the requirement that $\{\phi_k\}_{k=1}^{\infty}$ forms a basis, which gives the desired statement. □

Appendix B

Proofs for Section 3.3 and Section 4.3

This appendix concerns the proof of the tight upper and lower bounds for degree-two U- and V-statistics. Key ingredients of the proof are the degree-m polynomial construction used for proving the lower bound in Section 4.5 and the variance domination result of Section 4.3, which is also proved in Appendix B.4.

B.1 Matching upper and lower bounds for degree-two V-statistics

The next result states that the slow rate of approximation in Theorem 3.8 also holds for some V-statistic $v_n(X) = \frac{1}{n(n-1)} \sum_{i \neq j} k_v(X_i, X_j)$.

Theorem B.1. *Under the same setup as Theorem 3.8, there exists some $k_v : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ that depends on σ_n such that*

$$cn^{-\frac{\nu-2}{4\nu}} \;\leq\; \sup_{t \in \mathbb{R}} \Big| \mathbb{P}\big(n\, v_n(X) \leq t\big) - \mathbb{P}\big(\, \sigma_n \xi + \chi_1^2 \leq t\,\big) \Big| \;\leq\; C n^{-\frac{\nu-2}{4\nu}} \,.$$

B.2 Construction of k_u, k_v and X

We first recall the lower bound construction from Section 4.5 in the case $m = 2$. The construction involves a polynomial $p_n^* : (\mathbb{R}^2)^n \to \mathbb{R}$ given by

$$p_n^*(y_1, \ldots, y_n) \;\coloneqq\; \frac{1}{\sqrt{n}} \sum_{i=1}^n y_{i1} + \Big(\frac{1}{\sqrt{n}} \sum_{i=1}^n y_{i2} \Big)^2 \qquad \text{for } y_i = (y_{i1}, y_{i2}) \in \mathbb{R}^2 \,.$$

The collection of $\mathbb{R}^2$ random vectors $Y = (Y_1, \ldots, Y_n)$ are generated as follows. For a fixed $\sigma_0 > 0$ and $\nu \in (2, 3]$, write $\sigma_n = \min\{\sigma_0\, n^{-(\nu-2)/2\nu}\,,\, 1\}$. Note that this is the sequence $\sigma_n \to 0$ in Theorem 3.8. Let U_{σ_n} be the discrete random variable supported at three points with

$$U_{\sigma_n} \;=\; \begin{cases} -6^{-1/2} \sigma_n^{-2/(\nu-2)} & \text{with probability } 2\sigma_n^{2\nu/(\nu-2)} \,, \\ 0 & \text{with probability } 1 - 3\sigma_n^{2\nu/(\nu-2)} \,, \\ 2 \times 6^{-1/2} \sigma_n^{-2/(\nu-2)} & \text{with probability } \sigma_n^{2\nu/(\nu-2)} \,. \end{cases}$$

U_{σ_n} is constructed such that it becomes increasingly heavy-tailed as $\sigma_n \to 0$. Now let $U_{1;\sigma_n}, \ldots, U_{n;\sigma_n}$ be i.i.d. copies of U_{σ_n}. The i.i.d. random vectors $Y = (Y_i)_{i \le n}$ are generated by

$$Y_{i;\sigma_n} \coloneqq \left(\frac{1}{\sqrt{2}} U_{i;\sigma_n} + \frac{\sigma_n}{\sqrt{2}} \xi_{i1} \, , \, \xi_{i2}\right) \qquad \text{where} \quad \xi_{11}, \xi_{12}, \ldots, \xi_{n1}, \xi_{n2} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \, .$$

To adapt $p_n^*(Y)$ to our setup, we first observe that p_n^* can be rewritten as a V-statistic:

$$p_n^*(y_1, \ldots, y_n) \; = \frac{1}{n} \sum\nolimits_{i,j=1}^{n} \left(\frac{y_{i1}}{2\sqrt{n}} + \frac{y_{j1}}{2\sqrt{n}} + y_{i2} \, y_{j2} \right) \; = \; n \, \tilde{v}_n(y_1, \ldots, y_n) \, ,$$

where we have defined, for $y_1, \ldots, y_n \in \mathbb{R}^2$ and $a_1, a_2, b_1, b_2 \in \mathbb{R}$,

$$\tilde{v}_n(y_1, \ldots, y_n) \coloneqq \frac{1}{n^2} \sum\nolimits_{i,j=1}^{n} \tilde{k}_v(y_i, y_j) \, , \qquad \tilde{k}_v((a_1, a_2), (b_1, b_2)) \coloneqq \frac{a_1}{2\sqrt{n}} + \frac{b_1}{2\sqrt{n}} + a_2 b_2 \, .$$

Moreover, since we have no restrictions on how $d(n)$ depends on n, there are non-unique choices of a function $\phi_{d(n)} : \mathbb{R}^{d(n)} \to \mathbb{R}$ and a probability measure $\mu_{d(n)}$ on $\mathbb{R}^{d(n)}$ such that

$$X_1 \; \sim \mu_{d(n)} \qquad \Leftrightarrow \qquad \phi_{d(n)}(X_1) \; \overset{d}{=} Y_{1;\sigma_n} \, .$$

Our construction of the V-statistic is thus given by taking $X_i \overset{\text{i.i.d.}}{\sim} \mu_{d(n)}$ and $k_v(x_1, x_2) \coloneqq \tilde{k}_v(\phi_{d(n)}(x_1), \phi_{d(n)}(x_2))$, which gives

$$v_n(X) \; = \; \frac{1}{n^2} \sum\nolimits_{1 \le i,j \le n} k_v(X_i, X_j) \; \overset{d}{=} \tilde{v}_n(Y) \; = \; \frac{1}{n} \, p_n^*(Y) \, ,$$

where $\overset{d}{=}$ denotes equality in distribution. This makes the lower bound from Theorem 4.7 immediately applicable. For the U-statistics construction, we observe that

$$\begin{aligned} p_n^*(y_1, \ldots, y_n) \; &= \frac{1}{\sqrt{n(n-1)}} \sum\nolimits_{i \ne j}^{n} \left(\frac{y_{i1}}{2\sqrt{n-1}} + \frac{y_{i2}}{2\sqrt{n-1}} + \frac{\sqrt{n-1}}{\sqrt{n}} \, y_{i2} y_{j2} \right) + \frac{1}{n} \sum\nolimits_{i=1}^{n} y_{i2}^2 \\ &= \sqrt{n(n-1)} \, \tilde{u}_n(y_1, \ldots, y_n) + R_n(y_1, \ldots, y_n) \, , \end{aligned}$$

where we have defined, for $y_1, \ldots, y_n \in \mathbb{R}^2$ and $a_1, a_2, b_1, b_2 \in \mathbb{R}$,

$$\begin{aligned} \tilde{u}_n(y_1, \ldots, y_n) \; &\coloneqq \; \frac{1}{n(n-1)} \sum\nolimits_{i \ne j} \tilde{k}_u(y_i, y_j) \, , \qquad R_n(y_1, \ldots, y_n) \; \coloneqq \; \frac{1}{n} \sum\nolimits_{i=1}^{n} y_{i2}^2 \, , \\ \tilde{k}_u((a_1, a_2), (b_1, b_2)) \; &\coloneqq \; \frac{a_1}{2\sqrt{n-1}} + \frac{b_1}{2\sqrt{n-1}} + \frac{\sqrt{n-1}}{\sqrt{n}} \, a_2 b_2 \, . \end{aligned}$$

Thus, our construction of the U-statistic is to take $k_u(x_1, x_2) \coloneqq \tilde{k}_u(\phi_{d(n)}(x_1), \phi_{d(n)}(x_2))$, which gives

$$u_n(X) \; = \; \frac{1}{n(n-1)} \sum\nolimits_{1 \le i \ne j \le n} k_u(X_i, X_j) \; \overset{d}{=} \tilde{u}_n(Y) \; = \; p_n^*(Y) - R_n(Y) \, .$$

The main technical task is thus to show that $p_n^*(Y)$ approximates a chi-squared distribution and that $R_n(Y)$ has negligible effect other than centering the chi-squared distribution.

B.3 Proofs for Theorems 3.8 and B.1

Throughout this section, we denote the collection of Gaussian vectors

$$Z = (Z_1, \dots, Z_n) \quad \text{where} \quad Z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbb{E}[Y_{1;\sigma_n}], \text{Var}[Y_{1;\sigma_n}]) = \mathcal{N}\Big(\left(\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right), \left(\begin{smallmatrix} \sigma_n^2 & 0 \\ 0 & 1 \end{smallmatrix}\right) \Big) .$$

We state three intermediate results used in the proof of Theorems 3.8 and B.1. The first two results concerns the Gaussian universality approximations of $\tilde{v}_n(Y)$ and $\tilde{u}_n(Y)$. We improve the upper bound of Theorem 4.7, $n^{-\frac{\nu-2}{4\nu+2}}$, by using an argument specific to the construction instead of the generic the Lindeberg method. Note that the σ_0-dependence below arises because Y and Z are implicitly parameterised by σ_0.

Lemma B.2. *Fix $\nu \in (2, 3]$. Then there exist some absolute constants $c, C, \sigma_0 > 0$ and $N \in \mathbb{N}$, such that for all $n \geq N$,*

$$c n^{-\frac{\nu-2}{4\nu}} \leq \sup_{t \in \mathbb{R}} \left| \mathbb{P}(n\, \tilde{v}_n(Y) \leq t) - \mathbb{P}(n\, \tilde{v}_n(Z) \leq t) \right| \leq C n^{-\frac{\nu-2}{4\nu}} .$$

Lemma B.3. *Fix $\nu \in (2, 3]$ and let σ_0 be given as in Lemma B.2. Then there exist some absolute constants $c', C' > 0$ and $N' \in \mathbb{N}$, such that for all $n \geq N'$,*

$$c' n^{-\frac{\nu-2}{4\nu}} \leq \sup_{t \in \mathbb{R}} \left| \mathbb{P}\Big(\sqrt{n(n-1)}\, \tilde{u}_n(Y) \leq t \Big) - \mathbb{P}\Big(\sqrt{n(n-1)}\, \tilde{u}_n(Z) \leq t \Big) \right| \leq C' n^{-\frac{\nu-2}{4\nu}} .$$

Now observe that the universality approximations can be expressed as

$$\begin{aligned} n\, \tilde{v}_n(Z) &= \frac{1}{\sqrt{n}} \sum\nolimits_{i=1}^n Z_{i1} + \frac{1}{n} \sum\nolimits_{i,j=1}^n Z_{i2} Z_{j2} \overset{d}{=} \sigma_n \xi + \chi_1^2 , \\ \sqrt{n(n-1)}\, \tilde{u}_n(Z) &= \frac{1}{\sqrt{n}} \sum\nolimits_{i=1}^n Z_{i1} + \frac{1}{n} \sum\nolimits_{i \neq j} Z_{i2} Z_{j2} \overset{d}{=} \sigma_n \xi + \frac{1}{n} \sum\nolimits_{i \neq j} Z_{i2} Z_{j2} . \end{aligned}$$

The next result allows the quadratic part of $\sqrt{n(n-1)}\, \tilde{u}_n(Z)$ to be approximated by the centred chi-squared variable $\overline{\chi_1^2}$ plus a small Gaussian component $\sigma_n \xi$.

Lemma B.4. *There exists some absolute constant $C'' > 0$ such that for all $n \in \mathbb{N}$,*

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\Big(\sqrt{n(n-1)}\, \tilde{u}_n(Z) \leq t \Big) - \mathbb{P}\Big(\sigma_n \xi + \overline{\chi_1^2} \leq t \Big) \right| \leq C'' n^{-1/5} .$$

These results allow us to state the proofs for Theorems 3.8 and B.1.

Proof of Theorem B.1. Recall that by construction, $u_n(X) \overset{d}{=} \tilde{u}_n(Y)$, $v_n(X) \overset{d}{=} \tilde{v}_n(Y)$ and $n\tilde{v}_n(Z) \overset{d}{=} \sigma_n \xi + \chi_1^2$. The V-statistic bound then follows directly from Lemma B.2: There exist some constants $\sigma_0, c_1, C_1 > 0$ and $N_1 \in \mathbb{N}$ such that for all $n \geq N_1$,

$$c_1 n^{-\frac{\nu-2}{4\nu}} \leq \sup_{t \in \mathbb{R}} \left| \mathbb{P}\big(n\, v_n(X) \leq t \big) - \mathbb{P}\big(\sigma_n \xi + \chi_1^2 \leq t \big) \right| \leq C_1 n^{-\frac{\nu-2}{4\nu}} .$$

□

Proof of Theorem 3.8. By using the triangle inequality to combine the U-statistic bounds

from Lemma B.3 and Lemma B.4, there is another absolute constant $C_2 > 0$ such that

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}\big(\sqrt{n(n-1)}\, u_n(X) \le t\big) - \mathbb{P}\Big(\sigma_n\xi + \overline{\chi_1^2} \le t\Big)\right| \le C_1 n^{-\frac{\nu-2}{4\nu}} + C_2 n^{-\frac{1}{5}} ,$$
$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}\big(\sqrt{n(n-1)}\, u_n(X) \le t\big) - \mathbb{P}\Big(\sigma_n\xi + \overline{\chi_1^2} \le t\Big)\right| \ge c_1 n^{-\frac{\nu-2}{4\nu}} - C_2 n^{-\frac{1}{5}} .$$

Since $\frac{\nu-2}{4\nu} \le \frac{1}{8} < \frac{1}{5}$ for $\nu \in (2,3]$, the $n^{-1/5}$ term can be ignored when n is large. In particular, there exist $c \in (0, c_1]$, $C > C_1$ and an integer $N \ge N_1$ such that for all $n \ge N$,

$$cn^{-\frac{\nu-2}{4\nu}} \le \sup_{t\in\mathbb{R}} \left|\mathbb{P}\big(n\, v_n(X) \le t\big) - \mathbb{P}\big(\sigma_n\xi + \chi_1^2 \le t\big)\right| \le Cn^{-\frac{\nu-2}{4\nu}} .$$

□

The rest of the section proves the intermediate results, i.e. Lemmas B.2 to B.4.

B.3.1. Proof of Lemma B.2

The lower bound follows directly from Theorem 2 of Huang et al. (2024) by noting that $n\tilde{v}_n = p_n^*$, so it suffices to prove the upper bound. Denote

$$S_n \coloneqq \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\frac{1}{\sqrt{2}} U_{i;\sigma_n} + \frac{\sigma_n}{\sqrt{2}}\xi_{i1}\right)$$

such that, for $\xi \sim \mathcal{N}(0,1)$ independent of all other variables,

$$n\tilde{v}_n(Y) \overset{d}{=} S_n + \chi_1^2 \qquad \text{and} \qquad n\tilde{v}_n(Z) \overset{d}{=} \sigma_n\xi + \chi_1^2 .$$

By Lemmas 19 and 20 of Huang et al. (2024), under the choice of σ_0 and $\tilde{N}$ in Theorem 2 of Huang et al. (2024), we have that for all $n \ge \tilde{N}$,

$$\sup_{x\in\mathbb{R}} \left|\mathbb{P}(S_n < x) - \mathbb{P}(\sigma_n\xi < x) - \frac{A}{n^{1/2}\sigma_n^{\nu/(\nu-2)}}\left(1 - \frac{x^2}{\sigma_n^2}\right)e^{-x^2/(2\sigma_n^2)}\right| \le \frac{2}{n\sigma^{2\nu/(\nu-2)}} ,$$
$$\left|\mathbb{P}(S_n < x) - \mathbb{P}(\sigma_n\xi < x)\right| \le B\left(\frac{1}{n^{1/2}\sigma_n^{\nu/(\nu-2)}} e^{-x^2/16\sigma_n^2} + \frac{1}{n^{3/2}x^4\sigma_n^{(8-\nu)/(\nu-2)}}\right)$$

for some absolute constants $A, B > 0$. By splitting an integral and using these bounds,

$$\begin{aligned}
&\left|\mathbb{P}(n\,\tilde{v}_n(Y) < t) - \mathbb{P}(n\,\tilde{v}_n(Z) < t)\right| \\
&= \left|\int_{-\infty}^{\infty} \left(\mathbb{P}(S_n < t - y^2) - \mathbb{P}(\sigma_n\xi < t - y^2)\right) \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy\right| \\
&\le \int_{|t-y^2|<\sigma_n} \left|\mathbb{P}(S_n < t - y^2) - \mathbb{P}(\sigma_n\xi < t - y^2)\right| \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy \\
&\quad + \int_{|t-y^2|\ge\sigma_n} \left|\mathbb{P}(S_n < t - y^2) - \mathbb{P}(\sigma_n\xi < t - y^2)\right| \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy \\
&\le \frac{A}{\sqrt{2\pi}\, n^{1/2}\sigma_n^{\nu/(\nu-2)}} \int_{|t-y^2|<\sigma_n} \left|1 - \frac{(t-y^2)^2}{\sigma_n^2}\right| e^{-\frac{(t-y^2)^2}{2\sigma_n^2} - \frac{y^2}{2}} dy + \frac{4\sqrt{\sigma_n}}{n\sigma_n^{2\nu/(\nu-2)}} \\
&\quad + \frac{B}{\sqrt{2\pi}\, n^{1/2}\sigma_n^{\nu/(\nu-2)}} \int_{|t-y^2|\ge\sigma_n} e^{-\frac{(t-y^2)^2}{16\sigma_n^2} - \frac{y^2}{2}} dy
\end{aligned}$$

$$
\begin{aligned}
&\quad + \frac{B}{\sqrt{2\pi}\, n^{3/2}\sigma_n^{(8-\nu)/(\nu-2)}} \int_{|t-y^2|\geq\sigma_n} \frac{1}{(t-y^2)^4}\, e^{-\frac{y^2}{2}}\, dy \\
&\leq \frac{2A\sqrt{\sigma_n}}{\sqrt{2\pi}\, n^{1/2}\sigma_n^{\nu/(\nu-2)}} + \frac{4\sqrt{\sigma_n}}{n\sigma_n^{2\nu/(\nu-2)}} + \frac{B\sqrt{\sigma_n}}{\sqrt{2\pi}\, n^{1/2}\sigma_n^{\nu/(\nu-2)}} \int_{|\sigma_n^{-1}t-y^2|\geq 1} e^{-\frac{(\sigma_n^{-1}t-y^2)^2}{16}}\, dy \\
&\quad + \frac{B\sqrt{\sigma_n}}{\sqrt{2\pi}\, n^{3/2}\sigma_n^{(8-\nu)/(\nu-2)}\sigma_n^4} \int_{|\sigma_n^{-1}t-y^2|\geq 1} \frac{1}{(\sigma_n^{-1}t-y^2)^4}\, dy\ .
\end{aligned}
$$

In the last line, we have used a change-of-variable $y \mapsto \sqrt{\sigma_n}y$ and noted that $e^{-y^2/2} \leq 1$. Now fix $N \geq \tilde{N}$ such that $\sigma_n = \sigma_0 n^{-(\nu-2)/(2\nu)}$ for all $n \geq N$, which only depends on the absolute constant $\sigma_0 > 0$; in this case,

$$
\frac{\sqrt{\sigma_n}}{n^{1/2}\sigma_n^{\nu/(\nu-2)}} \;=\; \sigma_0^{\frac{1}{2}-\frac{\nu}{\nu-2}} n^{-\frac{\nu-2}{4\nu}}\ , \qquad \frac{\sqrt{\sigma_n}}{n^{3/2}\sigma_n^{(8-\nu)/(\nu-2)}\sigma_n^4} \;=\; \sigma_0^{\frac{1}{2}-\frac{3\nu}{\nu-2}} n^{-\frac{\nu-2}{4\nu}}\ .
$$

Then there exists some constant A' that depends only on σ_0 such that, for all $n > N$,

$$
\begin{aligned}
&\left|\mathbb{P}(n\,\tilde{v}_n(Y) < t) - \mathbb{P}(n\,\tilde{v}_n(Z) < t)\right| \\
&\leq\; A' n^{-\frac{\nu-2}{4\nu}} \Big(1 + \int_{|\sigma_n^{-1}t-y^2|\geq 1} e^{-\frac{(\sigma_n^{-1}t-y^2)^2}{16}}\, dy + \int_{|\sigma_n^{-1}t-y^2|\geq 1} \frac{1}{(\sigma_n^{-1}t-y^2)^4}\, dy\Big) \\
&=:\; A' n^{-\frac{\nu-2}{4\nu}} \left(1 + I_1(\sigma_n^{-1}t) + I_2(\sigma_n^{-1}t)\right)\ ,
\end{aligned}
$$

where we write $I_1(\tau) \coloneqq \int_{|\tau-y^2|\geq 1} e^{-(\tau-y^2)^2/16} dy$ and $I_2(\tau) \coloneqq \int_{|\tau-y^2|\geq 1} (\tau-y^2)^{-4} dy$. To handle $I_1(\tau)$, we split the integral further and use a change-of-variable with $z = y^2$ to obtain

$$
\begin{aligned}
I_1(\tau) \;&\leq\; \int e^{-(\tau-y^2)^2/16} dy \;=\; \int_{y^2<1} e^{-(\tau-y^2)^2/16} dy + 2\int_{y^2\geq 1, y\geq 0} e^{-(\tau-y^2)^2/16} dy \\
&\leq 2 \;+\; 2\int_{z\geq 1} e^{-(\tau-z)^2/16} \frac{1}{2\sqrt{z}}\, dz \\
&\leq 2 \;+\; \int_{z\geq 1} e^{-(\tau-z)^2/16}\, dz \;\leq\; 2 \;+\; \sqrt{2\pi\times 8} \int \frac{e^{-(\tau-z)^2/16}}{\sqrt{2\pi\times 8}}\, dz \;=\; 2 + 4\sqrt{\pi}\ .
\end{aligned}
$$

A similar strategy applied to $I_2(\tau)$ gives

$$
\begin{aligned}
I_2(\tau) \;&=\; \int_{|\tau-y^2|\geq 1, y^2<1} \frac{1}{(\tau-y^2)^4}\, dy + 2\int_{|\tau-y^2|\geq 1, y^2\geq 1, y\geq 0} \frac{1}{(\tau-y^2)^4}\, dy \\
&\leq 2 + 2\int_{|\tau-z|\geq 1, z\geq 1} \frac{1}{(\tau-z)^4} \frac{1}{2\sqrt{z}}\, dz \\
&\leq 2 + \int_{|\tau-z|\geq 1} \frac{1}{(\tau-z)^4}\, dz \;=\; 2 + \int_{|z'|\geq 1} \frac{1}{(z')^4}\, d(z') \;=\; \frac{8}{3}\ .
\end{aligned}
$$

Substituting these two bounds back and noting that the resulted bounds do not depend on t, we get that there exist some constants $C > 0$ and $N \in \mathbb{N}$ that depend only on the absolute constant $\sigma_0 > 0$ such that for all $n \geq N$,

$$
\sup_{t\in\mathbb{R}} \left|\mathbb{P}(n\,\tilde{v}_n(Y) < t) - \mathbb{P}(n\,\tilde{v}_n(Z) < t)\right| \;\leq\; C n^{-\frac{\nu-2}{4\nu}}\ .
$$

□

B.3.2. Proof of Lemma B.3

Notice that

$$\sqrt{n(n-1)}\tilde{u}_n \;=\; p_n^* - R_n \;=\; n\tilde{v}_n - R_n\;,$$

and we already have the universality approximation bound for $n\tilde{v}_n$ from Lemma B.2. It suffices to apply variance domination to approximate $\sqrt{n(n-1)}\tilde{u}_n$ by $n\tilde{v}_n$, which requires us to compute the relevant variances. Write $Z_i = (Z_{i1}, Z_{i2})$, where Z_{i1} is the Gaussian component that match the first two moments of $U_{i;\sigma_n} + 2^{-1/2}\sigma_n\xi_{i1}$ and Z_{i2} is the Gaussian component that matches $\xi_{i2} \sim \mathcal{N}(0,1)$ in distribution. Then by independence,

$$\begin{aligned}\sigma_*^2 \;:=\; \mathrm{Var}[p_n^*(Y)] \;&= \mathrm{Var}[p_n^*(Z)] \;=\; \mathrm{Var}\Big[\tfrac{1}{\sqrt{n}}\textstyle\sum_{i=1}^n Z_{i1}\Big] + \mathrm{Var}\Big[\Big(\tfrac{1}{\sqrt{n}}\textstyle\sum_{i=1}^n Z_{i2}\Big)^2\Big] \\ &\geq \mathrm{Var}\Big[\Big(\tfrac{1}{\sqrt{n}}\textstyle\sum_{i=1}^n Z_{i2}\Big)^2\Big] \;=\; \mathrm{Var}[Z_{12}^2] \;=\; 2\;.\end{aligned}$$

By independence again,

$$\mathrm{Var}[R_n(Y)] \;=\; \mathrm{Var}[R_n(Z)] \;=\; \mathrm{Var}\Big[\tfrac{1}{n}\textstyle\sum_{i=1}^n Z_{i2}^2\Big] \;=\; \dfrac{\mathrm{Var}[Z_{i2}^2]}{n} \;=\; \dfrac{2}{n}\;,$$

and also note that

$$\mathbb{E}[R_n(Y)] \;=\; \mathbb{E}[R_n(Z)] \;=\; 1\;.$$

We now prove the upper bound by combining the variance domination result in Corollary 4.5 and the upper bound of Lemma B.2. Let σ_0 be given as in Lemma B.2. Since

$$\sqrt{n(n-1)}\,\tilde{u}_n(Y) \;=\; (p_n^*(Y) - 1) - (R_n - \mathbb{E}[R_n(Y)])\;,$$

there are some absolute constants $\tilde{C}' > 0$ and $N \in \mathbb{N}$ such that for all $n \geq N$ and every $t \in \mathbb{R}$,

$$\begin{aligned}&\big|\mathbb{P}\big(\sqrt{n(n-1)}\,\tilde{u}_n(Y) \leq t\big) - \mathbb{P}\big(p_n^*(Y) - 1 \leq t\big)\big| \\ &\quad\leq\; \tilde{C}'\Big(\frac{\mathrm{Var}[R_n(Y)]}{\mathrm{Var}[p_n^*(Y)]}\Big)^{\frac{1}{5}} + 2\sup_{\tau\in\mathbb{R}}\big|\mathbb{P}\big(\sigma_*^{-1}(p_n^*(Y)-1) \leq \tau\big) - \mathbb{P}\big(\sigma_*^{-1}(p_n^*(Z)-1) \leq \tau\big)\big| \\ &\quad\leq\; \tilde{C}' n^{-\frac{1}{5}} + 2Cn^{-\frac{\nu-2}{4\nu}} \;\leq\; \tilde{C}'_* n^{-\frac{\nu-2}{4\nu}}\end{aligned}$$

for some absolute constants $C, \tilde{C}'_* > 0$; in the last line, we have used that $\frac{\nu-2}{4\nu} \leq \frac{1}{8} < \frac{1}{5}$ for $\nu \in (2,3]$. Similarly we have

$$\big|\mathbb{P}\big(\sqrt{n(n-1)}\,\tilde{u}_n(Z) \leq t\big) - \mathbb{P}\big((p_n^*(Z) - 1) \leq t\big)\big| \;\leq\; \tilde{C}'_* n^{-\frac{\nu-2}{4\nu}}\;.$$

Combining both bounds with the upper bound of Lemma B.2 by the triangle inequality, we get the desired upper bound that for some absolute constant $C' > 0$ and all $n \geq N$,

$$\sup_{t\in\mathbb{R}}\Big|\mathbb{P}\Big(\sqrt{n(n-1)}\,\tilde{u}_n(Y) \leq t\Big) - \mathbb{P}\Big(\sqrt{n(n-1)}\,\tilde{u}_n(Z) \leq t\Big)\Big| \;\leq\; C' n^{-\frac{\nu-2}{4\nu}}\;.$$

For the lower bound, we apply Lemma 21 of Huang et al. (2024): There exist some constants $c_* > 0$ and $\tilde{N} \in \mathbb{N}$ depending only on the fixed constant σ_0—which is chosen to be the one used in Theorem 2 of Huang et al. (2024) and also Lemma B.2—such that for any $n \geq \tilde{N}$,

$$\mathbb{P}\big(p_n^*(Z) < -2\sigma_n\big) - \mathbb{P}\big(p_n^*(Y) < -2\sigma_n\big) \;\geq\; c_*\, n^{-\frac{\nu-2}{4\nu}} \,.$$

To exploit this lower bound, we shall apply Lemma A.11 directly: For any $\epsilon > 0$,

$$\begin{aligned}
&\mathbb{P}\big(\sqrt{n(n-1)}\,\tilde{u}_n(Z) < -2\sigma_n - \sigma_*\epsilon - 1\big) - \mathbb{P}\big(\sqrt{n(n-1)}\,\tilde{u}_n(Y) < -2\sigma_n - \sigma_*\epsilon - 1\big) \\
&= \mathbb{P}\big(\sqrt{n(n-1)}\,\tilde{u}_n(Z) < -2\sigma_n - \sigma_*\epsilon - 1\big) - \mathbb{P}\big(p_n^*(Z) - 1 < -2\sigma_n - 2\sigma_*\epsilon - 1\big) \\
&\quad - \mathbb{P}\big(-2\sigma_n - 2\sigma_*\epsilon \leq p_n^*(Z) < -2\sigma_n\big) \\
&\quad + \mathbb{P}\big(p_n^*(Z) < -2\sigma_n\big) - \mathbb{P}\big(p_n^*(Y) < -2\sigma_n\big) \\
&\quad + \mathbb{P}\big(p_n^*(Y) - 1 < -2\sigma_n - 1\big) - \mathbb{P}\big(\sqrt{n(n-1)}\,\tilde{u}_n(Y) < -2\sigma_n - \sigma_*\epsilon - 1\big) \\
&\geq -2\mathbb{P}\big(\big|R_n(Y) - 1\big| \geq \sigma_*\epsilon\big) - \mathbb{P}\big(|p_n^*(Z) + 2\sigma_n + \sigma_*\epsilon| \leq \sigma_*\epsilon\big) + c_*\, n^{-\frac{\nu-2}{4\nu}} \\
&\overset{(a)}{\geq} -\frac{2\mathrm{Var}[R_n(Y)]}{\sigma_*^2\epsilon^2} - \frac{\tilde{c}'_*(\sigma_*\epsilon)^{1/2}}{(\mathbb{E}[|p_n^*(Z) + 2\sigma_n + \sigma_*\epsilon|^2])^{1/4}} + c_*\, n^{-\frac{\nu-2}{4\nu}} \\
&\overset{(b)}{\geq} -\frac{2}{\epsilon^2 n} - \tilde{c}'_*\epsilon^{1/2} + c_*\, n^{-\frac{\nu-2}{4\nu}}
\end{aligned}$$

for some absolute constant $\tilde{c}'_* > 0$. In (a), we have used the Markov's inequality and the Carbery-Wright inequality (Fact 4.4); in (b), we have plugged in the bounds on $\mathrm{Var}[R_n(Y)]$ and σ_*, and noted that $\mathbb{E}[|p_n^*(Z) + 2\sigma_n + \sigma_*\epsilon|^2] \geq \mathrm{Var}[p_n^*(Z)] = \sigma_*^2$. Taking $\epsilon = n^{-2/5}$ gives

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}\Big(\sqrt{n(n-1)}\,\tilde{u}_n(Y) \leq t\Big) - \mathbb{P}\Big(\sqrt{n(n-1)}\,\tilde{u}_n(Z) \leq t\Big)\right| \geq c_*\, n^{-\frac{\nu-2}{4\nu}} - \Big(\frac{4}{5} + \tilde{c}'_*\Big) n^{-\frac{1}{5}} \,,$$

Since $\frac{\nu-2}{4\nu} \leq \frac{1}{8} < \frac{1}{5}$, there are some absolute constants $c' > 0$ and $N' \geq \tilde{N}$ such that the desired lower bound holds for all $n \geq N'$. □

B.3.3. Proof of Lemma B.4

Recall from the proof of Lemma B.3 that $\mathbb{E}[R_n(Z)] = 1$, $\mathrm{Var}[R_n(Z)] = \frac{2}{n}$ and that $\mathrm{Var}[n\tilde{v}_n(Z)] = \mathrm{Var}[p_n^*(Z)] \geq 2$. Since

$$\sqrt{n(n-1)}\,\tilde{u}_n(Z) \;=\; n\tilde{v}_n(Z) - \frac{1}{n}\sum\nolimits_{i=1}^n Z_{i2}^2 \;=\; \big(n\tilde{v}_n(Z) - 1\big) - \Big(\frac{1}{n}\sum\nolimits_{i=1}^n Z_{i2}^2 - 1\Big) \,,$$

we can again apply variance domination (Corollary 4.5) to obtain that

$$\begin{aligned}
\sup_{t\in\mathbb{R}} \left|\mathbb{P}\Big(\sqrt{n(n-1)}\,\tilde{u}_n(Z) \leq t\Big) - \mathbb{P}\Big(n\tilde{v}_n(Z) - 1 \leq t\Big)\right| \;&\leq C\left(\frac{\mathrm{Var}[R_n(Z)]}{\mathrm{Var}[n\tilde{v}_n(Z) - 1]}\right)^{\frac{1}{5}} \\
&= Cn^{-\frac{1}{5}}
\end{aligned}$$

for some absolute constant $C > 0$. Noting that $n\tilde{v}_n(Z) - 1 \stackrel{d}{=} \sigma_n \xi + \overline{\chi_1^2}$ finishes the proof. □

B.4 Proof of Proposition 4.3

By Lemma A.11 followed by the Markov's inequality, we get that for every $t \in \mathbb{R}$ and $\epsilon > 0$,

$$\begin{aligned}
\mathbb{P}(X' + Y' \le t) \;&\le\; \mathbb{P}(X' \le t + \epsilon) + \mathbb{P}(|Y'| \ge \epsilon) \;\le\; \mathbb{P}(X' \le t + \epsilon) + \frac{\mathrm{Var}[Y']}{\epsilon^2}\,, \\
\mathbb{P}(X' + Y' \le t) \;&\ge\; \mathbb{P}(X' \le t - \epsilon) - \frac{\mathrm{Var}[Y']}{\epsilon^2}\,.
\end{aligned}$$

Subtracting $\mathbb{P}(X' \le t)$ from both sides, we get that

$$\begin{aligned}
&\big|\mathbb{P}(X' + Y' \le t) - \mathbb{P}(X' \le t)\big| \\
&\le\; \max\Big\{\, \mathbb{P}(X' \in (t - \epsilon, t])\,,\, \mathbb{P}(X' \in (t, t + \epsilon])\Big\} + \frac{\mathrm{Var}[Y']}{\epsilon^2}\,,
\end{aligned}$$

and taking an infimum over $\epsilon > 0$ gives the first bound. The second bound follows by rescaling t and ϵ at the same time by $\sigma'_X > 0$. □

Appendix C

Discussions and proofs for Chapters 4 and 5

This appendix provides additional results and proofs concerning both the general universality results in Chapter 4 and the applications considered in Chapter 5. The appendix is organised as follows:

- Appendix C.1 includes additional discussions. Appendix C.1.1 illustrates the intuition behind variance domination and changes in asymptotic regimes through a toy degree-three V-statistic. Appendix C.1.2 discuss when one expects Assumption 5.1 to hold;
- Appendix C.2 proves the upper bound result in Theorem 4.1;
- Appendix C.3 proves the pair of upper and lower bounds in Theorem 4.7;
- Appendix C.4 proves the variance domination result of Theorem 4.2;
- Appendix C.5 includes several results from Denker (1985), useful for computing the moments of higher-order U-statistics;
- Appendix C.6 proves the results in Section 5.1, which concern the universality of a simple V-statistic under δ-regularity;
- Appendix C.7 proves the results for the remaining applications in Chapter 5. Appendix C.7.1 proves Proposition 5.5 for delta method, Appendix C.7.2 proves Proposition 5.6 for U-statistics, and Appendix C.7.3 proves the results concerning subgraph count statistics in Section 5.3.
- Appendix C.8 proves the properties of several univariate distributions discussed in Section 4.5.

C.1 Additional results

C.1.1. A toy degree-three V-statistic

Given a feature map $\phi : \mathbb{R}^d \to \mathbb{R}$ and i.i.d. $\mathbb{R}^d$-valued random vectors Y_i's, consider a V-statistic

$$v(Y) \coloneqq \frac{1}{n^3} \sum_{i,j,k \le n} \phi(Y_i)\phi(Y_j)\phi(Y_k) \, .$$

Write $\mu_r \coloneqq \mathbb{E}[X_1^r]$. Since $v(Y)$ is a degree-three polynomial in the random variables $X_i \coloneqq \phi(Y_i)$, Theorem 4.1 approximates $p(X) \coloneqq v(Y)$ by replacing the $\mathbb{R}^3$ block tensors

$$\mathbf{X}_i \;=\; \left(\bar{X}_i \, , \, \overline{X_i^2} \, , \, \overline{X_i^3}\right) \;=\; \left(X_i - \mu_1 \, , \, X_i^2 - \mu_2 \, , \, X_i^3 - \mu_3\right)$$

by Gaussian surrogates ξ_i with the same mean and variance as $\mathbf{X}_i$. To understand the structure of the multilinear representation $q(\mathbf{X})$ defined in (4.2), we consider the decomposition

$$\begin{aligned}
v(Y) - \mathbb{E}\, v(Y) \;&=\; p(X) - \mathbb{E}\, p(X) \\
&= \frac{1}{n^3} \sum_{i,j,k \le n} X_i X_j X_k - \frac{(n-1)(n-2)}{n^2} \mu_1^3 - \frac{3(n-1)}{n^2} \mu_1 \mu_2 - \frac{1}{n^2} \mu_3 \\
&= \left(\frac{1}{n^3} \sum_{i,j,k \text{ distinct}} X_i X_j X_k - \frac{(n-1)(n-2)}{n^2} \mu_1^3 \right) + \left(\frac{3}{n^3} \sum_{i \ne j} X_i^2 X_j - \frac{3(n-1)}{n^2} \mu_1 \mu_2 \right) \\
&\quad + \left(\frac{1}{n^3} \sum_{i=1}^n X_i^3 - \frac{1}{n^2} \mu_3 \right) \\
&= \frac{3(n-1)}{n^2} \sum_{i \le n} \mu_1^2 \bar{X}_j \;+\; \frac{3}{n^2} \sum_{i,j \text{ distinct}} \mu_1 \bar{X}_i \bar{X}_j \;+\; \frac{1}{n^3} \sum_{i,j,k \text{ distinct}} \bar{X}_i \bar{X}_j \bar{X}_k \\
&\quad + \frac{3}{n^2} \sum_{i=1}^n \overline{X_i^2} \mu_1 \;+\; \frac{3}{n^2} \sum_{i=1}^n \mu_2 \bar{X}_i \;+\; \frac{3}{n^3} \sum_{i \ne j} \overline{X_i^2} \bar{X}_j \;+\; \frac{1}{n^3} \sum_{i=1}^n \overline{X_i^3} \\
&=: q_1(\mathbf{X}) + q_2(\mathbf{X}) + q_3(\mathbf{X}) + q_4(\mathbf{X}) + q_5(\mathbf{X}) + q_6(\mathbf{X}) + q_7(\mathbf{X}) \;=:\; q(\mathbf{X}) \, .
\end{aligned}$$

Note that $q_1(\mathbf{X})$, $q_2(\mathbf{X})$ and $q_3(\mathbf{X})$ correspond to the Hoeffding's decomposition of the U-statistic associated with $p(X) = v(Y)$ (see (5.6) in Section 5.3).

Theorem 4.1 gives the error of approximating $q(\mathbf{X})$ by $q(\Xi)$, where we recall that Ξ denotes the collection of the Gaussian surrogates $(\xi_1, \ldots, \xi_n)$. Variance domination (Theorem 4.2) says when we can make a further approximation by some $q_l(\Xi)$, depending on how the variances of $q_1(\mathbf{X}), \ldots, q_7(\mathbf{X})$ compare: They can be computed respectively as

$$\begin{aligned}
3n^{-3}(n-1)^2 \mu_1^4 \operatorname{Var}[X_1] \;&=:\; n^{-1}\sigma_1^2 \, , & 3n^{-3}(n-1)\mu_1^2 \operatorname{Var}[X_1]^2 \;&=:\; n^{-2}\sigma_2^2 \, , \\
n^{-5}(n-1)(n-2) \operatorname{Var}[X_1]^3 \;&=:\; n^{-3}\sigma_3^2 & 3n^{-3}\mu_1^2 \operatorname{Var}[X_1^2] \;&=:\; n^{-3}\sigma_4^2 \, , \\
3n^{-3}\mu_2^2 \operatorname{Var}[X_1] \;&=:\; n^{-3}\sigma_5^2 \, , & 3n^{-5}(n-1) \operatorname{Var}[X_1^2] \operatorname{Var}[X_1] \;&=:\; n^{-4}\sigma_6^2 \, , \\
n^{-5} \operatorname{Var}[X_1^3] \;&=:\; n^{-5}\sigma_7^2 \, .
\end{aligned}$$

Theorem 4.2 then provides the error of approximation by each $q_l(\Xi)$: For example, we

have

$$\sup_{t\in\mathbb{R}}\Big|\mathbb{P}\Big(\frac{n^{1/2}}{\sigma_1}\big(v(Y)-\mathbb{E}[v(Y)]\big)\le t\Big)-\mathbb{P}\Big(\frac{n^{1/2}}{\sigma_1}q_1(\Xi)\le t\Big)\Big|$$
$$\le C\bigg(\bigg(\frac{n^{3/2}\sigma_2+n(\sigma_3+\sigma_4+\sigma_5)+n^{1/2}\sigma_6+\sigma_7}{n^2\sigma_1}\bigg)^{\frac{2}{3}}+n^{-\frac{\nu-2}{2\nu+2}}\Big(\frac{\|\bar{X}_1\|_{L_\nu}}{\|\bar{X}_1\|_{L_2}}\Big)^{\frac{\nu}{\nu+1}}\bigg)\,,$$
$$\sup_{t\in\mathbb{R}}\Big|\mathbb{P}\Big(\frac{n}{\sigma_2}\big(v(Y)-\mathbb{E}[v(Y)]\big)\le t\Big)-\mathbb{P}\Big(\frac{n}{\sigma_2}q_2(\Xi)\le t\Big)\Big|$$
$$\le C\bigg(\bigg(\frac{n^2\sigma_1+n(\sigma_3+\sigma_4+\sigma_5)+n^{1/2}\sigma_6+\sigma_7}{n^{3/2}\sigma_2}\bigg)^{\frac{2}{5}}+n^{-\frac{\nu-2}{4\nu+2}}\Big(\frac{\|\bar{X}_1\|_{L_\nu}}{\|\bar{X}_1\|_{L_2}}\Big)^{\frac{2\nu}{2\nu+1}}\bigg)\,.$$

We may also characterise the different limits. $q_1(\Xi)$ is a Gaussian. Writing $\bar{\xi}_1=\frac{1}{n}\sum_{i=1}^n\xi_{i1}$, where ξ_{i1} is the first coordinate of ξ_i (corresponding to $\bar{X}_i$), we have

$$q_2(\Xi)\;=\;\big(\bar{\xi}_1\big)^2-\frac{1}{n}\Big(\frac{1}{n}\sum\nolimits_{i\le n}\xi_{i1}^2\Big)\;\approx\;\big(\bar{\xi}_1\big)^2-\frac{1}{n}\mathbb{E}\big[\big(\bar{\xi}_1\big)^2\big]\,,\tag{C.1}$$

i.e. $q_2(\Xi)$ asymptotically behaves like a centred and rescaled chi-square. An analogous argument shows that $q_3(\Xi)$ behaves like a centred and rescaled cubic power of a Gaussian, $q_6(\Xi)$ behaves like a centred and rescaled chi-square, and $q_4(\Xi)$, $q_5(\Xi)$ and $q_7(\Xi)$ are Gaussian.

Classically with d fixed and all σ_l^2 bounded, the limit of $v(Y)$ can be read off from variance domination: $q_1(\Xi)$ always dominates when $\sigma_1\neq 0$, $q_2(\Xi)$ dominates when $\sigma_1=0$ and $\sigma_2\neq 0$, and so on. Meanwhile, provided that X_1 is not constant almost surely, the only way to set some σ_l to zero is by requiring μ_1 to be zero. The only possible limits are therefore

(i) the Gaussian limit given by $q_1(\Xi)$ when $\mu_1\neq 0$;
(ii) a mixture limit corresponding to $q_3(\Xi)+q_5(\Xi)$ when $\mu_1=0$ and $\mu_2\neq 0$.

Notably since d is fixed, δ-regularity holds directly for the V-statistics associated with each $q_l(\mathbf{X})$, and Lemma 5.3 allows these two limits to be expressed equivalently by replacing each ξ_i by $(Z_i-\mathbb{E}X_i, Z_i^2-\mathbb{E}X_i^2, Z_i^3-\mathbb{E}Z_i^3)$. This agrees with classical results on asymptotics of non-degenerate ($\mu_1\neq 0$) and degenerate ($\mu_1=0$) V-statistics. We also remark that the presence of $q_5(\Xi)$ in (ii) shows that the limit of a degenerate V-statistic differs from that of its associated degenerate U-statistic.

In the high-dimensional setting, however, σ_l^2 are large relative to n through their dependence on $d\equiv d_n$. Since d is only an implicit parameter affecting the variances, d is allowed to be much larger than n (e.g. with exponential growth). In particular, we may have the mixture limit from $q_3(\Xi)+q_5(\Xi)$ even if $\mu_1\neq 0$: It suffices to ask μ_1 to be asymptotically negligible, and the threshold for comparison is exactly given by variance domination (Theorem 4.2).

Meanwhile, if $q_7(\mathbf{X})$ dominates, the implied limit from $q_7(\Xi)$ (Gaussian universality

with the augmented variables) is different from the limit given by replacing X_i's with Z_i's (Gaussian universality the original variables):

$$n^{5/2} q_7(\Xi) \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_{i3} \;\sim\; \mathcal{N}(0, \mathrm{Var}[X_1^3]) \;,$$

but

$$\mathrm{Var}\Big[\frac{1}{\sqrt{n}} \sum_{i=1}^n (Z_i^3 - \mathbb{E}X_i^3)\Big] \;=\; \mathrm{Var}[Z_i^3] \;=\; \mathbb{E}[X_1]^3 + 3\mathbb{E}[X_1]\mathrm{Var}[X_1] \;.$$

This is a case where δ-regularity is violated, and the limits obtained from the two notions of Gaussian universality disagree. In view of the continuous mapping theorem applied to $v(Y) = \big(\frac{1}{n}\sum_{i=1}^n X_i\big)^3$, we believe that $q_7(\mathbf{X})$ never dominates in $v(Y)$ for reasonable distributions of X_1. Notably if X_1 is sufficiently light-tailed such that $\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$ is asymptotically normal, $\mathrm{Var}[v(Y)]$ is either $\Theta(n^{-1})$ or $\Theta(n^{-3})$, i.e. larger than $\mathrm{Var}[q_7(\mathbf{X})] = \Theta(n^{-5})$.

C.1.2. Assumption 5.1 in L_2

Denote $L_2(\mathcal{E}, \mu)$ as the L_2 space of $\mathcal{E} \to \mathbb{R}$ functions under μ. We first show that Assumption 5.1 in Section 5.3, used for U-statistics, is very mild and holds under mild conditions when $\nu = 2$.

Lemma C.1. *Fix $n, m, d \in \mathbb{N}$, $\nu \geq 1$ and $(\mathcal{E}, \mu)$ a separable measure space. Assume that $\|u(Y_1, \ldots, Y_m)\|_{L_2} < \infty$. Then there exists a sequence of orthonormal $L_2(\mathcal{E}, \mu)$ functions, $\{\phi_k\}_{k=1}^\infty$, and an array of real values $\{\lambda_{k_1 \ldots k_m}\}_{k_1,\ldots,k_m=1}^\infty$, such that*

$$\varepsilon_{K;2} \;=\; \Big\| \sum_{k_1,\ldots,k_m=1}^K \lambda_{k_1\ldots k_m} \phi_{k_1}(Y_1) \ldots \phi_{k_m}(Y_m) - u(Y_1, \ldots, Y_m) \Big\|_{L_2} \xrightarrow{K\to\infty} 0 \;.$$

Proof. Separability of $(\mathcal{E}, \mu)$ implies that $L_2(\mathcal{E}, \mu)$ is a separable Hilbert space (see e.g. Exercise 10(b), Chapter 1, Stein and Shakarchi (2011)), which implies the existence of a countable orthonormal basis $\{\phi_k\}_{k=1}^\infty$. Consider the m-fold product space $(\mathcal{E}^m, \mu^m)$. Then the collection of pointwise products $\{\phi_{k_1} \times \ldots \times \phi_{k_m}\}_{k_1,\ldots,k_m=1}^\infty$ forms an orthonormal basis in $L_2(\mathcal{E}^m, \mu^m)$. The stated assumption implies $u \in L_2(\mathcal{E}^m, \mu^m)$, which implies the desired result. □

Remark C.1. We emphasise that the moment boundedness assumption is only required for every *fixed* n, m and d. Therefore, this does not contradict our overall analysis, which considers how the moments can be large relative to n (through e.g. dependence on d and m).

In the next lemma, we give another choice of approximating functions provided that u is well-approximated by its Taylor expansion in L_ν. For each $k \in K$, let $M_k \in \mathbb{N}$ be

the largest number such that

$$\sum_{M=1}^{M_k-1} d^M \;<\; k \;\leq\; \sum_{M=1}^{M_k} d^M \;.$$

For $y \in \mathbb{R}^d$, let $(y^{\otimes M})_t$ be the t-th coordinate of the tensor $y^{\otimes M}$ according to a fixed total order on $\mathbb{N}^m$, and define

$$\tilde{\phi}_k(y) \;:=\; \left(y^{\otimes M_k}\right)_{k-\sum_{M=1}^{M_k-1} d^M} \;.$$

For example, $\tilde{\phi}_1(y), \ldots, \tilde{\phi}_d(y)$ are the d coordinates of y, $\tilde{\phi}_{d+1}(y), \ldots, \tilde{\phi}_{d+d^2}(y)$ are the d^2 coordinates of $y^{\otimes 2}$ and so on.

Lemma C.2. *Fix $\nu \in (2, 3]$. Assume that u is infinitely differentiable and*

$$\left\| u(Y_1, \ldots, Y_m) - \sum_{l=0}^{L} \partial^l u(\mathbf{0})(Y_1, \ldots, Y_m)^{\otimes l} \right\|_{L_\nu} \xrightarrow{L\to\infty} 0 \;.$$

Then Assumption 5.1 holds with $\{\tilde{\phi}_k\}_{k\in\mathbb{N}}$ defined above and some $\{\tilde{\lambda}_{k_1\ldots k_m}\}_{k_1,\ldots,k_m\in\mathbb{N}}$.

Proof of Lemma C.2. Each coordinate of $(Y_1, \ldots, Y_m)^{\otimes l}$ can be written as a product of $\tilde{\phi}_{k_1}(Y_1), \ldots, \tilde{\phi}_{k_m}(Y_m)$ for some $k_1, \ldots, k_m \leq \sum_{M=1}^{l} d^M$. By identifying $\tilde{\lambda}_{k_1\ldots k_m}$'s as the corresponding coordinate in $\partial^l u(\mathbf{0})$, we see that the Taylor approximation error above is exactly $\epsilon_{K;\nu}$, which converges to zero by assumption. □

C.2 Proof of Theorem 4.1

The proof idea is standard: We first compare the distributions on a class of smooth functions via the Lindeberg method (e.g. Chapter 9 of Van Handel (2014)). The smooth function bound is then combined with the Carbery-Wright inequality (Carbery and Wright, 2001), an anti-concentration bound for a polynomial of Gaussians, to obtain the bound in the Kolmogorov metric. In this section, we also include several lemmas that will simplify subsequent proofs.

The next result allows us to control the difference in expectation of the above approximation functions evaluated at independent random quantities. We will be using the smooth approximation of an indicator function from Lemma A.10.

Lemma C.3. *Fix $t \in \mathbb{R}$, $\delta > 0$ and define $h_{t;\delta} \equiv h_{2;t;\delta}$ as in Lemma A.10. Let $\mathbf{V}, \mathbf{Z}, \mathbf{W}$ be some random vectors in $\mathbb{R}^b$ and Y be a random variable in $\mathbb{R}$, with dependence allowed. Then there exists an absolute constant C such that for any $\nu \in (2, 3]$ and $\delta > 0$, we have*

$$\left|\mathbb{E}[h_{t;\delta}(\mathbf{W}^\top \mathbf{V} + Y) - h_{t;\delta}(\mathbf{W}^\top \mathbf{Z} + Y)]\right| \;\leq\; Q_1 + Q_2 + Q_3 \;,$$

where

$$Q_1 \coloneqq \left|\mathbb{E}\left[h'_{t;\delta}(Y)\,\mathbf{W}^\top(\mathbf{V}-\mathbf{Z})\right]\right|, \quad Q_2 \coloneqq \frac{1}{2}\left|\mathbb{E}\left[h''_{t;\delta}(Y)\left(\mathbf{W}^\top\left(\mathbf{V}\mathbf{V}^\top-\mathbf{Z}\mathbf{Z}^\top\right)\mathbf{W}\right)\right]\right|,$$
$$Q_3 \coloneqq 54\delta^{-\nu}(\|\mathbf{W}^\top\mathbf{V}\|_{L_\nu}^\nu + \|\mathbf{W}^\top\mathbf{Z}\|_{L_\nu}^\nu).$$

Assume additionally that $(\mathbf{V},\mathbf{Z})$ *is independent of* $(\mathbf{W},Y)$. *If* $\mathbb{E}[\mathbf{V}]=\mathbb{E}[\mathbf{Z}]$, *then* $Q_1=0$. *If* $\mathrm{Var}[\mathbf{V}]=\mathrm{Var}[\mathbf{Z}]$, *then* $Q_2=0$.

Proof of Lemma C.3. Since $h_{t;\delta}$ is twice continuously differentiable, by a second-order Taylor expansion with the integral remainder, we get that for any $w,r\in\mathbb{R}$,

$$\begin{aligned}
h_{t;\delta}(u) &= h_{t;\delta}(r) + h'_{t;\delta}(r)\,(u-r) + \int_r^u h''_{t;\delta}(t)\,(u-t)\,dt \\
&= h_{t;\delta}(r) + h'_{t;\delta}(r)\,(u-r) + \int_0^1 h''_{t;\delta}\big((u-r)\theta+r\big)\,(u-r)^2(1-\theta)\,d\theta \\
&= h_{t;\delta}(r) + h'_{t;\delta}(r)\,(u-r) + \mathbb{E}\Big[h''_{t;\delta}\big((u-r)\Theta+r\big)\,(u-r)^2(1-\Theta)\Big], \quad \text{(C.2)}
\end{aligned}$$

where $\Theta\sim\mathrm{Uniform}[0,1]$. By applying this expansion once with $u=\mathbf{W}^\top\mathbf{V}+Y, r=Y$ and once with $u=\mathbf{W}^\top\mathbf{Z}+Y, r=Y$, we get that

$$\begin{aligned}
&\left|\mathbb{E}\left[h_{t;\delta}(\mathbf{W}^\top\mathbf{V}+Y)-h_{t;\delta}(\mathbf{W}^\top\mathbf{Z}+Y)\right]\right| \\
&= \Big|\mathbb{E}\Big[h'_{t;\delta}(Y)\,(\mathbf{W}^\top\mathbf{V}) - h'_{t;\delta}(Y)\,(\mathbf{W}^\top\mathbf{Z}) + h''_{t;\delta}\big(\Theta\mathbf{W}^\top\mathbf{V}+Y\big)\,(\mathbf{W}^\top\mathbf{V})^2(1-\Theta) \\
&\qquad - h''_{t;\delta}\big(\Theta\mathbf{W}^\top\mathbf{Z}+Y\big)\,(\mathbf{W}^\top\mathbf{Z})^2(1-\Theta)\Big]\Big|.
\end{aligned}$$

By adding and subtracting two second derivative terms evaluated at t, we can further obtain

$$\begin{aligned}
&\left|\mathbb{E}\left[h_{t;\delta}(\mathbf{W}^\top\mathbf{V}+Y)-h_{t;\delta}(\mathbf{W}^\top\mathbf{Z}+Y)\right]\right| \\
&\quad= \Big|\mathbb{E}\left[h'_{t;\delta}(Y)\,\mathbf{W}^\top(\mathbf{V}-\mathbf{Z})\right] + \mathbb{E}\left[h''_{t;\delta}(Y)\left((\mathbf{W}^\top\mathbf{V})^2-(\mathbf{W}^\top\mathbf{Z})^2\right)(1-\Theta)\right] \\
&\qquad + \mathbb{E}\left[\left(h''_{t;\delta}\big(\Theta\mathbf{W}^\top\mathbf{V}+Y\big)-h''_{t;\delta}(Y)\right)(\mathbf{W}^\top\mathbf{V})^2(1-\Theta)\right] \\
&\qquad - \mathbb{E}\left[\left(h''_{t;\delta}\big(\Theta\mathbf{W}^\top\mathbf{Z}+Y\big)-h''_{t;\delta}(Y)\right)(\mathbf{W}^\top\mathbf{Z})^2(1-\Theta)\right]\Big| \\
&\quad=: |Q_1+Q_2+Q_V-Q_Z| \;\le\; |Q_1|+|Q_2|+|Q_V|+|Q_Z|. \quad \text{(C.3)}
\end{aligned}$$

We handle the four terms individually. Q_1 is already in the desired form. Q_2 can be simplified by noting that Θ is independent of all other variables:

$$Q_2 \;=\; \frac{1}{2}\,\mathbb{E}\left[h''_{t;\delta}(Y)\left(\mathbf{W}^\top\left(\mathbf{V}\mathbf{V}^\top-\mathbf{Z}\mathbf{Z}^\top\right)\mathbf{W}\right)\right].$$

Now to handle Q_V and Q_Z, we use the Jensen's inequality to move the absolute sign inside the expectation and note that $1-\Theta$ is bounded in norm by 1:

$$\begin{aligned}
|Q_V| &= \left|\mathbb{E}\left[\left(h''_{t;\delta}\big(\Theta\mathbf{W}^\top\mathbf{V}+Y\big)-h''_{t;\delta}(Y)\right)(\mathbf{W}^\top\mathbf{V})^2(1-\Theta)\right]\right| \\
&\le \mathbb{E}\left[\left|h''_{t;\delta}\big(\Theta\mathbf{W}^\top\mathbf{V}+Y\big)-h''_{t;\delta}(Y)\right|\times\left(\mathbf{W}^\top\mathbf{V}\right)^2\right].
\end{aligned}$$

Recall the Hölder property of $h''_{t;\delta}$ from Lemma A.10: For any $\epsilon \in [0,1]$ and $x, y \in \mathbb{R}$, we have

$$\left|h''_{t;\delta}(x) - h''_{t;\delta}(y)\right| \le 54\delta^{-2-\epsilon}\,|x-y|^{\epsilon}\ ,$$

Applying this to the bound above with $\epsilon = \nu - 2 \in (0,1]$, $x = \Theta\mathbf{W}^\top\mathbf{V} + Y$ and $y = Y$, while noting that $|\Theta|$ is bounded above by 1 almost surely, we get that

$$|Q_V| \le 54\,\delta^{-\nu}\mathbb{E}\big[|\Theta\mathbf{W}^\top\mathbf{V}|^{\nu-2}\,(\mathbf{W}^\top\mathbf{V})^2\big] \le 54\,\delta^{-\nu}\left\|\mathbf{W}^\top\mathbf{V}\right\|_{L_\nu}^{\nu}.$$

To deal with Q_Z, the argument is the same except $\mathbf{V}$ is replaced by $\mathbf{Z}$:

$$|Q_Z| \le 54\,\delta^{-\nu}\left\|\mathbf{W}^\top\mathbf{Z}\right\|_{L_\nu}^{\nu}.$$

Applying the bounds to (C.3) then gives the first desired bound:

$$\begin{aligned}\left|\mathbb{E}\big[h_{t;\delta}(\mathbf{W}^\top\mathbf{V}+Y) - h_{t;\delta}(\mathbf{W}^\top\mathbf{Z}+Y)\big]\right| &\\ \le |Q_1| + |Q_2| + 54\,\delta^{-\nu}\big(\left\|\mathbf{W}^\top\mathbf{V}\right\|_{L_\nu}^{\nu} + \left\|\mathbf{W}^\top\mathbf{Z}\right\|_{L_\nu}^{\nu}\big)\ .&\end{aligned}$$

In the case where $\mathbb{E}[\mathbf{V}] = \mathbb{E}[\mathbf{Z}]$ and $(\mathbf{V},\mathbf{Z})$ is independent of $(\mathbf{W}, Y)$, we get that

$$Q_1 = \mathbb{E}\big[h'_{t;\delta}(Y)\,\mathbf{W}^\top(\mathbf{V}-\mathbf{Z})\big] = \mathbb{E}\big[h'_{t;\delta}(Y)\,\mathbf{W}^\top\mathbb{E}[\mathbf{V}-\mathbf{Z}]\big] = 0\ .$$

Similarly in the case where $\mathrm{Var}[\mathbf{V}] = \mathrm{Var}[\mathbf{Z}]$ and $(\mathbf{V},\mathbf{Z})$ is independent of $(\mathbf{W}, Y)$, we have

$$Q_2 = \tfrac{1}{2}\,\mathbb{E}\big[h''_{t;\delta}(Y)\,\big(\mathbf{W}^\top\,\mathbb{E}[\mathbf{V}\mathbf{V}^\top - \mathbf{Z}\mathbf{Z}^\top]\,\mathbf{W}\big)\big] = 0\ . \qquad \square$$

The next lemma is convenient for simplifying moment terms involving Gaussians:

Lemma C.4. *Consider a zero-mean $\mathbb{R}^b$-valued Gaussian vector η. Suppose $\mathrm{Var}[\eta] = \mathrm{Var}[V]$ for some $\mathbb{R}^b$ zero-mean random vector V and let W be a random vector in $\mathbb{R}^b$ independent of η and V. Then for any real number $\nu \ge 2$, we have*

$$\mathbb{E}\big[\big|W^\top\eta\big|^\nu\big] \le \frac{2^{\nu/2}\,\Gamma\big(\frac{\nu+1}{2}\big)}{\sqrt{\pi}}\,\mathbb{E}\big[|W^\top V|^\nu\big]\ ,$$

where Γ represents the Gamma function. In the case $\nu = 2$, we have $\mathrm{Var}[W^\top\eta] = \mathrm{Var}[W^\top V]$.

Proof of Lemma C.4. Note that W is independent of η and V. Conditioning on W, $W^\top\eta$ is a zero-mean normal random variable with variance given by

$$W^\top\mathrm{Var}[\eta]W = W^\top\mathrm{Var}[V]W = \mathbb{E}\big[W^\top VV^\top W\big|W\big] = \mathbb{E}\big[(W^\top V)^2\big|W\big]\ .$$

Applying the formula of the ν-th moment of a Gaussian random variable followed by the Jensen's inequality with $\nu \ge 2$, we get that

$$\mathbb{E}\big[\big|W^\top\eta\big|^\nu\big] = \mathbb{E}\big[\mathbb{E}\big[\big|W^\top\eta\big|^\nu\big|W\big]\big] = \frac{2^{\nu/2}\,\Gamma\big(\frac{\nu+1}{2}\big)}{\sqrt{\pi}}\mathbb{E}\big[\mathbb{E}\big[(W^\top V)^2\big|W\big]^{\nu/2}\big]$$

$$\leq \frac{2^{\nu/2}\,\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi}}\mathbb{E}\left[|W^\top V|^\nu\right] .$$

For $\nu = 2$, the above becomes an equality and we get $\mathrm{Var}[W^\top \eta] = \mathrm{Var}[W^\top V]$. □

Proof of Theorem 4.1. We first note that since q_m is affine in each argument, $\mathbb{E}[q_m(\Xi)] = \mathbb{E}[q_m(\mathbf{X})] = 0$. Meanwhile, by applying Lemma C.4 repeatedly, $\mathrm{Var}[q_m(\mathbf{X})] = \mathrm{Var}[q_m(\Xi)]$. This proves the second part of Theorem 4.1.

To prove the first part, we denote $\tilde{q}_m \coloneqq \sigma^{-1} q_m$ for simplicity. We first approximate the probability terms. For $\tau \in \mathbb{R}$ and $\delta > 0$, let $h_{\tau;\delta} \equiv h_{2;\tau;\delta}$ be the twice continuously differentiable function defined in Lemma A.10, which satisfies that $h_{\tau+\delta;\delta}(x) \leq \mathbb{I}_{\{x>\tau\}} \leq h_{\tau;\delta}(x)$ for any $\tau \in \mathbb{R}$. This allows us to bound

$$\begin{aligned}\mathbb{P}(\tilde{q}_m(\mathbf{X}) > t) - \mathbb{P}(\tilde{q}_m(\Xi) > t - \delta) &= \mathbb{E}\left[\mathbb{I}_{\{\tilde{q}_m(\mathbf{X})>t\}} - \mathbb{I}_{\{\tilde{q}_m(\Xi)>t-\delta\}}\right] \\ &\leq \mathbb{E}\left[h_{t;\delta}\left(\tilde{q}_m(\mathbf{X})\right) - h_{t;\delta}\left(\tilde{q}_m(\Xi)\right)\right] ,\end{aligned}$$

and similarly

$$\mathbb{P}(\tilde{q}_m(\Xi) > t + \delta) - \mathbb{P}(\tilde{q}_m(\mathbf{X}) > t) \leq \mathbb{E}\left[h_{t+\delta;\delta}\left(\tilde{q}_m(\Xi)\right) - h_{t+\delta;\delta}\left(\tilde{q}_m(\mathbf{X})\right)\right] .$$

By expressing $\mathbb{P}(\tilde{q}_m(\Xi) > t - \delta) = \mathbb{P}(\tilde{q}_m(\Xi) > t) + \mathbb{P}(t - \delta < \tilde{q}_m(\Xi) \leq t)$ and performing a similar decomposition for $\mathbb{P}(\tilde{q}_m(\Xi) > t + \delta)$, we obtain that

$$\begin{aligned}\left|\mathbb{P}(\sigma^{-1}q_m(\mathbf{X}) \leq t) - \mathbb{P}(\sigma^{-1}q_m(\Xi) \leq t)\right| &= \left|\mathbb{P}(\tilde{q}_m(\mathbf{X}) > t) - \mathbb{P}(\tilde{q}_m(\Xi) > t)\right| \\ &\leq \max\{\mathbb{P}(t - \delta < \tilde{q}_m(\Xi) \leq t)\,,\, \mathbb{P}(t \leq \tilde{q}_m(\Xi) < t + \delta)\} + \max\{E_t, E_{t+\delta}\} \\ &\leq \mathbb{P}(t - \delta < \tilde{q}_m(\Xi) < t + \delta) + E_t + E_{t+\delta} \,, \qquad \text{(C.4)}\end{aligned}$$

where we have defined $E_\tau \coloneqq \left|\mathbb{E}[h_{\tau;\delta}(\tilde{q}_m(\mathbf{X})) - h_{\tau;\delta}(\tilde{q}_m(\Xi))]\right|$ for $\tau \in \mathbb{R}$. The next step is to control quantities of the form E_τ by the Lindeberg method. We recall $\mathbf{W}_i = (\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{0}, \xi_{i+1}, \ldots, \xi_n) \in \mathbb{R}^{nD}$ and define $\mathbf{W}_i^* = (\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_i, \xi_{i+1}, \ldots, \xi_n)$. Then by expanding the difference into a telescoping sum and applying the triangle inequality,

$$\begin{aligned}E_\tau &= \left|\mathbb{E}[h_{\tau;\delta}(\tilde{q}_m(\mathbf{X})) - h_{\tau;\delta}(\tilde{q}_m(\Xi))]\right| \\ &\leq \textstyle\sum_{i=1}^n \left|\mathbb{E}[h_{\tau;\delta}(\tilde{q}_m(\mathbf{W}_i^*)) - h_{\tau;\delta}(\tilde{q}_m(\mathbf{W}_{i-1}^*))]\right| \\ &\overset{(a)}{=} \textstyle\sum_{i=1}^n \left|\mathbb{E}[h_{\tau;\delta}(\tilde{q}_m(\mathbf{W}_i^*)) - h_{\tau;\delta}(\tilde{q}_m(\mathbf{W}_{i-1}^*))]\right| \;=:\; \textstyle\sum_{i=1}^n E_{\tau;i} \,.\end{aligned}$$

We focus on bounding each $E_{\tau;i}$. Since q_m is affine in the i-th argument, we get

$$\tilde{q}_m(\mathbf{W}_i^*) = \partial_i \tilde{q}_m(\mathbf{W}_i)^\top \mathbf{X}_i + \tilde{q}_m(\mathbf{W}_i) \quad \text{and} \quad \tilde{q}_m(\mathbf{W}_{i-1}^*) = \partial_i \tilde{q}_m(\mathbf{W}_i)^\top \xi_i + \tilde{q}_m(\mathbf{W}_i) \,.$$

Note that $\mathbf{X}_i$ and ξ_i are zero-mean with the same variance, and are independent of $\mathbf{W}_i$.

This allows us to apply Lemma C.3: For any $\nu \in (2, 3]$ and $\delta > 0$, we have

$$\begin{aligned} E_{\tau;i} &= \left|\mathbb{E}\left[h_{\tau;\delta}\left(\partial_i \tilde{q}_m(\mathbf{W}_i)^\top \mathbf{X}_i + \tilde{q}_m(\mathbf{W}_i)\right) - h_{\tau;\delta}\left(\partial_i \tilde{q}_m(\mathbf{W}_i)^\top \xi_i + \tilde{q}_m(\mathbf{W}_i)\right)\right]\right| \\ &\leq 54\,\delta^{-\nu}\left(\left\|\partial_i \tilde{q}_m(\mathbf{W}_i)^\top \mathbf{X}_i\right\|_{L_\nu}^{\nu} + \left\|\partial_i \tilde{q}_m(\mathbf{W}_i)^\top \xi_i\right\|_{L_\nu}^{\nu}\right). \end{aligned} \tag{C.5}$$

Since ξ_i is a Gaussian with the same mean and variance as $\mathbf{X}_i$ and both are independent of $\partial_i f(\mathbf{W}_i)$, by Lemma C.4 and noting that $\nu \leq 3$, there is an absolute constant $C' > 0$ such that

$$\mathbb{E}\left[|\partial_i \tilde{q}_m(\mathbf{W}_i)^\top \xi_i|^\nu\right] \leq C' \left\|\partial_i \tilde{q}_m(\mathbf{W}_i)^\top \mathbf{X}_i\right\|_{L_\nu}^{\nu}.$$

By additionally noting that $M_{\nu;i} = \sigma \|\partial_i \tilde{q}_m(\mathbf{W}_i)^\top \mathbf{X}_i\|_{L_\nu}$ by definition, we get that

$$E_{\tau;i} \leq 54(C'+1)\,\delta^{-\nu}\sigma^{-\nu}M_{\nu;i}^{\nu}.$$

Summing over $i = 1, \ldots, n$ then gives

$$E_\tau \leq \sum_{i=1}^n E_{\tau;i} r \leq 54(C'+1)\,\frac{\sum_{i=1}^n M_{\nu;i}^{\nu}}{\delta^\nu \sigma^\nu}. \tag{C.6}$$

On the other hand, by Lemma 4.4, there exists an absolute constant $C^* > 0$ such that

$$\begin{aligned} \mathbb{P}(t - \delta < \tilde{q}_m(\Xi) < t + \delta) &\leq C^* m\,\delta^{1/m}(\mathbb{E}[|\tilde{q}_m(\Xi) - t|^2])^{-1/2m} \\ &\overset{(a)}{=} C^* m\,\delta^{1/m}(\sigma^{-2}\mathrm{Var}[q_m(\Xi)] + t^2)^{-1/2m} \overset{(b)}{=} C^* m\,\delta^{1/m}(1+t^2)^{-1/2m}. \end{aligned} \tag{C.7}$$

In (a) and (b), we have used that $\mathbb{E}[f(\Xi)] = 0$ and $\mathrm{Var}[f(\Xi)] = \mathrm{Var}[f(\mathbf{X})] = \sigma^2$. Substituting (C.6) and (C.7) into (C.4), we get that there exists some absolute constant $C > 0$ such that

$$\left|\mathbb{P}(\sigma^{-1}q_m(\mathbf{X}) \leq t) - \mathbb{P}(\sigma^{-1}q_m(\Xi) \leq t)\right| \leq \frac{Cm}{2}\left(\frac{\delta^{1/m}}{(1+t^2)^{1/2m}} + \frac{\sum_{i=1}^n M_{\nu;i}^{\nu}}{\delta^\nu \sigma^\nu}\right).$$

Finally by choosing $\delta = (1+t^2)^{\frac{1}{2+2\nu m}}\left(\sigma^{-\nu}\sum_{i=1}^n M_{\nu;i}^{\nu}\right)^{\frac{m}{\nu m+1}}$, we get the desired bound

$$\left|\mathbb{P}(\sigma^{-1}f(\mathbf{X}) \leq t) - \mathbb{P}(\sigma^{-1}f(\Xi) \leq t)\right| \leq Cm\left(\frac{\sum_{i=1}^n M_{\nu;i}^{\nu}}{(1+t^2)^{\nu/2}\sigma^\nu}\right)^{\frac{1}{\nu m+1}}. \qquad \square$$

C.3 Proofs for Theorem 4.7

Recall that q_m^* is the multilinear representation of p_m^* with respect to X. The proof consists of three main steps:

(i) By carefully exploiting the heavy tail of V_i's and the asymmetry of $v_{m'}^*$, we can obtain the first lower bound for the approximation of $p_m^*(X)$ by $p_m^*(Z)$ (Lemma C.5);

(ii) By applying Theorem 4.1 to approximate $p_m^*(X) = q_m^*(\mathbf{X}) + \mathbb{E}[p_m^*(X)]$ by $q_m^*(\Xi) + \mathbb{E}[p_m^*(X)]$, the upper bound involving q_m^* in Theorem 4.7 reduces to a moment

control (Lemma C.7);

(iii) By verifying δ-regularity and modifying the argument of Proposition 5.2, we can approximate $q_m^*(\Xi) + \mathbb{E}[p_m^*(X)]$ further by $p_m^*(Z)$ (Lemma C.8).

We introduce some more notation. Note that we can express $X_i = (V_i, Y_i)$ and $\mathbf{X}_i = (V_i, \mathbf{Y}_i)$, where $\mathbf{Y}_i = (Y_i, \dots, Y_i^m - \mathbb{E}[Y_i^m])$. Correspondingly, we express

$$\begin{aligned} Z_i &= (Z_{V;i}, Z_{Y;i}) \qquad \text{where} \qquad Z_{V;i} \sim \mathcal{N}(0, \operatorname{Var} V_i) \text{ and } Z_{Y;i} \sim \mathcal{N}(0, \operatorname{Var} Y_i) \,, \\ \xi_i &= (\xi_{V;i}, \xi_{\mathbf{Y};i}) \qquad \text{where} \qquad \xi_{V;i} \sim \mathcal{N}(0, \operatorname{Var} V_i) \text{ and } \xi_{\mathbf{Y};i} \sim \mathcal{N}(\mathbf{0}, \operatorname{Var} \mathbf{Y}_i) \,. \end{aligned}$$

Denote the collections $Z_V := (Z_{V;i})_{i \le n}$, $Z_Y := (Z_{Y;i})_{i \le n}$ and $\Xi_{\mathbf{Y}} := (\xi_{\mathbf{Y};i})_{i \le n}$. It is also convenient to denote $q_{v;m}^*$ as the multilinear representation of v_m^* with respect to Y, since

$$q_m^*(\mathbf{X}) = v_1^*(V) + q_{v;m}^*(\mathbf{Y}) \qquad \text{and} \qquad q_m^*(\Xi) \overset{d}{=} v_1^*(Z_V) + q_{v;m}^*(\Xi_{\mathbf{Y}}) \,.$$

Lemma C.5. *Fix $\nu > 2$. Assume that m is even with $m = o(\log n)$. Then there exist some absolute constants $c, \sigma_0 > 0$ and $N \in \mathbb{N}$ such that, for any $n \ge N$,*

$$\mathbb{P}\big(p_m^*(Z) < -2\sigma_n\big) - \mathbb{P}\big(p_m^*(X) < -2\sigma_n\big) \ge c n^{-\frac{\nu-2}{2\nu m}} \,.$$

Proof of Lemma C.5. Since the second coordinate of X_i is already Gaussian, the main hurdle is to approximate the heavy-tailed average by a Gaussian. Let F_V be the c.d.f. of $\frac{1}{\sqrt{n}} \sum_{i=1}^n V_i$ – an empirical average of the heavy-tailed coordinates – and F_Z be the c.d.f. of $\frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i$, where Z_i's are i.i.d. zero-mean random variables with the same variance as V_1. Write φ be the p.d.f. of $\frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i \sim \mathcal{N}(0,1)$. Recalling that m is even and denoting

$$I(x) := F_Z(-2\sigma_n - x^m) - F_V(-2\sigma_n - x^m) \,,$$

we can bound the quantity of interest as

$$\begin{aligned} \mathbb{P}\big(p_m^*(Z) < -2\sigma_n\big) - \mathbb{P}\big(p_m^*(X) < -2\sigma_n\big) &= \int_{0 \le x^m < \infty} I(x)\varphi(x)dx \\ &\ge \int_{0 \le x^m \le \kappa^m \sigma_n} I(x)\varphi(x)dx - \Big| \int_{|x| > \kappa \sigma_n^{1/m}} I(x)\varphi(x)dx \Big| \\ &=: J_1 - J_2 \,, \end{aligned}$$

for some $\kappa \ge 1$ to be chosen later.

Bounding J_1. Recall that for $y \in \mathbb{R}$, Lemma 4.9 provides an error bound for approximating the difference $F_Z(y) - F_n(y)$ by

$$-F_q(y) = \frac{A}{n^{1/2} \sigma_n^{\nu/(\nu-2)}} \left(\frac{y^2}{\sigma_n^2} - 1 \right) e^{-y^2/(2\sigma_n^2)}$$

for some absolute constant $A > 0$. Therefore

$$J_1 \ge \int_{x^m \le \kappa^m \sigma_n} -F_q(-2\sigma_n - x^m)\varphi(x)dx - \frac{2}{n\sigma_n^{2\nu/(\nu-2)}} \int_{x^m \le \kappa^m \sigma_n} \varphi(x)dx$$

$$
\begin{aligned}
&\overset{(a)}{\geq} \frac{A}{n^{1/2}\sigma_n^{\nu/(\nu-2)}} \int_{x^m \leq \sigma_n} \left(\frac{(2\sigma_n + x^m)^2}{\sigma_n^2} - 1 \right) e^{-(2\sigma_n + x^m)^2/(2\sigma_n^2)} \varphi(x) dx \\
&\quad - \frac{2}{n\sigma_n^{2\nu/(\nu-2)}} \int_{x^m \leq \kappa^m \sigma_n} \varphi(x) dx \\
&\overset{(b)}{\geq} \frac{3e^{-9/2}A}{n^{1/2}\sigma_n^{\nu/(\nu-2)}} \int_{x^m \leq \sigma_n} \varphi(x) dx - \frac{2}{n\sigma_n^{2\nu/(\nu-2)}} \int_{x^m \leq \kappa^m \sigma_n} \varphi(x) dx \, .
\end{aligned}
$$

In (a), we have restricted the first integral to a smaller ball by noting that $\kappa \geq 1$; in (b), we have noted in the first integral that $4\sigma_n^2 \leq (2\sigma_n + x^m)^2 \leq 9\sigma_n^2$ in the domain of integration. Since both integrals involve c.d.f. of a standard normal, we obtain

$$
\begin{aligned}
J_1 \; &\geq \frac{3e^{-9/2}A}{n^{1/2}\sigma_n^{\nu/(\nu-2)}} \frac{2\sigma_n^{1/m}}{\sqrt{2\pi}} \exp\left(- \frac{\sigma_n^{2/m}}{2} \right) - \frac{4\kappa\sigma_n^{1/m}}{n\sigma_n^{2\nu/(\nu-2)}} \\
&\geq \frac{\sigma_n^{1/m}}{n^{1/2}\sigma_n^{\nu/(\nu-2)}} \left(A_1 e^{-\frac{1}{2}\sigma_n^{2/m}} - \frac{A_2\kappa}{n^{1/2}\sigma_n^{\nu/(\nu-2)}} \right)
\end{aligned}
$$

for some absolute constants $A_1, A_2 > 0$.

Bounding J_2. Recall that $\sigma_n = \min\left\{\sigma_0 n^{-\frac{\nu-2}{2\nu}}, 1\right\}$, which implies

$$
\sigma_n^{\frac{\nu}{\nu-2}} \; \geq \; \min\left\{\sigma_0^{\frac{\nu}{\nu-2}} n^{-1/2}, 1\right\} \; \geq \; \sigma_0^{\frac{\nu}{\nu-2}} n^{-1/2} \, .
$$

Suppose we choose σ_0 such that $\sigma_0^{\frac{\nu}{\nu-2}} > 6^{-1/2}$. Since $6(6^{-1/2})^2 = 1 \leq n$, by applying Lemma 4.10 with $M = 6^{-1/2}$, there exists some absolute constant $C > 0$ such that

$$
\begin{aligned}
J_2 \; &= \left| \int_{x^m > \kappa^m \sigma_n} \big(F_Z(-2\sigma_n - x^m) - F_n(-2\sigma_n - x^m) \big) \varphi(x) dx \right| \\
&\leq \frac{C}{n^{1/2}\sigma_n^{\nu/(\nu-2)}} \int_{x^m > \kappa^m \sigma_n} e^{-\frac{(2\sigma_n + x^m)^2}{16\sigma_n^2}} \varphi(x) dx \\
&\qquad + \frac{C}{n^{3/2}\sigma_n^{(8-\nu)/(\nu-2)}} \int_{x^m > \kappa^m \sigma_n} \frac{1}{(2\sigma_n + x^m)^4} \varphi(x) dx \\
&=: J_{21} + J_{22} \, .
\end{aligned}
$$

By applying a change-of-variable, we get that

$$
\begin{aligned}
J_{21} \; &\leq \frac{C}{n^{1/2}\sigma_n^{\nu/(\nu-2)}} \int_{x^m > \kappa^m \sigma_n} e^{-\frac{x^{2m}}{16\sigma_n^2}} \varphi(x) dx \\
&= \frac{C\sigma_n^{1/m}}{n^{1/2}\sigma_n^{\nu/(\nu-2)}} \int_{x^m > \kappa^m} e^{-\frac{x^{2m}}{16}} \varphi\big(\sigma_n^{1/m} x\big) dx \\
&\overset{(a)}{\leq} \frac{2C\sigma_n^{1/m}}{n^{1/2}\sigma_n^{\nu/(\nu-2)}} \int_{x > \kappa} e^{-\frac{x^{2m}}{16}} dx \\
&\leq \frac{2C\sigma_n^{1/m}}{n^{1/2}\sigma_n^{\nu/(\nu-2)}} \int_{x > \kappa} (2m) \left(\frac{x}{\kappa} \right)^{2m-1} e^{-\frac{x^{2m}}{16}} dx \\
&\leq \frac{A_3 \, \sigma_n^{1/m}}{n^{1/2}\sigma_n^{\nu/(\nu-2)}\kappa^{2m-1}}
\end{aligned}
$$

for some absolute constant $A_3 > 0$. In (a), we have noted that $\varphi(x) \leq 1$ and the integrand

is symmetric in x. Similarly, we can bound

$$\begin{aligned} J_{22} &\leq \frac{C}{n^{3/2}\sigma_n^{(8-\nu)/(\nu-2)}} \int_{x^m > \kappa^m \sigma_n} \frac{1}{x^{4m}} \varphi(x) dx \\ &\leq \frac{C}{n^{3/2}\sigma_n^{(8-\nu)/(\nu-2)}} \int_{x^m > \kappa^m \sigma_n} \frac{1}{x^{4m}} dx \\ &= \frac{(4m-1)C}{n^{3/2}\sigma_n^{(8-\nu)/(\nu-2)} \, \sigma_n^{4-1/m} \, \kappa^{4m-1}} \\ &= \frac{A_4 \, \sigma_n^{1/m}(4m-1)}{n^{3/2}\sigma_n^{3\nu/(\nu-2)} \kappa^{4m-1}} \end{aligned}$$

for some absolute constant $A_4 > 0$.

Combining the bounds. Combining the bounds, we get that if $\sigma_0^{\frac{\nu}{\nu-2}} > 6^{-1/2}$, then

$$\begin{aligned} &\mathbb{P}\big(p_m^*(Z) < -2\sigma_n\big) - \mathbb{P}\big(p_m^*(X) < -2\sigma_n\big) \;\geq\; J_1 - J_2 \\ &\geq \frac{\sigma_n^{1/m}}{n^{1/2}\sigma_n^{\nu/(\nu-2)}} \left(A_1 e^{-\frac{1}{2}\sigma_n^{2/m}} - \frac{A_2 \kappa}{n^{1/2}\sigma_n^{\nu/(\nu-2)}} - \frac{A_3}{\kappa^{2m-1}} - \frac{A_4(4m-1)}{n \, \sigma_n^{2\nu/(\nu-2)} \, \kappa^{4m-1}} \right). \end{aligned}$$

Recall that $\sigma_n = \min\left\{\sigma_0 n^{-\frac{\nu-2}{2\nu}}, 1\right\}$ and $m = o(\log n)$. Take the absolute constant $N_2 \in \mathbb{N}$ sufficiently large such that $\sigma_{N_2} = \sigma_0 N_2^{-\frac{\nu-2}{2\nu}} < 1$ and therefore

$$N_2^{1/2} \sigma_{N_2}^{\nu/(\nu-2)} \;=\; \sigma_0^{\nu/(\nu-2)} \,.$$

Also take the absolute constants $\kappa > 2$ sufficiently large such that $\frac{A_3}{\kappa^{2m-1}} < \frac{1}{9}A_1$, $\sigma_0 > 0$ sufficiently large such that $\frac{A_2\kappa}{\sigma_0^{\nu/(\nu-2)}} < \frac{1}{9}A_1$ and $\frac{A_4}{\sigma_0^{2\nu/(\nu-2)}} \frac{4m-1}{\kappa^{4m-1}} < \frac{1}{9}A_1$, and finally $N_1 \geq N_2$ sufficiently large such that $A_1 e^{-\frac{1}{2}\sigma_{N_1}^{2/m}} = A_1 e^{-\frac{1}{2}\sigma_0^{2/m} N_1^{-\frac{\nu-2}{m\nu}}} \geq \frac{1}{2}A_1$. Then we get the desired bound that, for some absolute constant $c > 0$ and all $n \geq N \coloneqq N_1$,

$$\mathbb{P}\big(p_m^*(Z) < -2\sigma_n\big) - \mathbb{P}\big(p_m^*(X) < -2\sigma_n\big) \;\geq\; \frac{A_1}{6} \frac{\sigma_n^{1/m}}{n^{1/2}\sigma_n^{\nu/(\nu-2)}} \;\geq\; c n^{-\frac{\nu-2}{2\nu m}} \,. \qquad \square$$

To obtain the desired moment controls, we will need to use tight bounds on the moments of a univariate Gaussian. The proof of the following result is included in Appendix C.8.1.

Lemma C.6. *Let $Z \sim \mathcal{N}(0,1)$. Then there exist some absolute constants $C, c > 0$ such that*

(i) for any $\nu \geq 1$, $\mathbb{E}|Z|^\nu \leq C\nu^{\nu/2}$;

(ii) for any $m_1, m_2 \in \mathbb{N}$ with different parities, $\mathrm{Cov}[Z^{m_1}, Z^{m_2}] = 0$*;*

(iii) for $m_1, m_2 \in \mathbb{N}$ with the same parity, $\mathrm{Cov}[Z^{m_1}, Z^{m_2}] > c^{m_1+m_2}(m_1+m_2)^{(m_1+m_2)/2}$.

The next result controls the moment ratio arising from Theorem 4.1.

Lemma C.7. *Fix $\nu \in (2,3]$ and let $n \geq 2m^2$. Then for some absolute constant $C > 0$,*

$$\frac{\sum_{i=1}^n \left\|\partial_i\, q_m^*(\mathbf{W}_i)^\top \mathbf{X}_i\right\|_{L_\nu}^\nu}{\left(\mathrm{Var}[p_m^*(X)]\right)^{\nu/2}} \leq C^m n^{-\frac{\nu-2}{2}}, \qquad \textit{where } \mathbf{W}_i \coloneqq \left(\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, 0, \xi_{i+1}, \ldots, \xi_n\right).$$

Proof of Lemma C.7. First note that by independence,

$$\sigma \;=\; \sqrt{\mathrm{Var}[q_m^*(\mathbf{X})]} \;=\; \sqrt{\mathrm{Var}[v_1^*(V)] + \mathrm{Var}[q_{v;m}^*(\Xi_{\mathbf{Y}})]} \;\geq\; \sqrt{\mathrm{Var}[q_{v;m}^*(\Xi_{\mathbf{Y}})]}\,.$$

Next to bound the ν-th moment, denote $W_{V;i} \coloneqq (V_1, \ldots, V_{i-1}, 0, Z_{V;i+1}, \ldots, Z_{V;n})$ and $\mathbf{W}_{\mathbf{Y};i} \coloneqq (\mathbf{Y}_1, \ldots, \mathbf{Y}_{i-1}, 0, \xi_{\mathbf{Y};i+1}, \ldots, \xi_{\mathbf{Y};n})$. By the triangle inequality, we get that

$$\begin{aligned}\left\|\partial_i\, q_m^*(\mathbf{W}_i)^\top \mathbf{X}_i\right\|_{L_\nu} &= \left\|\partial_i\, v_1^*(W_{V;i}) V_i + \partial_i\, q_{v;m}^*(\mathbf{W}_{Y;i})^\top \mathbf{Y}_i\right\|_{L_\nu} \\ &\leq n^{-1/2}\|V_i\|_{L_\nu} + \left\|\partial_i\, q_{v;m}^*(\mathbf{W}_{Y;i})^\top \mathbf{Y}_i\right\|_{L_\nu}.\end{aligned}$$

Since $\nu \in (2,3]$, the quantity to control can be bounded for some absolute constant $C_1 > 0$ as

$$\begin{aligned}\frac{\sum_{i=1}^n \left\|\partial_i\, q_m^*(\mathbf{W}_i)^\top \mathbf{X}_i\right\|_{L_\nu}^\nu}{\left(\mathrm{Var}[p_m^*(X)]\right)^{\nu/2}} &\overset{(a)}{\leq} \frac{4n^{-\frac{\nu-2}{2}}\|V_1\|_{L_\nu} + 4\sum_{i=1}^n \left\|\partial_i\, q_{v;m}^*(\mathbf{W}_{Y;i})^\top \mathbf{Y}_i\right\|_{L_\nu}^\nu}{\left(\mathrm{Var}[q_{v;m}^*(\Xi_{\mathbf{Y}})]\right)^{\nu/2}} \\ &\overset{(b)}{\leq} \frac{C_1 n^{-\frac{\nu-2}{2}}}{\left(\mathrm{Var}[v_m^*(Y)]\right)^{\nu/2}} + \frac{4\sum_{i=1}^n \left\|\partial_i\, q_{v;m}^*(\mathbf{W}_{Y;i})^\top \mathbf{Y}_i\right\|_{L_\nu}^\nu}{\left(\mathrm{Var}[v_m^*(Y)]\right)^{\nu/2}}. \qquad \text{(C.8)}\end{aligned}$$

In (a), we have used the Jensen's inequality to note that $(a+b)^\nu \leq 2^{\nu-1}(a^\nu + b^\nu) \leq 4(a^\nu + b^\nu)$ and noted that V_i's are i.i.d.; in (b), we have used the moment bound on V_1 from Lemma 4.8 and noted that $\mathrm{Var}[q_{v;m}^*(\Xi_{\mathbf{Y}})] = \mathrm{Var}[q_{v;m}^*(\mathbf{Y})] = \mathrm{Var}[v_m^*(Y)]$ by a repeated application of Lemma C.4 and property of the multilinear representation. The first term can be controlled by using the lower bound from Lemma C.6:

$$\mathrm{Var}[v_m^*(Y)] \;=\; \mathrm{Var}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i\right)^m\right] \;=\; \mathrm{Var}[Y_1^m] \;\geq\; (c_2')^m (2m)^m \;\geq\; c_2 \qquad \text{(C.9)}$$

for some absolute constants $c_2', c_2 > 0$. The second term can be controlled by applying Lemma 5.4: Since $n \geq 2m^2$, there exists some absolute constant $C_3 > 0$ such that

$$\frac{\sum_{i=1}^n \left\|\partial_i\, q_{v;m}^*(\mathbf{W}_{Y;i})^\top \mathbf{Y}_i\right\|_{L_\nu}^\nu}{\left(\mathrm{Var}[q_{v;m}^*(\Xi_{\mathbf{Y}})]\right)^{\nu/2}} \;\leq\; (C_3)^m\, n^{-\frac{\nu-2}{2}} \left(\frac{\sum_{k=1}^m \binom{n}{k}(\alpha_\nu^*(k))^2}{\sum_{k=1}^m \binom{n}{k}(\alpha_2^*(k))^2}\right)^{\frac{\nu}{2}}, \qquad \text{(C.10)}$$

where

$$\alpha_\nu^*(k) \;\coloneqq\; \left\| \sum_{\substack{p_1+\ldots+p_k=m \\ p_1,\ldots,p_k \geq 1}} \prod_{l=1}^k \left(Y_l^{p_l} - \mathbb{E}Y_l^{p_l}\right) \right\|_{L_\nu}.$$

We are left with bounding $\alpha_2^*(k)$ and $\alpha_\nu^*(k)$. Since Y_i's are i.i.d. standard normals, the proof is down to using the moment bounds for standard normals from Lemma C.6. For convenience, we denote $\overline{Y_l^{p_l}} \coloneqq Y_l^{p_l} - \mathbb{E}Y_l^{p_l}$ from now on. We can then express

$$(\alpha_2^*(k))^2 \;=\; \sum_{\substack{p_1+\ldots+p_k=m \\ p_1,\ldots,p_k \geq 1}} \sum_{\substack{q_1+\ldots+q_k=m \\ q_1,\ldots,q_k \geq 1}} \mathrm{Cov}\left[\prod_{l=1}^k \overline{Y_l^{p_l}}\,,\, \prod_{l=1}^k \overline{Y_l^{p_l}}\right].$$

We now show by induction that for some absolute constant $c > 0$,

$$\mathrm{Cov}\left[\prod_{l=1}^{k} \overline{Y_l^{p_l}}\,,\, \prod_{l=1}^{k} \overline{Y_l^{p_l}}\right] \;\geq\; \prod_{l=1}^{k} \mathbb{I}_{\{p_l \equiv q_l \,(\mathrm{mod}\, 2)\}}\, c^{p_l+q_l}\,(p_l+q_l)^{(p_l+q_l)/2}\,.$$

For $k = 1$, this directly holds for all $p_l, q_l \in [m]$ by the lower bound of Lemma C.6. Suppose this holds for $k-1$. By the total law of covariance and noting that all variables are centred and independent,

$$\begin{aligned}
\mathrm{Cov}\left[\prod_{l=1}^{k} \overline{Y_l^{p_l}}\,,\, \prod_{l=1}^{k} \overline{Y_l^{p_l}}\right] &= \mathbb{E}\,\mathrm{Cov}\left[\prod_{l=1}^{k} \overline{Y_l^{p_l}}\,,\, \prod_{l=1}^{k} \overline{Y_l^{p_l}}\,\middle|\, Y_1,\ldots,Y_{k-1}\right] \\
&= \mathbb{E}\left[\overline{Y_k^{p_k}} \times \overline{Y_k^{q_k}}\right] \mathrm{Cov}\left[\prod_{l=1}^{k-1} \overline{Y_l^{p_l}}\,,\, \prod_{l=1}^{k-1} \overline{Y_l^{p_l}}\right] \\
&\geq \mathrm{Cov}[Y_k^{p_k}\,,\, Y_k^{q_k}] \prod_{l=1}^{k-1} \mathbb{I}_{\{p_l \equiv q_l \,(\mathrm{mod}\, 2)\}}\, c^{p_l+q_l}\,(p_l+q_l)^{(p_l+q_l)/2} \\
&\geq \prod_{l=1}^{k} \mathbb{I}_{\{p_l \equiv q_l \,(\mathrm{mod}\, 2)\}}\, c^{p_l+q_l}\,(p_l+q_l)^{(p_l+q_l)/2}\,.
\end{aligned}$$

This finishes the induction, and in particular implies

$$\begin{aligned}
(\alpha_2^*(k))^2 &\geq c^{2m} \sum_{\substack{p_1+\ldots+p_k=m \\ p_1,\ldots,p_k\geq 1}} \sum_{\substack{q_1+\ldots+q_k=m \\ q_1,\ldots,q_k\geq 1}} \prod_{l=1}^{k} \mathbb{I}_{\{p_l \equiv q_l \,(\mathrm{mod}\, 2)\}}\,(p_l+q_l)^{(p_l+q_l)/2} \\
&\overset{(a)}{\geq} c^{2m} \sum_{\substack{p_1+\ldots+p_k=m \\ p_1,\ldots,p_k\geq 1}} \sum_{\substack{q_1+\ldots+q_k=m \\ q_1,\ldots,q_k\geq 1}} \\
&\qquad \times \prod_{l=1}^{k} \mathbb{I}_{\{p_l \equiv q_l \,(\mathrm{mod}\, 2)\}}\, \frac{(p_l+q_l)^{(p_l+q_l)/2} + (c')^{p_l+q_l}(p_l+q_l+1)^{(p_l+q_l+1)/2}}{2} \\
&\geq (c'')^{2m} \sum_{\substack{p_1+\ldots+p_k=m \\ p_1,\ldots,p_k\geq 1}} \sum_{\substack{q_1+\ldots+q_k=m \\ q_1,\ldots,q_k\geq 1}} (p_l+q_l)^{(p_l+q_l)/2}
\end{aligned}$$

for some absolute constant $c'' > 0$. In (a), we have noted that

$$\begin{aligned}
(p_l+q_l)^{(p_l+q_l)/2} &= (p_l+q_l)^{-1/2}(p_l+q_l)^{(p_l+q_l+1)/2} \\
&\geq \left((p_l+q_l)^{-1/2} 2^{-(p_l+q_l+1)/2}\right)(p_l+q_l+1)^{(p_l+q_l+1)/2} \\
&\geq (c')^{p_l+q_l}(p_l+q_l+1)^{(p_l+q_l+1)/2}
\end{aligned}$$

for some absolute constant $c' > 0$. On the other hand, by the triangle inequality and noting that Y_i's are i.i.d., we have

$$\begin{aligned}
\left(\alpha_\nu^*(k)\right)^2 &= \left\| \sum_{\substack{p_1+\ldots+p_k=m \\ p_1,\ldots,p_k\geq 1}} \prod_{l=1}^{k} \overline{Y_l^{p_l}} \right\|_{L_\nu}^2 \\
&\leq \sum_{\substack{p_1+\ldots+p_k=m \\ p_1,\ldots,p_k\geq 1}} \sum_{\substack{q_1+\ldots+q_k=m \\ q_1,\ldots,q_k\geq 1}} \left\| \prod_{l=1}^{k} \overline{Y_l^{p_l}} \right\|_{L_\nu} \left\| \prod_{l=1}^{k} \overline{Y_l^{q_l}} \right\|_{L_\nu} \\
&= \sum_{\substack{p_1+\ldots+p_k=m \\ p_1,\ldots,p_k\geq 1}} \sum_{\substack{q_1+\ldots+q_k=m \\ q_1,\ldots,q_k\geq 1}} \prod_{l=1}^{k} \left\| \overline{Y_1^{p_l}} \right\|_{L_\nu} \left\| \overline{Y_1^{q_l}} \right\|_{L_\nu}\,.
\end{aligned}$$

By noting that $\nu \leq 3$ and using the moment bound in Lemma C.6, we get that for some absolute constants $C', C'' > 0$,

$$\begin{aligned}
\left\| \overline{Y_1^{p}} \right\|_{L_\nu}^{\nu} \left\| \overline{Y_1^{q}} \right\|_{L_\nu}^{\nu} &= \mathbb{E}\left[\left|Y_1^p - \mathbb{E}[Y_1^p]\right|^\nu\right] \mathbb{E}\left[\left|Y_1^q - \mathbb{E}[Y_1^q]\right|^\nu\right] \\
&\leq C' \mathbb{E}[|Y_1|^{p\nu}] \mathbb{E}[|Y_1|^{q\nu}] \;\leq\; C'' p^{p\nu/2} q^{q\nu/2} \;\leq\; C''(p+q)^{(p+q)\nu/2}\,.
\end{aligned}$$

This implies that

$$\big(\alpha_\nu^*(k)\big)^2 \;\le\; \sum_{\substack{p_1+\ldots+p_k=m\\ p_1,\ldots,p_k\ge 1}} \sum_{\substack{q_1+\ldots+q_k=m\\ q_1,\ldots,q_k\ge 1}} \prod_{l=1}^k (C'')^{1/\nu}(p_l+q_l)^{(p_l+q_l)/2} \;\le\; (\tilde{c})^{2m}\big(\alpha_2^*(k)\big)^2$$

for some absolute constant $\tilde{c} > 0$. (C.10) then becomes

$$\frac{\sum_{i=1}^n \big\|\partial_i\, q_{v;m}^*(\mathbf{W}_{Y;i})^\top \mathbf{Y}_i\big\|_{L_\nu}^\nu}{(\mathrm{Var}[q_{v;m}^*(\Xi_{\mathbf{Y}})])^{\nu/2}} \;\le (C_3)^m\, n^{-\frac{\nu-2}{2}} \left(\frac{\sum_{k=1}^m \binom{n}{k}(\alpha_\nu^*(k))^2}{\sum_{k=1}^m \binom{n}{k}(\alpha_2^*(k))^2}\right)^{\frac{\nu}{2}} \;\le\; (\tilde{c}^\nu C_3)^{\nu m}\, n^{-\frac{\nu-2}{2}}\,.$$

Plugging in this bound and (C.9) into (C.8), we get that

$$\frac{\sum_{i=1}^n \big\|\partial_i\, q_m^*(\mathbf{W}_i)^\top \mathbf{X}_i\big\|_{L_\nu}^\nu}{\big(\mathrm{Var}[p_m^*(X)]\big)^{\nu/2}} \;\le \frac{C_1 n^{-\frac{\nu-2}{2}}}{c_2^{\nu/2}} + 4(\tilde{c}^\nu C_3)^{\nu m}\, n^{-\frac{\nu-2}{2}} \;\le\; C^m n^{-\frac{\nu-2}{2}}$$

for some absolute constant $C > 0$. This finishes the proof. □

The next result allows $q_m^*(\Xi) + \mathbb{E}[p_m^*(X)]$ to be approximated by $p_m^*(Z)$.

Lemma C.8. *Let C and c be the absolute constants in the upper and lower bounds in Lemma C.6, and assume that $m \le \delta(\log n)(2 + \max\{\log c, 0\} + \max\{\log C, 0\})^{-1}$ for some $\delta \in [0,1)$. Also assume that Z_Y and $\Xi_{\mathbf{Y}}$ are coupled such that $Z_{Y;i} = \xi_{Y;i1}$ almost surely. Then there exists some absolute constant $C' > 0$ such that*

$$\frac{\|q_m^*(\Xi) + \mathbb{E}[p_m^*(X)] - p_m^*(Z)\|_{L_2}^2}{\mathrm{Var}[q_m^*(\Xi)]} \;\le\; \frac{(C')^m}{n^{1-\delta}}\,.$$

Proof of Lemma C.8. We first compute the quantities from (7) and (8) in the definition of δ-regularity in Section 5.1, which can be bounded together as

$$\begin{aligned}
&\max_{q_1+\ldots+q_j=m\,,\, q_l\in\mathbb{N}} \max\left\{\left(\mathbb{E}\Big[n^{m/2}\textstyle\prod_{l=1}^j Y_l^{q_l}\Big]\right)^2, \mathrm{Var}\Big[n^{m/2}\textstyle\prod_{l=1}^j Y_l^{q_l}\Big]\right\} \\
&\quad\le\; n^m \max_{q_1+\ldots+q_j=m\,,\, q_l\in\mathbb{N}} \Big\|\textstyle\prod_{l=1}^j Y_l^{q_l}\Big\|_{L_2}^2 \;=\; n^m \max_{q_1+\ldots+q_j=m\,,\, q_l\in\mathbb{N}} \textstyle\prod_{l=1}^j \|Y_1^{q_l}\|_{L_2}^2 \\
&\quad\le\; n^m \max_{q_1+\ldots+q_j=m\,,\, q_l\in\mathbb{N}} \textstyle\prod_{l=1}^j \big(C(2q_l)^{q_l}\big) \;\le\; n^m\,(2C)^m \max_{q_1+\ldots+q_j=m\,,\, q_l\in\mathbb{N}} \textstyle\prod_{l=1}^j q_l^{q_l}
\end{aligned}$$

for some absolute constant $C > 0$. Note that we have used independence followed by Lemma C.6 again, and that we only need to bound the above for $j \in [m-1]$. We shall now perform an induction to show that

$$Q(j,m) \;:=\; \max_{q_1+\ldots+q_j=m\,,\, q_l\in\mathbb{N}} \prod_{l=1}^j q_l^{q_l} \;=\; (m-j+1)^{m-j+1}\,.$$

This holds trivially for $j = 1 < m$. To prove this for $j \ge 2$ and $m > j$, suppose the statement holds for all $j' < j$ and all $m' > j'$. By applying the pigeonhole principle, q_1 can only take values in $[m-j+1]$, and

$$Q(j,m) \;=\; \max_{q_1\in[m-j+1]} q_1^{q_1} Q(j-1, m-q_1)$$

$$= \max_{q_1 \in [m-j+1]} q_1^{q_1} (m - q_1 - j + 2)^{m-q_1-j+2} = \max_{q_1 \in [m-j+1]} \psi_{m-j+2}(q_1) ,$$

where we have denoted $\psi_{m'}(x) = x^x (m' - x)^{m'-x}$, defined for $x \in [0, m']$. Since $\psi_{m'}(x)$ is strictly convex with a minimum at $x = m'/2$, its maximum is obtained at the boundary, so

$$Q(j, m) = \psi_{m-j+2}(1) = (m - 1 - j + 2)^{m-1-j+2} = (m - j + 1)^{m-j+1} .$$

This finishes the induction. By additionally recalling that, by using the lower bound from Lemma C.6 (as proved in Equation (C.9)), we have

$$\text{Var}[v_m^*(Y)] = \text{Var}\Big[\Big(\frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i\Big)^m\Big] = \text{Var}[Y_1^m] \geq (1/c)^m m^m$$

for some absolute constant $c > 0$. This implies that

$$\begin{aligned}
&\max_{q_1+\ldots+q_j=m\,,\, q_l \in \mathbb{N}} \max \left\{ \left(\mathbb{E}\Big[n^{m/2} \textstyle\prod_{l=1}^j Y_l^{q_l} \Big] \right)^2 , \text{Var}\Big[n^{m/2} \textstyle\prod_{l=1}^j Y_l^{q_l} \Big] \right\} \\
&\leq n^m (2C)^m m^{m-j+1} \\
&\leq n^m (2 \max\{C, 1\})^m m^{m-j+1} c^m m^{-m} \text{Var}[v_m^*(Y)] \\
&\leq (4 \max\{c, 1\} \max\{C, 1\})^m n^m m^{-j} \text{Var}[v_m^*(Y)] \\
&= e^{m(\log 4 + \max\{\log c, 0\} + \max\{\log C, 0\})} n^m m^{-j} \text{Var}[v_m^*(Y)] \\
&< e^{m(2 + \max\{\log c, 0\} + \max\{\log C, 0\})} n^m m^{-j} \text{Var}[v_m^*(Y)] \\
&\leq n^{m+\delta} m^{-j} \text{Var}[v_m^*(Y)] ,
\end{aligned}$$

where we have used the assumption on m in the last line. This implies that v_m^* in (6) of Section 5.1 is δ-regular with respect to Y. By Lemma 5.3 in Section 5.1, there is some absolute constant $C' > 0$ such that

$$\frac{\|q_{v;m}^*(\Xi_{\mathbf{Y}}) + \mathbb{E}[v_m^*(Y)] - v_m^*(Z_Y)\|_{L_2}^2}{\text{Var}[q_{v;m}^*(\Xi_{\mathbf{Y}})]} \leq \frac{(C')^m}{n^{1-\delta}} .$$

Recall that $v_m^*(Z_Y) - \mathbb{E}[v_m^*(Y)] = p_m^*(Z) - \mathbb{E}[p_m^*(X)] - v_1^*(Z_V)$ and, by the coupling assumption, $q_{v;m}^*(\Xi_{\mathbf{Y}}) = q_m^*(\Xi) - v_1^*(Z_V)$. Also, since $v_m^*(Y) = \big(n^{-1/2} \sum_{i=1}^n Y_i\big)^m \stackrel{d}{=} Y_1^m$, by Lemma C.6, $\text{Var}[q_{v;m}^*(\Xi_{\mathbf{Y}})] = \text{Var}[v_m^*(Y)] > c^m (2m)^m$ for some absolute constant $c > 0$. These imply the desired bound that

$$\frac{\|q_{v;m}^*(\Xi) - (p_m^*(Z) - \mathbb{E}[p_m^*(X)])\|_{L_2}^2}{\text{Var}[f^*(\Xi)]} = \frac{\|q_{v;m}^*(\Xi_{\mathbf{Y}}) + \mathbb{E}[v_m^*(Y)] - v_m^*(Z_Y)\|_{L_2}^2}{\text{Var}[v_1^*(Z_V)] + \text{Var}[q_{v;m}^*(\Xi_{\mathbf{Y}})]} \leq \frac{(C')^m}{n^{1-\delta}} . \quad \square$$

We will also make use of the result in Lemma A.11, which effectively plays the role of Proposition 4.3 for ignoring random variables with negligible variances. Combining the results in this section allows us to prove Theorem 4.7.

Proof of Theorem 4.7 . By a direct application of Theorem 4.1 to $q_m^*(\mathbf{X}) = p_m^*(X) - \mathbb{E}[p_m^*(X)]$ followed by the moment bound computed in Lemma C.7, we get that if $n \geq$

$2m^2$, there is an absolute constant $C' > 0$ such that

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}(q_m^*(\mathbf{X}) \le t) - \mathbb{P}(q_m^*(\Xi) \le t)\right| \le C' m n^{-\frac{\nu-2}{2\nu m+2}} . \tag{C.11}$$

Since $m = o(\log n)$, there is some absolute constant $N' \in \mathbb{N}$ such that the above holds for all $n \ge N'$. Meanwhile since m is even, by Lemma C.5, there are absolute constants $c', \sigma_0 > 0$ and $N'' \in \mathbb{N}$ such that, for any $n \ge N''$,

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}(p_m^*(X) \le t) - \mathbb{P}(p_m^*(Z) \le t)\right| \ge c' n^{-\frac{\nu-2}{2\nu m}} . \tag{C.12}$$

We are left with applying δ-regularity. WLOG assume that $Z_{Y;i} = \xi_{Y;i1}$ almost surely. By Lemma C.8, there are absolute constants $c_*, C_*, C_{**} > 0$ such that, if $m \le \delta(\log n)(2 + \max\{\log c_*, 0\} + \max\{\log C_*, 0\})^{-1}$ for any $\delta \in [0, 1)$, we have

$$\frac{\|q_m^*(\Xi) + \mathbb{E}[p_m^*(X)] - p_m^*(Z)\|_{L_2}^2}{\mathrm{Var}[q_m^*(\Xi)]} \le \frac{(C_{**})^m}{n^{1-\delta}} . \tag{C.13}$$

Since $m = \log(n)$, there is some absolute constant $N''' \in \mathbb{N}$ such that the above holds for all $n \ge N'''$ and any fixed $\delta \in [0, 1)$ to be specified. Now by Lemma A.11, for any $\epsilon > 0$ and $t \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{P}\big(p_m^*(Z) \le t\big) &\le \mathbb{P}\big(q_m^*(\Xi) + \mathbb{E}[p_m^*(X)] \le t + \epsilon\big) + \mathbb{P}\big(|q_m^*(\Xi) + \mathbb{E}[p_m^*(X)] - p_m^*(Z)| \ge \epsilon\big) , \\
\mathbb{P}\big(p_m^*(Z) \le t\big) &\ge \mathbb{P}\big(q_m^*(\Xi) + \mathbb{E}[p_m^*(X)] \le t - \epsilon\big) - \mathbb{P}\big(|q_m^*(\Xi) + \mathbb{E}[p_m^*(X)] - p_m^*(Z)| \ge \epsilon\big) .
\end{aligned}$$

This implies that

$$\begin{aligned}
&\left|\mathbb{P}\big(q_m^*(\Xi) + \mathbb{E}[p_m^*(X)] \le t\big) - \mathbb{P}\big(p_m^*(Z) \le t\big)\right| \\
&\le \mathbb{P}\big(t - \epsilon \le q_m^*(\Xi) + \mathbb{E}[p_m^*(X)] \le t + \epsilon\big) + \mathbb{P}\big(|q_m^*(\Xi) + \mathbb{E}[p_m^*(X)] - p_m^*(Z)| \ge \epsilon\big) \\
&\overset{(a)}{\le} C_1 m \Big(\frac{\epsilon}{\sqrt{\mathrm{Var}[q_m^*(\Xi)]}}\Big)^{1/m} + \frac{\|q_m^*(\Xi) + \mathbb{E}[p_m^*(X)] - p_m^*(Z)\|_{L_2}^2}{\epsilon^2} \\
&\overset{(b)}{\le} C_1 m \Big(\frac{\epsilon}{\sqrt{\mathrm{Var}[q_m^*(\Xi)]}}\Big)^{1/m} + \frac{(C_{**})^m}{n^{1-\delta}} \Big(\frac{\sqrt{\mathrm{Var}[q_m^*(\Xi)]}}{\epsilon}\Big)^2
\end{aligned}$$

for some absolute constant $C_1 > 0$. In (a), we used the Carbery-Wright argument in (C.7) in the proof of Theorem 4.1 for the first term, and Markov's inequality for the second term. In (b), we have applied (C.13). Choosing

$$\epsilon = (C_{**})^{m^2/(2m+1)} n^{-m(1-\delta)/(2m+1)} \sqrt{\mathrm{Var}[q_m^*(\Xi)]} ,$$

we get that for some absolute constant $C_{***} > 0$,

$$\begin{aligned}
\left|\mathbb{P}\big(q_m^*(\Xi) + \mathbb{E}[p_m^*(X)] \le t\big) - \mathbb{P}\big(p_m^*(Z) \le t\big)\right| &\le (C_1 m + 1)(C_{**})^{\frac{m}{2m+1}} n^{-\frac{1-\delta}{2m+1}} \\
&\le C_{***} m n^{-\frac{1-\delta}{2m+1}} .
\end{aligned} \tag{C.14}$$

Combining this with (C.11) via the triangle inequality, we get that for all $n \ge \max\{N', N'''\}$,

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}(p_m^*(X) \le t) - \mathbb{P}(p_m^*(Z) \le t)\right| \le C''' m n^{-\frac{\nu-2}{2\nu m+2}} + C_{***} m n^{-\frac{1-\delta}{2m+1}} ,$$

whereas combining the bound with (C.12) via a reverse triangle inequality, we get that for all $n \geq \max\{N', N''\}$,

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}\big(q_m^*(\mathbf{X}) \leq t\big) - \mathbb{P}\big(q_m^*(\Xi) \leq t\big)\right| \;\geq c'' n^{-\frac{\nu-2}{2\nu m}} - C_{***} m n^{-\frac{1-\delta}{2m+1}} \,.$$

Now choose $\delta = \frac{1}{2}$. Since $\nu \in (2, 3]$ and m is even, this implies

$$\frac{1-\delta}{2m+1} \;=\; \frac{1}{4m+2} \;>\; \frac{1}{6m} \;\geq\; \frac{\nu-2}{2\nu m} \;\geq\; \frac{\nu-2}{2\nu m+2} \,,$$

and that the $m n^{-\frac{1-\delta}{2m+1}} = o\big((\log n) n^{-\frac{1-\delta}{2m+1}}\big)$ term can be ignored. This finishes the proof. □

C.4 Proof of Theorem 4.2

By Lemma A.11, for any $\epsilon > 0$ and $t \in \mathbb{R}$, we have

$$\begin{aligned}
\mathbb{P}\big(\sigma^{-1} f(X) > t\big) \;&\leq \mathbb{P}\big(\sigma^{-1} q_m(\mathbf{X}) > t - \epsilon\big) + \mathbb{P}\big(\sigma^{-1}|f(X) - q_m(\mathbf{X})| \geq \epsilon\big) \,, \\
\mathbb{P}\big(\sigma^{-1} f(X) > t\big) \;&\geq \mathbb{P}\big(\sigma^{-1} q_m(\mathbf{X}) > t + \epsilon\big) - \mathbb{P}\big(\sigma^{-1}|f(X) - q_m(\mathbf{X})| \geq \epsilon\big) \,.
\end{aligned}$$

By additionally expressing

$$\begin{aligned}
\mathbb{P}\big(\sigma^{-1} q_m(\Xi) > t\big) \;&= \mathbb{P}\big(\sigma^{-1} q_m(\Xi) > t + \epsilon\big) + \mathbb{P}\big(t \leq \sigma^{-1} q_m(\Xi) < t + \epsilon\big) \\
&= \mathbb{P}\big(\sigma^{-1} q_m(\Xi) > t - \epsilon\big) - \mathbb{P}\big(t - \epsilon < \sigma^{-1} q_m(\Xi) \leq t\big) \,,
\end{aligned}$$

we can bound

$$\begin{aligned}
\left|\mathbb{P}\big(\sigma^{-1} f(X) \leq t\big) - \mathbb{P}\big(\sigma^{-1} q_m(\Xi) \leq t\big)\right| \;&=\; \left|\mathbb{P}\big(\sigma^{-1} f(X) > t\big) - \mathbb{P}\big(\sigma^{-1} q_m(\Xi) > t\big)\right| \\
&\leq \mathbb{P}\big(\sigma^{-1}|f(X) - q_m(\mathbf{X})| \geq \epsilon\big) + \mathbb{P}\big(|\sigma^{-1} q_m(\Xi) - t| < \epsilon\big) \\
&\quad + \max_{s\in\{+1,-1\}} \left|\mathbb{P}\big(\sigma^{-1} q_m(\mathbf{X}) > t + s\epsilon\big) - \mathbb{P}\big(\sigma^{-1} q_m(\Xi) > t + s\epsilon\big)\right| \\
&=: T_1 + T_2 + T_3 \,.
\end{aligned}$$

By Markov's inequality, we have

$$T_1 \;\leq\; \frac{\|f(X) - q_m(\mathbf{X})\|_{L_2}^2}{\sigma^2 \epsilon^2} \,.$$

By the Carbery-Wright inequality from Lemma 4.4 and noting that $\mathbb{E} q_m(\mathbf{X}) = 0$ and $\operatorname{Var} q_m(\mathbf{X}) = \operatorname{Var} q_m(\Xi)$ by Theorem 4.1, we have that for some absolute constant $C' > 0$,

$$T_2 \;\leq C' m\, \epsilon^{1/m} (\mathbb{E}[(\sigma^{-1} q_m(\Xi) - t)^2])^{-1/2m} \;\leq\; C' m\, \frac{\epsilon^{1/m}}{(1+t^2)^{1/2m}} \,.$$

By Theorem 4.1, there exists some absolute constant $C'' > 0$ such that

$$T_3 \;\leq C'' m \max_{s\in\{+1,-1\}} \left(\frac{\sum_{i=1}^n M_{\nu;i}^\nu}{(1 + (t+s\epsilon)^2)^{\nu/2}\, \sigma^\nu}\right)^{\frac{1}{\nu m+1}} \,.$$

Combining the bounds and choosing $\epsilon = \epsilon_t = (\frac{\|f(X)-q_m(\mathbf{X})\|_{L_2}}{\sigma})^{2m/(2m+1)}(1+t^2)^{1/2(2m+1)}$, we obtain the non-uniform bound: There exists some absolute constant $C > 0$ such that

$$\begin{aligned}&\big|\mathbb{P}\big(\sigma^{-1}f(\mathbf{X}) \le t\big) - \mathbb{P}\big(\sigma^{-1}q_m(\Xi) \le t\big)\big| \\ &\qquad \le Cm\bigg(\Big(\frac{\|f(X) - q_m(\mathbf{X})\|_{L_2}}{\sigma(1+t^2)^{1/2}}\Big)^{\frac{2}{2m+1}} + \max_{s\in\{+1,-1\}} \Big(\frac{\sum_{i=1}^n M_{\nu;i}^\nu}{(1+(t+s\epsilon_t)^2)^{\nu/2}\,\sigma^\nu}\Big)^{\frac{1}{\nu m+1}}\bigg)\,.\end{aligned}$$

Taking a supremum over t gives the desired bound. □

C.5 Moment computation for U-statistics

The proofs for Sections 5.1 to 5.4 rely extensively on moment formula for U-statistics, which are collected next. We follow the notation of Section 5.3: $Y = (Y_i)_{i\le n}$ are i.i.d. variables in a general measurable space $\mathcal{E}$, and $U_j^{\mathrm{H}}(Y)$ and $u_j^{\mathrm{H}}(y_1,\dots,y_j)$ are defined as in Hoeffding's decomposition (11) of u_m in (10). The next result states that Hoeffding decompositions of different orders are orthogonal in L_2.

Lemma C.9 (Lemma 1.2.3 of Denker (1985))**.** *Fix some $j \ne j' \in [m]$. Suppose $\|U_j^{\mathrm{H}}(Y)\|_{L_2}$ and $\|U_{j'}^{\mathrm{H}}(Y)\|_{L_2}$ are bounded. Then $\mathbb{E}[U_j^{\mathrm{H}}(Y)U_{j'}^{\mathrm{H}}(Y)] = 0$.*

Denote $\sigma_r^2 \coloneqq \operatorname{Var}\mathbb{E}[u(Y_1,\dots,Y_m)\,|\,Y_1,\dots,Y_r]$, where u is the kernel of u_m in (10). The next two results compute the variances associated with u_j^{H}, U_j^{H} and u_m.

Lemma C.10 (Lemma 1.2.4 of Denker (1985))**.** *For each $j \in [m]$, we have*

$$\operatorname{Var}\big[u_j^{\mathrm{H}}(Y_1,\dots,Y_j)\big] = \sum_{r=1}^j \binom{j}{j-r}(-1)^{j-r}\sigma_r^2\,.$$

Lemma C.11 (Theorem 4.1.2.2 of Denker (1985))**.** *We have that*

$$\operatorname{Var}\big[U_j^{\mathrm{H}}(Y)\big] = \binom{n}{j}^{-1}\sigma_j^2 \quad \textit{and} \quad \operatorname{Var}[u_m(Y)] = \binom{n}{m}^{-1}\sum_{r=1}^m \binom{m}{r}\binom{n-m}{m-r}\sigma_r^2\,.$$

C.6 Proofs for Section 5.1

The proofs for δ-regularity are mathematically straightforward but tedious, as it involves expanding out a degree-m V-statistic and comparing moment terms of different orders. Throughout, we express $\xi_i = (\xi_{i1},\dots,\xi_{im})$, where each ξ_{ij} matches $\overline{X_i^{\otimes j}} = X_i^{\otimes j} - \mathbb{E}[X_i^{\otimes j}]$ in mean and variance. We require two auxiliary results. The first is the martingale moment bound of Lemma A.4 used in Appendix A.2, and the second expresses a V-statistic as a sum of U-statistics:

Lemma C.12 (Theorem 4.1, p.183, Lee (1990)). *Given a function $u : (\mathbb{R}^d)^m \to \mathbb{R}$, we have*

$$\frac{1}{n^m}\sum\nolimits_{i_1,\ldots,i_m\in[n]} u(y_{i_1},\ldots,y_{i_m}) \;= n^{-m}\sum\nolimits_{j=1}^{m}\binom{n}{j}\tilde{u}_{j;m}(y_1,\ldots,y_n)\;,$$

where $\tilde{u}_{j;m}$ is a degree-j U-statistic defined by

$$\begin{aligned}\tilde{u}_{j;m}(y_1,\ldots,y_n) &\coloneqq \frac{1}{n(n-1)\ldots(n-j+1)}\sum\nolimits_{i_1,\ldots,i_j\in[n]\ \textit{distinct}} u_{j;m}(y_{i_1},\ldots,y_{i_j})\\ u_{j;m}(y_1,\ldots,y_j) &\coloneqq \sum_{\substack{(p_1,\ldots,p_m)\in[j]^m\\ \textit{with exactly } j \textit{ distinct elements}}} u(y_{p_1},\ldots,y_{p_m})\;.\end{aligned}$$

C.6.1. Proof of Lemma 5.3

By Lemma C.12, we have

$$v_m(x_1,\ldots,x_n) \;= n^{-m}\sum\nolimits_{j=1}^{m}\binom{n}{j}\tilde{u}_{j;m}(x_1,\ldots,x_n)\;,$$

where

$$\begin{aligned}\tilde{u}_{j;m}(x_1,\ldots,x_n) &\coloneqq \frac{1}{n(n-1)\ldots(n-j+1)}\sum\nolimits_{i_1,\ldots,i_j\in[n]\ \text{distinct}} u_{j;m}(x_{i_1},\ldots,x_{i_j})\\ u_{j;m}(x_1,\ldots,x_j) &\coloneqq \sum_{\substack{(p_1,\ldots,p_m)\in[j]^m\\ \text{with exactly } j \text{ distinct elements}}} \langle S,\, x_{p_1}\otimes\ldots\otimes x_{p_m}\rangle\;.\end{aligned}$$

By the symmetry of S, we can express

$$\begin{aligned}u_{j;m}(x_1,\ldots,x_j) \;&=\; \sum\nolimits_{q_1+\ldots+q_j=m\,,\,q_l\in\mathbb{N}} \langle S,\, x_1^{\otimes q_1}\otimes\ldots\otimes x_j^{\otimes q_j}\rangle\\ =\sum_{\substack{q_1+\ldots+q_j=m\\ q_l\in\mathbb{N}}} &\langle S,\, \left(x_1^{\otimes q_1}-\mathbb{E}[X_1^{\otimes q_1}]+\mathbb{E}[X_1^{\otimes q_1}]\right)\otimes\ldots\otimes\left(x_j^{\otimes q_j}-\mathbb{E}[X_1^{\otimes q_j}]+\mathbb{E}[X_1^{\otimes q_j}]\right)\rangle\;.\end{aligned}$$

Analogously, we can express, for $y_i=(y_{i1},\ldots,y_{im})$ with $y_{ij}\in\mathbb{R}^{d^j}$,

$$q_m^v(y_1,\ldots,y_n)+\mathbb{E}[v_m(X)] \;= n^{-m}\sum\nolimits_{j=1}^{m}\binom{n}{j}\tilde{u}^q_{j;m}(y_1,\ldots,y_n)\;,$$

where

$$\begin{aligned}\tilde{u}^q_{j;m}(y_1,\ldots,y_n) &\coloneqq \frac{1}{n(n-1)\ldots(n-j+1)}\sum\nolimits_{i_1,\ldots,i_j\in[n]\ \text{distinct}} u^q_{j;m}(y_{i_1},\ldots,y_{i_j})\\ u^q_{j;m}(y_1,\ldots,y_j) &\coloneqq \sum\nolimits_{\substack{q_1+\ldots+q_j=m\\ q_l\in\mathbb{N}}} \left\langle S,\, \left(y_{1q_1}+\mathbb{E}[X_1^{\otimes q_1}]\right)\otimes\ldots\otimes\left(y_{jq_j}+\mathbb{E}[X_1^{\otimes q_j}]\right)\right\rangle\;.\end{aligned}$$

This allows us to express the difference of interest as a sum of centred U-statistics:

$$\begin{aligned}\|q_m^v(\Xi)+\mathbb{E}[v_m(X)]-v_m(Z)\|_{L_2}^2 \;&=\; \left\|n^{-m}\sum\nolimits_{j=1}^{m}\binom{n}{j}\tilde{u}^{\Delta}_{j;m}(\Xi)\right\|_{L_2}^2\\ &\le \left(n^{-m}\sum\nolimits_{j=1}^{m}\binom{n}{j}\left\|\tilde{u}^{\Delta}_{j;m}(\Xi)\right\|_{L_2}\right)^2\\ &\le m\,n^{-2m}\sum\nolimits_{j=1}^{m}\binom{n}{j}^2\left\|\tilde{u}^{\Delta}_{j;m}(\Xi)\right\|_{L_2}^2\;,\end{aligned}$$

where we have noted the coupling $\xi_{i1} = Z_i - \mathbb{E}Z_i$ a.s. and defined the degree-j U-statistic

$$\begin{aligned}
\tilde{u}^{\Delta}_{j;m}(y_1, \dots, y_n) &\coloneqq \frac{1}{n(n-1)\dots(n-j+1)} \sum\nolimits_{i_1,\dots,i_j \in [n] \text{ distinct}} u^{\Delta}_{j;m}(y_{i_1}, \dots, y_{i_j}) \,, \\
u^{\Delta}_{j;m}(y_1, \dots, y_j) &\coloneqq \sum\nolimits_{\substack{q_1+\dots+q_j=m \\ q_l \in \mathbb{N}}} \Big(\big\langle S,\, \big(y_{1q_1} + \mathbb{E}[X_1^{\otimes q_1}]\big) \otimes \dots \otimes \big(y_{jq_j} + \mathbb{E}[X_1^{\otimes q_j}]\big) \big\rangle \\
&\qquad\qquad - \big\langle S,\, y_{11}^{\otimes q_1} \otimes \dots \otimes y_{j1}^{\otimes q_j} \big\rangle \Big) \,.
\end{aligned}$$

The task now is to control $\|\tilde{u}^{\Delta}_{j;m}(\Xi)\|_{L_2}$. By the variance formula in Lemma C.11, we have that

$$\begin{aligned}
\| \tilde{u}^{\Delta}_{j;m}(\Xi)\|^2_{L_2} &= \big(\mathbb{E}\big[\tilde{u}^{\Delta}_{j;m}(\Xi)\big]\big)^2 + \mathrm{Var}\big[\tilde{u}^{\Delta}_{j;m}(\Xi)\big] \\
&= \big(\mathbb{E}\big[u^{\Delta}_{j;m}(\Xi)\big]\big)^2 + \binom{n}{j}^{-1} \sum\nolimits_{r=1}^{j} \binom{j}{r}\binom{n-j}{j-r} \mathrm{Var}\, \mathbb{E}\big[u^{\Delta}_{j;m}(\xi_1, \dots, \xi_j) \,\big|\, \xi_1, \dots, \xi_r\big] \,.
\end{aligned}$$

To compute the moments of $u^{\Delta}_{j;m}$, first note that we can express

$$\begin{aligned}
&\mathbb{E}\big[u^{\Delta}_{j;m}(\xi_1, \dots, \xi_j) \,\big|\, \xi_1, \dots, \xi_r\big] \qquad\qquad (C.15) \\
&\quad = \sum_{\substack{q_1+\dots+q_j=m \\ q_l \in \mathbb{N}}} \Bigg(\Bigg\langle S \,, \bigotimes_{l=1}^{r} \big(\xi_{lq_l} + \mathbb{E}[X_1^{\otimes q_l}]\big) \otimes \bigotimes_{l'=r+1}^{j} \mathbb{E}[X_1^{\otimes q_{l'}}] \Bigg\rangle \\
&\qquad\qquad - \Bigg\langle S \,, \bigotimes_{l=1}^{r} \xi_{l1}^{\otimes q_l} \otimes \bigotimes_{l'=r+1}^{j} \mathbb{E}\big[\xi_{11}^{\otimes q_{l'}}\big] \Bigg\rangle \Bigg) \,.
\end{aligned}$$

Meanwhile, recall that $\xi_{i1} = Z_1$ almost surely,

$$\mathbb{E}[\xi_{i1}] \;=\; \mathbb{E}[X_1] \;=\; 0 \,,$$

and, since $\mathbb{E}[X_i^{\otimes 2}] = \mathbb{E}[Z_i^{\otimes 2}] = \mathbb{E}[\xi_{i1}^{\otimes 2}]$ and $\mathbb{E}[\xi_{i2}] = 0$, we have

$$\begin{aligned}
\mathbb{E}\big[\xi_{i2} + \mathbb{E}\big[X_1^{\otimes 2}\big]\big] - \mathbb{E}\Big[\big(\xi_{i1} + \mathbb{E}[X_1]\big)^{\otimes 2}\Big] &= \mathbb{E}\Big[\mathbb{E}\big[X_1^{\otimes 2}\big] - (Z_1 - \mathbb{E}[Z_1])^{\otimes 2} - \mathbb{E}[X_1]^{\otimes 2}\Big] \\
&= 0 \,.
\end{aligned}$$

This implies that the summand in (C.15) vanishes if

(i) $q_{l'} = 1$ for any $l = r+1, \dots, j$, or

(ii) $q_l = 1$ for all $l = 1, \dots, r$ and $q_{l'} = 2$ for all $l = r+1, \dots, j$.

For (C.15) to be non-zero, (i) and (ii) imply that $r + 2(j-r) < m$ by the pigeonhole principle. In other words, $r \geq 2j - m + 1$, which allows us to rewrite

$$\begin{aligned}
\mathrm{Var}\big[\tilde{u}^{\Delta}_{j;m}(\Xi)\big] &= \binom{n}{j}^{-1} \sum\nolimits_{r=2j-m+1}^{j} \binom{j}{r}\binom{n-j}{j-r} \mathrm{Var}\, \mathbb{E}\big[u^{\Delta}_{j;m}(\xi_1, \dots, \xi_j) \,\big|\, \xi_1, \dots, \xi_r\big] \,, \\
\mathbb{E}\big[u^{\Delta}_{j;m}(\Xi)\big] &= \mathbb{E}\big[u^{\Delta}_{j;m}(\Xi)\big]\, \mathbb{I}_{\{j \leq \lfloor \frac{m-1}{2} \rfloor\}} \,.
\end{aligned}$$

Therefore by the Jensen's inequality combined with the above bound, we get that

$$\|q^v_m(\Xi) + \mathbb{E}[v_m(X)] - v_m(Z)\|^2_{L_2}$$

$$
\begin{aligned}
&\leq \frac{m}{n^{2m}} \sum_{j=1}^{\lfloor \frac{m-1}{2} \rfloor} \binom{n}{j}^2 \big(\mathbb{E}\big[u^{\Delta}_{j;m}(\Xi)\big]\big)^2 \\
&\quad + \frac{m}{n^{2m}} \sum_{j=1}^{m} \binom{n}{j} \sum_{r=2j-m+1}^{j} \binom{j}{r}\binom{n-j}{j-r} \operatorname{Var} \mathbb{E}\big[u^{\Delta}_{j;m}(\xi_1,\ldots,\xi_j) \,\big|\, \xi_1,\ldots,\xi_r\big] \\
&\leq \frac{m}{n^{2m}} \sum_{j=1}^{\lfloor \frac{m-1}{2} \rfloor} \frac{e^{2j} n^{2j}}{j^{2j}} \big(\mathbb{E}\big[u^{\Delta}_{j;m}(\Xi)\big]\big)^2 \\
&\quad + \frac{m}{n^{2m}} \sum_{j=1}^{m-1} \frac{e^j n^j}{j^j} \sum_{r=2j-m+1}^{j} \binom{j}{r} n^{m-1-j} \operatorname{Var}\big[u^{\Delta}_{j;m}(\xi_1,\ldots,\xi_j)\big] \\
&\leq \frac{m}{n^{2m}} \sum_{j=1}^{\lfloor \frac{m-1}{2} \rfloor} \frac{e^{2j} n^{2j}}{j^{2j}} \big(\mathbb{E}\big[u^{\Delta}_{j;m}(\Xi)\big]\big)^2 \\
&\quad + \frac{m}{n^{m+1}} \sum_{j=1}^{m-1} \frac{(2e)^j}{j^j} \operatorname{Var}\big[u^{\Delta}_{j;m}(\xi_1,\ldots,\xi_j)\big] \,. \qquad \text{(C.16)}
\end{aligned}
$$

To get a further control on the moment terms, note that

$$
\begin{aligned}
\operatorname{Var}\big[u^{\Delta}_{j;m}(\xi_1,\ldots,\xi_j)\big] &= \operatorname{Var}\big[u^{q}_{j;m}(\xi_1,\ldots,\xi_j) - u_{j;m}(\xi_{11},\ldots,\xi_{j1})\big] \\
&\leq 2 \operatorname{Var}\big[u^{q}_{j;m}(\xi_1,\ldots,\xi_j)\big] + 2 \operatorname{Var}\big[u_{j;m}(\xi_{11},\ldots,\xi_{j1})\big] \,,
\end{aligned}
$$

and similarly

$$
\big(\mathbb{E}\big[u^{\Delta}_{j;m}(\Xi)\big]\big)^2 \;\leq\; 2\big(\mathbb{E}\big[u^{q}_{j;m}(\Xi)\big]\big)^2 + 2\big(\mathbb{E}\big[u_{j;m}(\Xi)\big]\big)^2 \,.
$$

Meanwhile, note that for any $j \in [m-1]$,

$$
\begin{aligned}
\operatorname{Var}\big[u^{q}_{j;m}(\xi_{11},\ldots,\xi_{j1})\big] \;&=\; \operatorname{Var}\Bigg[\sum_{\substack{q_1+\ldots+q_j=m \\ q_l\in\mathbb{N}}} \Big\langle S\,, \bigotimes_{l=1}^{j} \big(\xi_{1q_l} + \mathbb{E}\big[X_1^{\otimes q_l}\big]\big)\Big\rangle\Bigg] \\
&\leq \binom{m-1}{j-1}^2 \max_{q_1+\ldots+q_j=m\,,\,q_l\in\mathbb{N}} \operatorname{Var}\Bigg[\Big\langle S\,, \bigotimes_{l=1}^{j} \big(\xi_{lq_l} + \mathbb{E}\big[X_1^{\otimes q_l}\big]\big)\Big\rangle\Bigg] \\
&\leq m^{2j} \max_{q_1+\ldots+q_j=m\,,\,q_l\in\mathbb{N}} \operatorname{Var}\Bigg[\Big\langle S\,, \bigotimes_{l=1}^{j} \big(\xi_{lq_l} + \mathbb{E}\big[X_1^{\otimes q_l}\big]\big)\Big\rangle\Bigg] \\
&\overset{(a)}{=} m^{2j} \max_{q_1+\ldots+q_j=m\,,\,q_l\in\mathbb{N}} \operatorname{Var}\Bigg[\Big\langle S\,, \bigotimes_{l=1}^{j} X_l^{\otimes q_l}\Big\rangle\Bigg] \\
&\overset{(b)}{\leq} m^{2j}\big(C'\, n^{m+\delta} m^{-j} \operatorname{Var}[v_m(X)]\big) \;=\; C'\, m^j\, n^{m+\delta} \operatorname{Var}[q^{v}_{m}(\Xi)]
\end{aligned}
$$

for some absolute constant $C' > 0$, where we have used Lemma C.4 in (a) and δ-regularity in (b). Similarly by δ-regularity again,

$$
\begin{aligned}
\operatorname{Var}\big[u_{j;m}(\xi_{11},\ldots,\xi_{j1})\big] &\leq m^{2j} \max_{q_1+\ldots+q_j=m\,,\,q_l\in\mathbb{N}} \operatorname{Var}\Bigg[\Big\langle S\,, \bigotimes_{l=1}^{j} Z_l^{\otimes q_l}\Big\rangle\Bigg] \\
&\leq C'\, m^j\, n^{m+\delta} \operatorname{Var}[q^{v}_{m}(\Xi)] \,,
\end{aligned}
$$

whereas

$$
\max\Big\{\big(\mathbb{E}\big[u^{q}_{j;m}(\Xi)\big]\big)^2\,,\, \big(\mathbb{E}\big[u_{j;m}(\Xi)\big]\big)^2\Big\}
$$

$$
\begin{aligned}
&\leq\ m^{2j} \max_{\substack{q_1+\ldots+q_j=m \\ q_l\in\mathbb{N}}} \\
&\qquad \max\Big\{\Big(\mathbb{E}\Big[\Big\langle S\,,\bigotimes_{l=1}^{j}\big(\xi_{lq_l}+\mathbb{E}\big[X_1^{\otimes q_l}\big]\big)\Big\rangle\Big]\Big)^2,\Big(\mathbb{E}\Big[\Big\langle S\,,\bigotimes_{l=1}^{j} Z_l^{\otimes q_l}\Big\rangle\Big]\Big)^2\Big\} \\
&=\ m^{2j} \max_{q_1+\ldots+q_j=m\,,\,q_l\in\mathbb{N}} \max\Big\{\Big(\mathbb{E}\Big[\Big\langle S\,,\bigotimes_{l=1}^{j} X_1^{\otimes q_l}\Big\rangle\Big]\Big)^2,\Big(\mathbb{E}\Big[\Big\langle S\,,\bigotimes_{l=1}^{j} Z_l^{\otimes q_l}\Big\rangle\Big]\Big)^2\Big\} \\
&\leq\ C'\,m^{2j}\,n^{m+\delta}\,\mathrm{Var}[q_m^v(\Xi)]\,.
\end{aligned}
$$

Substituting these bounds into (C.16), we get that

$$
\begin{aligned}
&\|q_m^v(\Xi)+\mathbb{E}[v_m(X)]-v_m(Z)\|_{L_2}^2 \\
&\qquad\leq 4C'\,m\,n^{-2m}\sum_{j=1}^{\lfloor\frac{m-1}{2}\rfloor}\frac{e^{2j}n^{2j+m+\delta}m^{2j}}{j^{2j}}\,\mathrm{Var}[q_m^v(\Xi)] \\
&\qquad\quad+4C'\,m\,n^{-m-1}\sum_{j=1}^{m-1}\frac{(2e)^j m^j n^{m+\delta}}{j^j}\,\mathrm{Var}[q_m^v(\Xi)] \\
&\qquad\leq 4C'\,m\,n^{-(1-\delta)}\,\mathrm{Var}[q_m^v(\Xi)]\Big(\sum_{j=1}^{\lfloor\frac{m-1}{2}\rfloor}\frac{e^{2j}m^{2j}}{j^{2j}}+\sum_{j=1}^{m-1}\frac{(2e)^j m^j}{j^j}\Big) \\
&\qquad\leq 4C'\,m\,n^{-(1-\delta)}\,\mathrm{Var}[q_m^v(\Xi)]\Big(\sum_{j=0}^{m}e^{2j}\binom{m}{j}^2+\sum_{j=0}^{m}(2e)^j\binom{m}{j}\Big) \\
&\qquad\leq 4C'\,m\,n^{-(1-\delta)}\,\mathrm{Var}[q_m^v(\Xi)]\Big(\Big(\sum_{j=0}^{m}e^{j}\binom{m}{j}\Big)^2+\sum_{j=0}^{m}(2e)^j\binom{m}{j}\Big) \\
&\qquad\leq C^m\,n^{-(1-\delta)}\,\mathrm{Var}[q_m^v(\Xi)]
\end{aligned}
$$

for some absolute constant $C>0$. This proves the first bound. To show the second bound, we use the triangle inequality to get that

$$
\begin{aligned}
\mathrm{Var}[v_m(Z)]\ &=\ \|q_m^v(\Xi)+\mathbb{E}[v_m(X)]-v_m(Z)+\mathbb{E}[v_m(Z)]-q_m^v(\Xi)-\mathbb{E}[v_m(X)]\|_{L_2}^2 \\
&\leq\big(\|q_m^v(\Xi)+\mathbb{E}[v_m(X)]-v_m(Z)\|_{L_2}+\|q_m^v(\Xi)\|_{L_2}+|\mathbb{E}[v_m(Z)]-\mathbb{E}[v_m(X)]|\big)^2 \\
&=\Big(\|q_m^v(\Xi)+\mathbb{E}[v_m(X)]-v_m(Z)\|_{L_2}+\sqrt{\mathrm{Var}[q_m^v(\Xi)]} \\
&\qquad\qquad+\big|\mathbb{E}\big[v_m(Z)-q_m^v(\Xi)-\mathbb{E}[v_m(X)]\big]\big|\Big)^2 \\
&\leq\Big(\,2\|q_m^v(\Xi)+\mathbb{E}[v_m(X)]-v_m(Z)\|_{L_2}+\sqrt{\mathrm{Var}[q_m^v(\Xi)]}\,\Big)^2 \\
&\leq\mathrm{Var}[q_m^v(\Xi)]\big(1+(C_*)^m n^{-(1-\delta)/2}\big)^2\ =\ \mathrm{Var}[v_m(X)]\big(1+(C_*)^m n^{-(1-\delta)/2}\big)^2\,,
\end{aligned}
$$

where we have written $(C_*)^m=2C^{m/2}$. Similarly by a reverse triangle inequality,

$$
\begin{aligned}
\mathrm{Var}[v_m(Z)]\ &\geq\big(\sqrt{\mathrm{Var}[q_m^v(\Xi)]}-2\big\|q_m^v(\Xi)+\mathbb{E}[v_m(X)]-v_m(Z)\big\|_{L_2}\big)^2 \\
&\geq\mathrm{Var}[v_m(X)]\big(1-(C_*)^m n^{-(1-\delta)/2}\big)^2\,.
\end{aligned}
$$

□

C.6.2. Proof of Lemma 5.4

The first result holds directly by Theorem 4.1: There exists an absolute constant $C > 0$ such that

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}(\sigma^{-1} q_m^v(\mathbf{X}) \le t) - \mathbb{P}(\sigma^{-1} q_m^v(\Xi) \le t)\right| \le Cm \left(\frac{\sum_{i=1}^n M_{\nu;i}^\nu}{\sigma^\nu}\right)^{\frac{1}{\nu m+1}} . \quad \text{(C.17)}$$

We now show that, if $n \ge 2m^2$, there exists some absolute constant $C' > 0$ such that

$$\left(\frac{\sum_{i=1}^n M_{\nu;i}^\nu}{\sigma^\nu}\right)^{\frac{1}{\nu m+1}} \le C'\, n^{-\frac{\nu-2}{2\nu m+2}} \left(\frac{\sum_{k=1}^m \binom{n}{k}(\alpha_\nu(S,k))^2}{\sum_{k=1}^m \binom{n}{k}(\alpha_2(S,k))^2}\right)^{\frac{\nu}{2\nu m+2}} . \quad \text{(C.18)}$$

We proceed by induction. For $i \le n$ and $\mathcal{I}_0, \mathcal{I}_1 \subseteq [n]$ with $|\mathcal{I}_1| \le m$ and $\mathcal{I}_0 \cap \mathcal{I}_1 = \emptyset$, define

$$T(i,\mathcal{I}_0,\mathcal{I}_1) := \sum_{\substack{p_1+\ldots+p_n=m \\ 0\le p_l\le m \ \forall l\in[n] \\ p_{l'}\ge 1 \ \forall l'\in\mathcal{I}_1 \\ p_{l'}=0 \ \forall l'\in\mathcal{I}_0}} \binom{m}{p_1 \ \ldots \ p_n} \left\langle S\,,\, \overline{X_1^{\otimes p_1}} \otimes \ldots \otimes \overline{X_i^{\otimes p_i}} \otimes \xi_{(i+1)p_{i+1}} \otimes \ldots \otimes \xi_{np_n} \right\rangle ,$$

where $\binom{m}{p_1 \ldots p_n} := \frac{m!}{p_1!\ldots p_n!}$ is the multinomial coefficient. The quantities of interest become

$$\sigma = \sqrt{\mathrm{Var}[v_m(X)]} = \frac{\|T(n,\emptyset,\emptyset)\|_{L_2}}{n^m} , \quad M_{\nu;i} = \left\|\partial_i q_m^v(\mathbf{W}_i)^\top \mathbf{X}_i\right\|_{L_\nu} = \frac{\|T(i,\{i\},\emptyset)\|_{L_\nu}}{n^m} .$$

Write $\alpha_\nu(k) = \alpha_\nu(S,k)$ for short. We first show by induction on $|\mathcal{I}_1|$ that for $\nu \in [2,3]$

$$\left\|T(i,\mathcal{I}_0,\mathcal{I}_1)\right\|_{L_\nu}^2 \le (C_*)^m \sum_{k=0}^{m-|\mathcal{I}_1|} \frac{(C'')^{k+1}}{k!} (n - |\mathcal{I}_0| - |\mathcal{I}_1| + 1)^k (\alpha_\nu(|\mathcal{I}_1|+k))^2 , \quad \text{(C.19)}$$

$$\begin{aligned} \left\|T(i,\mathcal{I}_0,\mathcal{I}_1)\right\|_{L_2}^2 \ge \sum_{k=0}^{m-|\mathcal{I}_1|} \Bigg(& \frac{(c'')^{k+1}}{k!} (\alpha_2(|\mathcal{I}_1|+k))^2 \\ & \times \max\left\{(n - |\mathcal{I}_0| - |\mathcal{I}_1| + 1)^k \right. \\ & \left. - k^2 (n - |\mathcal{I}_0| - |\mathcal{I}_1| + 1)^{k-1}, 0\right\}\Bigg) . \end{aligned} \quad \text{(C.20)}$$

$C_* \ge 1$ is the supremum of the constant in the upper bound of Lemma C.4 over $\nu \in [2,3]$, C'' is the supremum of the constant in the first upper bound of Lemma A.4 over $\nu \in [2,3]$, and c'' is the constant in the lower bound of Lemma A.4 when $\nu = 2$.

For the base case, if $|\mathcal{I}_1| = n - |\mathcal{I}_0|$, i.e. $\mathcal{I}_0 = [n] \setminus \mathcal{I}_1$, by using the independence of $X_1, \ldots, X_n$ and applying Lemma C.4 to replace each ξ_i by $\left(\overline{X_i^{\otimes 1}}, \ldots, \overline{X_i^{\otimes m}}\right)$, we get that

$$\begin{aligned} \|T(i,\mathcal{I}_0,\mathcal{I}_1)\|_{L_2}^2 &= \left\| \sum_{\substack{\sum_{l\in\mathcal{I}_1} p_l = m \\ p_l \ge 1 \ \forall l\in\mathcal{I}_1}} \frac{m!}{\prod_{l\in\mathcal{I}_1}(p_l!)} \left\langle S\,,\, \bigotimes_{l\in\mathcal{I}_1} \overline{X_l^{\otimes p_l}} \right\rangle \right\|_{L_2}^2 \\ &\ge \left(\alpha_2(|\mathcal{I}_1|)\right)^2 . \end{aligned}$$

By applying Lemma C.4 again with $\nu \in [2,3]$, we get that

$$\begin{aligned}\|T(i,\mathcal{I}_0,\mathcal{I}_1)\|_{L_\nu}^2 &\le (C_*)^{|\mathcal{I}_1|}\Big\| \sum_{\substack{\sum_{l\in\mathcal{I}_1} p_l = m \\ p_l\ge 1 \ \forall l\in\mathcal{I}_1}} \frac{m!}{\prod_{l\in\mathcal{I}_1}(p_l!)} \Big\langle S\,,\, \bigotimes_{l\in\mathcal{I}_1} \overline{X_l^{\otimes p_l}} \Big\rangle \Big\|_{L_\nu}^2 \\ &\le (C_*)^m \big(\alpha_\nu(|\mathcal{I}_1|)\big)^2.\end{aligned}$$

Meanwhile, if $|\mathcal{I}_1| = m$, we automatically have $|\mathcal{I}_1| = n - |\mathcal{I}_0|$, so the above also holds.

For the inductive step, fix $\mathcal{I}_1 \subseteq [n]$ with $|\mathcal{I}_1| < \min\{n - |\mathcal{I}_0|, m\}$. Suppose (C.19) and (C.20) hold for all disjoint $\mathcal{I}_0', \mathcal{I}_1' \subseteq [n]$ with $|\mathcal{I}_1'| > |\mathcal{I}_1|$. For convenience, write $\mathbf{V}_j^{(i)} := \mathbf{X}_j$ for $j \le i$ and $\mathbf{V}_j^{(i)} := \xi_j$ for $j > i$. We also denote $V_{jp_j}^{(i)} = \overline{X_j^{\otimes p_j}}$ for $j \le i$ and $V_{jp_j}^{(i)} = \xi_{jp_j}$ for $j > i$.

Notice that by definition, $T(i,\mathcal{I}_0,\mathcal{I}_1)$ does not depend on $\mathbf{V}_j^{(i)}$ for $j \in \mathcal{I}_0$. We enumerate the elements of $[n] \setminus (\mathcal{I}_0 \cup \mathcal{I}_1)$ as $i_1 < \ldots < i_N$ for $N := n - |\mathcal{I}_0| - |\mathcal{I}_1|$. Also denote $\mathbf{V}_{\mathcal{I}}^{(i)} := \big\{\mathbf{V}_j^{(i)} \,\big|\, j \in \mathcal{I}\big\}$. Since $\mathbb{E}[T(i,\mathcal{I}_0,\mathcal{I}_1)] = 0$, we can define a martingale difference sequence by

$$D_0(i,\mathcal{I}_0,\mathcal{I}_1) \;:=\; \mathbb{E}\Big[T(i,\mathcal{I}_0,\mathcal{I}_1) \,\Big|\, \mathbf{V}_{\mathcal{I}_1}^{(i)}\Big]$$

and, for $1 \le j \le N$,

$$\begin{aligned}D_j(i,\mathcal{I}_0,\mathcal{I}_1) \;:=\; &\mathbb{E}\Big[T(i,\mathcal{I}_0,\mathcal{I}_1) \,\Big|\, \mathbf{V}_{\mathcal{I}_1}^{(i)}, \mathbf{V}_{i_1}^{(i)}, \ldots, \mathbf{V}_{i_j}^{(i)}\Big] \\ &- \mathbb{E}\Big[T(i,\mathcal{I}_0,\mathcal{I}_1) \,\Big|\, \mathbf{V}_{\mathcal{I}_1}^{(i)}, \mathbf{V}_{i_1}^{(i)}, \ldots, \mathbf{V}_{i_{j-1}}^{(i)}\Big].\end{aligned}$$

Then almost surely,

$$\bar{T}(i,\mathcal{I}_0,\mathcal{I}_1) \;=\; \sum_{j=0}^N \bar{D}_j(i,\mathcal{I}_0,\mathcal{I}_1)\,.$$

By applying Lemma A.4 with $\nu \in [2,3]$ followed by the triangle inequality, we get that

$$\big\|T(i,\mathcal{I}_0,\mathcal{I}_1)\big\|_{L_\nu}^2 \le C_*'' \Big\| \sum_{j=0}^N D_j(i,\mathcal{I}_0,\mathcal{I}_1)^2 \Big\|_{L_{\nu/2}} \le C'' \sum_{j=0}^N \|D_j(i,\mathcal{I}_0,\mathcal{I}_1)\|_{L_\nu}^2\,, \tag{C.21}$$

$$\big\|T(i,\mathcal{I}_0,\mathcal{I}_1)\big\|_{L_2}^2 \ge c'' \sum_{j=0}^N \|D_j(i,\mathcal{I}_0,\mathcal{I}_1)\|_{L_2}^2\,. \tag{C.22}$$

To control the martingale difference terms, recall that $\mathbb{E}\big[V_{jp_j}^{(i)}\big] = 0$ for all $i,j \in [n]$. Therefore

$$\begin{aligned}D_0(i,\mathcal{I}_0,\mathcal{I}_1) \;&=\; \sum_{\substack{p_1+\ldots+p_n=m \\ 0\le p_l\le m \ \forall l\in[n] \\ p_{l'}\ge 1 \ \forall l'\in\mathcal{I}_1 \\ p_{l'}=0 \ \forall l'\in\mathcal{I}_0}} \mathbb{E}\Big[\big\langle S\,,\, V_{1p_1}^{(i)} \otimes \ldots \otimes V_{np_n}^{(i)}\big\rangle \,\Big|\, \mathbf{V}_{\mathcal{I}_1}^{(i)}\Big] \\ &= T(i,\,[n]\setminus\mathcal{I}_1,\,\mathcal{I}_1)\,,\end{aligned}$$

i.e. there is no dependence on $\mathbf{V}_j^{(i)}$ for any $j \notin \mathcal{I}_1$. Similarly for $1 \le j \le N$,

$$\begin{aligned} D_j(i,\mathcal{I}_0,\mathcal{I}_1) \;&=\; \sum_{\substack{p_1+\ldots+p_n=m \\ 0\le p_l\le m \;\forall l\in[n] \\ p_{l'}\ge 1 \;\forall l'\in\mathcal{I}_1 \\ p_{l'}=0 \;\forall l'\in\mathcal{I}_0}} \Big(\mathbb{E}\Big[\big\langle S\,,\, V_{1p_1}^{(i)} \otimes \ldots \otimes V_{np_n}^{(i)} \big\rangle \,\Big|\, \mathbf{V}_{\mathcal{I}_1}^{(i)}, \mathbf{V}_{i_1}^{(i)}, \ldots, \mathbf{V}_{i_j}^{(i)} \Big] \\ &\qquad\qquad - \mathbb{E}\Big[\big\langle S\,,\, V_{1p_1}^{(i)} \otimes \ldots \otimes V_{np_n}^{(i)} \big\rangle \,\Big|\, \mathbf{V}_{\mathcal{I}_1}^{(i)}, \mathbf{V}_{i_1}^{(i)}, \ldots, \mathbf{V}_{i_{j-1}}^{(i)} \Big] \Big) \\ &= T(i,\mathcal{I}_0\cup\{i_{j+1},\ldots,i_N\},\mathcal{I}_1\cup\{i_j\})\;. \end{aligned}$$

By noting that $N = n - |\mathcal{I}_0| - |\mathcal{I}_1|$ and applying the inductive statement , we get that

$$\begin{aligned} &\textstyle\sum_{j=0}^N \|D_j(i,\mathcal{I}_0,\mathcal{I}_1)\|_{L_\nu}^2 \\ &= \big\|T(i,[n]\setminus\mathcal{I}_1,\mathcal{I}_1)\big\|_{L_\nu}^2 + \textstyle\sum_{j=1}^N \big\|T(i,\mathcal{I}_0\cup\{i_{j+1},\ldots,i_N\},\mathcal{I}_1\cup\{i_j\})\big\|_{L_\nu}^2 \\ &\le (C_*)^m(\alpha_\nu(|\mathcal{I}_1|))^2 + (C_*)^m \textstyle\sum_{j=1}^N \sum_{k=0}^{m-(|\mathcal{I}_1|+1)} \Big(\frac{(C'')^{k+1}}{k!} \times \\ &\qquad\qquad (n-(|\mathcal{I}_0|+N-j)-(|\mathcal{I}_1|+1)+1)^k(\alpha_\nu(|\mathcal{I}_1|+1+k))^2 \Big) \\ &= (C_*)^m(\alpha_\nu(|\mathcal{I}_1|))^2 + (C_*)^m \textstyle\sum_{j=1}^N \sum_{k=1}^{m-|\mathcal{I}_1|} \frac{(C'')^k}{(k-1)!} j^{k-1}(\alpha_\nu(|\mathcal{I}_1|+k))^2 \\ &= (C_*)^m(\alpha_\nu(|\mathcal{I}_1|))^2 + (C_*)^m \textstyle\sum_{k=1}^{m-|\mathcal{I}_1|} \frac{(C'')^k}{(k-1)!} \big(\sum_{j=1}^N j^{k-1} \big)(\alpha_\nu(|\mathcal{I}_1|+k))^2\;. \end{aligned}$$

Since $k-1 \ge 0$, we have

$$\textstyle\sum_{j=1}^N j^{k-1} \;\le\; \sum_{j=1}^N \int_j^{j+1} x^{k-1}dx \;=\; \sum_{j=1}^N \frac{(j+1)^k - j^k}{k} \;<\; \frac{(N+1)^k}{k}\;.$$

Substituting these into (C.21), we get that

$$\begin{aligned} \big\|T(i,\mathcal{I}_0,\mathcal{I}_1)\big\|_{L_\nu}^2 \;&\le C'' \textstyle\sum_{j=0}^N \|D_j(i,\mathcal{I}_0,\mathcal{I}_1)\|_{L_\nu}^2 \\ &\le C''(C_*)^m \textstyle\sum_{k=0}^{m-|\mathcal{I}_1|} \frac{(C'')^k}{k!}(N+1)^k(\alpha_\nu(|\mathcal{I}_1|+k))^2 \\ &= (C_*)^m \textstyle\sum_{k=0}^{m-|\mathcal{I}_1|} \frac{(C'')^{k+1}}{k!}(n-|\mathcal{I}_0|-|\mathcal{I}_1|+1)^k(\alpha_\nu(|\mathcal{I}_1|+k))^2\;, \end{aligned}$$

which finishes the induction for (C.19). Similarly for (C.20), using the inductive statement gives

$$\begin{aligned} &\textstyle\sum_{j=0}^N \|D_j(i,\mathcal{I}_0,\mathcal{I}_1)\|_{L_2}^2 \\ &= \big\|T(i,[n]\setminus\mathcal{I}_1,\mathcal{I}_1)\big\|_{L_2}^2 + \textstyle\sum_{j=1}^N \big\|T(i,\mathcal{I}_0\cup\{i_{j+1},\ldots,i_N\},\mathcal{I}_1\cup\{i_j\})\big\|_{L_2}^2 \\ &\ge (\alpha_2(|\mathcal{I}_1|))^2 + \sum_{j=1}^N \sum_{k=0}^{m-(|\mathcal{I}_1|+1)} \Big(\frac{(c'')^{k+1}}{k!}(\alpha_2(|\mathcal{I}_1|+1+k))^2 \max\{j^k - k^2 j^{k-1}\,,\,0\} \Big) \\ &= (\alpha_2(|\mathcal{I}_1|))^2 + \sum_{k=1}^{m-|\mathcal{I}_1|} \frac{(c'')^k}{(k-1)!} \left(\sum_{j=1}^N \max\left\{ j^{k-1} - (k-1)^2 j^{k-2}\,,\,0 \right\} \right) (\alpha_2(|\mathcal{I}_1|+k))^2\;. \end{aligned}$$

Since $k-1 \geq 0$, we have

$$\begin{aligned}
&\textstyle\sum_{j=1}^{N} \max\left\{j^{k-1} - (k-1)^2 j^{k-2}\,,\, 0\right\} \\
&\geq\ \max\left\{\textstyle\sum_{j=1}^{N}\left(j^{k-1} - (k-1)^2 j^{k-2}\right),\, 0\right\} \\
&=\ \max\left\{\textstyle\sum_{j=1}^{N+1} j^{k-1} - (N+1)^{k-1} - \sum_{j=1}^{N}(k-1)^2 j^{k-2}\,,\, 0\right\} \\
&\geq\ \max\left\{\textstyle\sum_{j=1}^{N+1}\int_{j-1}^{j} x^{k-1}dx - (N+1)^{k-1} - \sum_{j=1}^{N}(k-1)^2\int_{j}^{j+1} x^{k-2}dx\,,\, 0\right\} \\
&=\ \max\left\{\sum_{j=1}^{N+1}\frac{j^k-(j-1)^k}{k} - (N+1)^{k-1} - \sum_{j=1}^{N}(k-1)((j+1)^{k-1}-j^{k-1})\,,\, 0\right\} \\
&>\ \max\left\{\frac{(N+1)^k}{k} - (N+1)^{k-1} - (k-1)(N+1)^{k-1}\,,\, 0\right\} \\
&=\ \frac{1}{k}\max\left\{(N+1)^k - k^2(N+1)^{k-1}\,,\, 0\right\},
\end{aligned}$$

Substituting these into (C.22), we get that

$$\begin{aligned}
\left\|T(i,\mathcal{I}_0,\mathcal{I}_1)\right\|_{L_2}^2 \ &\geq\ c''\textstyle\sum_{j=0}^{N}\|D_j(i,\mathcal{I}_0,\mathcal{I}_1)\|_{L_2}^2 \\
&\geq c''\textstyle\sum_{k=0}^{m-|\mathcal{I}_1|}\frac{(c'')^k}{k!}\max\{(N+1)^k - k^2(N+1)^{k-1},0\}(\alpha_2(|\mathcal{I}_1|+k))^2 \\
&=\ \textstyle\sum_{k=0}^{m-|\mathcal{I}_1|}\Big(\frac{(c'')^{k+1}}{k!}(\alpha_2(|\mathcal{I}_1|+k))^2 \\
&\qquad\qquad \times\max\left\{(n-|\mathcal{I}_0|-|\mathcal{I}_1|+1)^k - k^2(n-|\mathcal{I}_0|-|\mathcal{I}_1|+1)^{k-1},0\right\}\Big)\,.
\end{aligned}$$

This finishes the induction for (C.20). In particular, we can now obtain

$$\begin{aligned}
\frac{\sum_{i=1}^{n} M_{\nu;i}^{\nu}}{\sigma^{\nu}} \ &=\ \frac{\sum_{i=1}^{n}\|T(i,\{i\},\emptyset)\|_{L_\nu}^{\nu}}{\|T(n,\emptyset,\emptyset)\|_{L_2}^{\nu}} \\
&\leq \frac{n\Big((C_*)^m\sum_{k=0}^{m-1}\frac{(C'')^{k+1}(\alpha_\nu(k+1))^2}{k!}n^k\Big)^{\frac{\nu}{2}}}{\Big(\sum_{k=0}^{m}\frac{(c'')^{k+1}(\alpha_2(k))^2}{k!}\max\left\{(n+1)^k - k^2(n+1)^{k-1},0\right\}\Big)^{\frac{\nu}{2}}}\,.
\end{aligned}$$

Since $n \geq 2m^2$ by assumption, we have

$$(n+1)^k - k^2(n+1)^{k-1} \ \geq (n+1)^{k-1}(n+1-m^2)\ \geq\ \frac{1}{2}(n+1)^k\,.$$

By further noting that $\alpha_2(0)=0$, we can simplify the ratio as

$$\begin{aligned}
\frac{\sum_{i=1}^{n} M_{\nu;i}^{\nu}}{\sigma^{\nu}} \ &\leq \frac{2^{\nu}(C_*)^{\frac{m\nu}{2}}n\Big(\sum_{k=0}^{m-1}\frac{(C'')^{k+1}}{k!}n^k(\alpha_\nu(k+1))^2\Big)^{\nu/2}}{\Big(\sum_{k=0}^{m}\frac{(c'')^{k+1}}{k!}(n+1)^k(\alpha_2(k))^2\Big)^{\nu/2}} \\
&= \frac{2^{\nu}(C_*)^{\frac{m\nu}{2}}n\Big(\sum_{k=1}^{m}\frac{(C'')^{k}}{(k-1)!}n^{k-1}(\alpha_\nu(k))^2\Big)^{\nu/2}}{\Big(\sum_{k=1}^{m}\frac{(c'')^{k+1}}{k!}(n+1)^k(\alpha_2(k))^2\Big)^{\nu/2}} \\
&= \frac{2^{\nu}(C_*)^{\frac{m\nu}{2}}}{n^{\frac{\nu-2}{2}}}\left(\frac{\sum_{k=1}^{m}\frac{(C'')^{k}}{(k-1)!}n^k(\alpha_\nu(k))^2}{\sum_{k=1}^{m}\frac{(c'')^{k+1}}{k!}(n+1)^k(\alpha_2(k))^2}\right)^{\frac{\nu}{2}}\,.
\end{aligned}$$

Meanwhile by Stirling's approximation, $\frac{n^k}{(k-1)!} \le A^m \binom{n}{k}$ and $\frac{(n+1)^k}{k!} \ge \frac{n^k}{k!} \ge B^m \binom{n}{k}$ for some absolute constants $A, B > 0$. Since $\nu < 3$, we get that for some absolute constant $C' > 0$,

$$\left(\frac{\sum_{i=1}^n M_{\nu;i}^\nu}{\sigma^\nu}\right)^{\frac{1}{\nu m+1}} \le C' n^{-\frac{\nu-2}{2\nu m+2}} \left(\frac{\sum_{k=1}^m \binom{n}{k}(\alpha_\nu(S,k))^2}{\sum_{k=1}^m \binom{n}{k}(\alpha_2(S,k))^2}\right)^{\frac{\nu}{2\nu m+2}} = C' n^{-\frac{\nu-2}{2\nu m+2}} \beta_{m,\nu}^{\frac{\nu}{2\nu m+2}} .$$

By applying (C.19) and (C.20) to $\sigma^2 = \mathrm{Var}[v_m(X)]$ and using the same argument as above, we also get that if $n \ge 2m^2$, there exist some absolute constants $C_1, C_2 > 0$ such that

$$\frac{(C_1)^m}{n^{2m}} \sum_{k=1}^m \binom{n}{k} (\alpha_2(S,k))^2 \le \mathrm{Var}[v_m(X)] \le \frac{(C_2)^m}{n^{2m}} \sum_{k=1}^m \binom{n}{k} (\alpha_2(S,k))^2 . \quad \square$$

C.6.3. Proof of Proposition 5.2

Recall that q_m^v is the multilinear representation of v_m, and WLOG assume the coupling between ξ_i and Z_i considered in Lemma 5.3. By the exact same argument as the proof of Theorem 4.7 leading up to (C.14) (δ-regularity, Lemma A.11, a Markov's inequality and choosing ϵ appropriately), we get that for some absolute constant $C_1 > 0$,

$$\left|\mathbb{P}\big(\sigma^{-1}(q_m^v(\Xi) + \mathbb{E}[v_m(X)]) \le t\big) - \mathbb{P}\big(\sigma^{-1} v_m(Z) \le t\big)\right| \le C_1 m n^{-\frac{1-\delta}{2m+1}} . \tag{C.23}$$

By applying Lemma 5.4 to replace $q_m^v(\Xi)$ by $q_m^v(\mathbf{X})$, noting that $v_m(X) = q_m^v(\mathbf{X}) + \mathbb{E}[v_m(X)]$ and using the triangle inequality to combine the bounds, we get that for some absolute constant $C' > 0$,

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}(\sigma^{-1} v_m(X) \le t) - \mathbb{P}(\sigma^{-1} v_m(Z) \le t)\right| \le C_1 m n^{-\frac{1-\delta}{2m+1}} + C' m \Delta_\delta .$$

To prove the final bound, we first note that by Lemma 5.4, (C.23) and the triangle inequality,

$$\begin{aligned}
&\sup_{t\in\mathbb{R}} \left|\mathbb{P}(\sigma^{-1}(v_m(X) - \mathbb{E}[v_m(X)]) \le t) - \mathbb{P}(\sigma^{-1}(v_m(Z) - \mathbb{E}[v_m(Z)]) \le t)\right| \\
&\le C' m \Delta_\delta + \sup_{t\in\mathbb{R}} \left|\mathbb{P}(\sigma^{-1} q_m^v(\Xi) \le t) - \mathbb{P}(\sigma^{-1}(v_m(Z) - \mathbb{E}[v_m(Z)]) \le t)\right| \\
&\le C' m \Delta_\delta + C_1 m n^{-\frac{1-\delta}{2m+1}} \\
&\qquad + \sup_{t\in\mathbb{R}} \left|\mathbb{P}(\sigma^{-1} q_m^v(\Xi) \le t) - \mathbb{P}(\sigma^{-1}(q_m^v(\Xi) + \mathbb{E}[v_m(X)] - \mathbb{E}[v_m(Z)]) \le t)\right| .
\end{aligned}$$

By the same Carbery-Wright inequality argument used in the proof of Theorem 4.1 in (C.7), the last term can be bounded by

$$\begin{aligned}
&\sup_{t\in\mathbb{R}} \mathbb{P}\big(|\sigma^{-1} q_m^v(\Xi) - t| \le \sigma^{-1} |\mathbb{E}[v_m(X) - v_m(Z)]|\big) \\
&\le C_2 m \frac{|\sigma^{-1}\mathbb{E}[v_m(X)] - \sigma^{-1}\mathbb{E}[v_m(Z)]|^{1/m}}{\mathbb{E}[(\sigma^{-1} q_m^v(\Xi) - t)^2]^{1/2m}} = C_2 m \sigma^{-1/m} \left|\mathbb{E}[v_m(X)] - \mathbb{E}[v_m(Z)]\right|^{1/m} \\
&= C_2 m \sigma^{-1/m} \left|\mathbb{E}\big[q_m^v(\Xi) + \mathbb{E}[v_m(X)] - v_m(Z)\big]\right|^{1/m}
\end{aligned}$$

$$\leq\ C_2 m\left(\frac{\left\|q_m^v(\Xi)+\mathbb{E}[v_m(X)]-v_m(Z)\right\|_{L_2}^2}{\mathrm{Var}[q_m^v(\Xi)]}\right)^{1/(2m)}\ \leq\ C_3 m n^{-\frac{1-\delta}{2m}}\ \leq\ C_3 m n^{-\frac{1-\delta}{2m+1}}$$

for some absolute constants $C_2, C_3 > 0$; in the last line, we noted that $\sigma^2 = \mathrm{Var}[v_m(X)] = \mathrm{Var}[q_m^v(\Xi)]$ and applied Lemma 5.3 again. By taking $C = C_1 + C_3 > 0$, we get that

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}(\sigma^{-1}(v_m(X)-\mathbb{E}[v_m(X)])\leq t)-\mathbb{P}(\sigma^{-1}(v_m(Z)-\mathbb{E}[v_m(Z)])\leq t)\right| \leq\ C m n^{-\frac{1-\delta}{2m+1}} + C' m \Delta_\delta\,. \quad \square$$

C.7 Proofs for Sections 5.2, 5.3 and 5.4

C.7.1. Proof of Proposition 5.5

Since g is $(m+1)$-times continuously differentiable, by an m-th order Taylor expansion around $\mathbb{E}X_1$, we have that almost surely

$$\begin{aligned}
&\left|\hat{g}(X)-\textstyle\sum_{l=0}^m \mu_l-\left(\hat{g}_m^{(\mathbb{E}X_1)}(X)-\mu_m\right)\right| \\
&=\left|\mu_0+\textstyle\sum_{j=1}^m \hat{g}_j^{(\mathbb{E}X_1)}(X)\ +\ (m+1)\mathbb{E}\Big[(1-\Theta)^{m+1}\,\hat{g}_{m+1}^{\left(\mathbb{E}X_1+\Theta n^{-1}\sum_{i\leq n}\bar{X}_i\right)}(X)\,\Big|\,X\Big]\right. \\
&\qquad\left.-\textstyle\sum_{l=0}^m \mu_l-\left(\hat{g}_m^{(\mathbb{E}X_1)}(X)-\mu_m\right)\right| \\
&=\left|\textstyle\sum_{l=1}^{m-1}\left(\hat{g}_l^{(\mathbb{E}X_1)}(X)-\mu_l\right)\ +\ (m+1)\mathbb{E}\Big[(1-\Theta)^{m+1}\,\hat{g}_{m+1}^{\left(\mathbb{E}X_1+\Theta n^{-1}\sum_{i\leq n}\bar{X}_i\right)}(X)\,\Big|\,X\Big]\right| \\
&=:R\,.
\end{aligned}$$

Now let $\mathbf{X}_i = (\bar{X}_i, \ldots, \overline{X}_i^{\otimes m})$ and ξ_i be defined as in Theorem 4.2, f_1 be the multilinear representation of $\hat{g}_m^{(\mathbb{E}X_1)}$ and denote $\sigma_m \coloneqq \mathrm{Var}\big[\hat{g}_m^{(\mathbb{E}X_1)}(X)\big] = \mathrm{Var}\big[f_1(\mathbf{X})\big]$. Then by Theorem 4.2, there is some absolute constant $C' > 0$,

$$\begin{aligned}
&\sup_{t\in\mathbb{R}}\left|\mathbb{P}\big(\hat{g}(X)-\textstyle\sum_{l=0}^m\mu_l\leq t\big)-\mathbb{P}(f_1(\Xi)\leq t)\right| \\
&\leq C' m\left(\left(\frac{\|R\|_{L_2}^2}{\sigma_m^2}\right)^{\frac{1}{2m+1}}\right. \\
&\qquad\left.+\left(\frac{\sum_{i=1}^n\left\|\partial_i f_1(\mathbf{X}_1,\ldots,\mathbf{X}_{i-1},\mathbb{E}[\mathbf{X}_1],\xi_{i+1},\ldots,\xi_n)(\mathbf{X}_i-\mathbb{E}[\mathbf{X}_1])\right\|_{L_\nu}^\nu}{\sigma_m^\nu}\right)^{\frac{1}{\nu m+1}}\right).
\end{aligned}$$

The first term can be controlled by applying the triangle inequality twice, using the Jensen's inequality and noting the definition of ϵ_m:

$$\begin{aligned}
&\frac{\|R\|_{L_2}^2}{\sigma_m^2} \\
&\leq\ \frac{\left(\left\|\sum_{l=1}^{m-1}\left(\hat{g}_l^{(\mathbb{E}X_1)}(X)-\mu_l\right)\right\|_{L_2}+(m+1)\left\|\mathbb{E}\Big[(1-\Theta)^{m+1}\,\hat{g}_{m+1}^{\left(\mathbb{E}X_1+\Theta n^{-1}\sum_{i\leq n}\bar{X}_i\right)}(X)\,\Big|\,X\Big]\right\|_{L_2}\right)^2}{\sigma_m^2}
\end{aligned}$$

$$\leq \frac{2\mathrm{Var}\Big[\sum_{l=1}^{m-1}\hat{g}_l^{(\mathbb{E}X_1)}(X)\Big]}{\sigma_m^2} + \frac{2(m+1)^2\Big\|\mathbb{E}\Big[(1-\Theta)^{m+1}\,\hat{g}_{m+1}^{\left(\mathbb{E}X_1+\Theta n^{-1}\sum_{i\leq n}\bar{X}_i\right)}(X)\Big|X\Big]\Big\|_{L_2}^2}{\sigma_m^2}$$
$$= 2\epsilon_m\,.$$

By noting that Lemma 5.4 applies to $\hat{g}_m^{(\mathbb{E}X_1)}(X)$, the second term can be bounded by

$$C''\,n^{-\frac{\nu-2}{2\nu m+2}}\left(\frac{\sum_{k=1}^m\binom{n}{k}(\alpha_{m,\nu}(S))^2}{\sum_{k=1}^m\binom{n}{k}(\alpha_{m,2}(k))^2}\right)^{\frac{\nu}{2\nu m+2}} \;=\; C''n^{-\frac{\nu-2}{2\nu m+2}}\beta_{m,\nu}^{\frac{\nu}{2\nu m+2}}$$

for some absolute constant $C''>0$. Moreover by the δ-regularity assumption, the first bound of Proposition 5.2 implies that for some absolute constant $C'''>0$,

$$\sup_{t\in\mathbb{R}}\Big|\,\mathbb{P}(f_1(\Xi)\leq t)-\mathbb{P}\big(\hat{g}_m^{(\mathbb{E}X_1)}(Z)-\mathbb{E}[\hat{g}_m^{(\mathbb{E}X_1)}(Z)]\leq t\big)\,\Big| \;\leq\; C'''mn^{-\frac{1-\delta}{2m+1}}\,.$$

Combining all three bounds gives that, for some absolute constant $C>0$,

$$\sup_{t\in\mathbb{R}}\Big|\mathbb{P}\big(\hat{g}(X)-\textstyle\sum_{l=0}^m\mu_l\leq t\big)-\mathbb{P}\big(\hat{g}_m^{(\mathbb{E}X_1)}(Z)-\mathbb{E}[\hat{g}_m^{(\mathbb{E}X_1)}(Z)]\leq t\big)\Big|$$
$$\leq Cm\big(\epsilon_m^{\frac{1}{2m+1}}+n^{-\frac{1-\delta}{2m+1}}+n^{-\frac{\nu-2}{2\nu m+2}}\,\beta_{m,\nu}^{\frac{\nu}{2\nu m+2}}\big)\,.$$

To prove the fourth moment bound, applying Proposition 4.6 and Remark 4.3 imply that

$$\sup_{t\in\mathbb{R}}\Big|\mathbb{P}\big(\hat{g}_m^{(\mathbb{E}X_1)}(Z)-\mathbb{E}[\hat{g}_m^{(\mathbb{E}X_1)}(Z)]\leq t\big)-\Phi(\sigma_Z^{-1}t)\Big| \;\leq\; \Big(\frac{4m-4}{3m}\Big|\mathrm{Kurt}\big[\hat{g}_m^{(\mathbb{E}X_1)}(Z)\big]\Big|\Big)^{1/2}\,,$$

where $\sigma_Z^2 \coloneqq \mathrm{Var}\big[\hat{g}_m^{(\mathbb{E}X_1)}(Z)\big]$. Write $\sigma_X^2 \coloneqq \mathrm{Var}\big[\hat{g}_m^{(\mathbb{E}X_1)}(X)\big]$. Note that if $t=0$,

$$\big|\Phi(\sigma_Z^{-1}t)-\Phi(\sigma_X^{-1}t)\big| \;=\; 0\,.$$

Suppose $t\neq 0$ and write $\eta\sim\mathcal{N}(0,1)$. By Fact 4.4, we have that

$$\big|\Phi(\sigma_Z^{-1}t)-\Phi(\sigma_X^{-1}t)\big| \;\leq C'\,\frac{\big|\sigma_Z^{-1}t-\sigma_X^{-1}t\big|/2}{\big(\mathbb{E}\big[(\eta-(\sigma_Z^{-1}t+\sigma_X^{-1}t)/2)^2\big]\big)^{1/2}} \;\leq\; C'\,\frac{\big|\sigma_Z^{-1}t-\sigma_X^{-1}t\big|}{\big|\sigma_Z^{-1}t+\sigma_X^{-1}t\big|}$$

for some absolute constant $C'>0$. Rearranging and applying the second bound of Lemma 5.3,

$$\big|\Phi(\sigma_Z^{-1}t)-\Phi(\sigma_X^{-1}t)\big| \;\leq C'\,\frac{\big|1-\sigma_X^{-1}\sigma_Z\big|}{1+\sigma_X^{-1}\sigma_Z} \;\leq\; C'\frac{(C'')^m\,n^{-(1-\delta)/2}}{2-(C'')^m\,n^{-(1-\delta)/2}}$$

for some absolute constant $C''>0$. Applying the triangle inequality and replacing t with $\sigma_X t$, we obtain the final bound that

$$\sup_{t\in\mathbb{R}}\Big|\mathbb{P}\Big(\sigma_X^{-1}\big(\hat{g}(X)-\textstyle\sum_{l=0}^m\mu_l\big)\leq t\Big)-\Phi(t)\Big|$$
$$\leq \Delta_{\hat{g},m}+C'\frac{(C'')^m\,n^{-(1-\delta)/2}}{2-(C'')^m\,n^{-(1-\delta)/2}}+\Big(\frac{4m-4}{3m}\Big|\mathrm{Kurt}\big[\hat{g}_m^{(\mathbb{E}X_1)}(Z)\big]\Big|\Big)^{1/2}.\qquad\square$$

C.7.2. Proof of Proposition 5.6

Fix $K \in \mathbb{N}$. Denote

$$\mu_k \coloneqq \mathbb{E}[\phi_k^{(K)}(Y_1)] \in \mathbb{R}\,, \quad \bar{u} \coloneqq \mathbb{E}[u(Y_1,\dots,Y_m)] \quad \text{and} \quad V^K \coloneqq \{V_1^K,\dots,V_n^K\}\,,$$
$$\text{where} \quad V_i^K \coloneqq (\phi_1^{(K)}(Y_i)-\mu_1,\dots,\phi_K^{(K)}(Y_i)-\mu_K)\,.$$

By Hoeffding's decomposition (5.6) and writing $j \neq M$ in short for $j \in [m]\setminus\{M\}$, we have

$$u_m(Y)-\bar{u} \;=\; \binom{m}{M} U_M^{(K)}(V^K) + \sum\nolimits_{j\neq M}\binom{m}{j} U_j^{\mathrm{H}}(Y) + \binom{m}{M}\Big(U_M^{\mathrm{H}}(Y) - U_M^{(K)}(V^K)\Big)\,.$$

By Theorem 4.2, we get that for some absolute constant $C'>0$ and for every $t\in\mathbb{R}$,

$$\begin{aligned}
&\Big|\mathbb{P}\big(u_m(Y)-\bar{u}\le t\big) - \mathbb{P}\Big(\binom{m}{M}U_M^{(K)}(\Xi^{(K)})\le t\Big)\Big| \\
&\le C'm\left(\frac{\sum_{i=1}^n \big\|\partial_i U_M^{(K)}\big(V_1^K,\dots,V_{i-1}^K,\mathbf{0},\xi_{i+1}^{(K)},\dots,\xi_n^{(K)}\big)V_i^K\big\|_{L_\nu}^\nu}{\big\|U_M^{(K)}(V^K)\big\|_{L_2}^\nu}\right)^{\frac{1}{\nu M+1}} \\
&\quad + C'm\left(\frac{\Big\|\sum_{j\neq M}\binom{m}{j}U_j^{\mathrm{H}}(Y)+\binom{m}{M}\Big(U_M^{\mathrm{H}}(Y)-U_M^{(K)}(V^K)\Big)\Big\|_{L_2}^2}{\Big\|\binom{m}{M}U_M^{(K)}(V^K)\Big\|_{L_2}^2}\right)^{\frac{1}{2M+1}} =: C'm(R_1+R_2)\,.
\end{aligned}$$

Before proceeding, first note that $U_M^{(K)}(V^K)$ is a U-statistic with the kernel

$$u_M^{(K)}(v_1,\dots,v_j) \;=\; \sum\nolimits_{k_1,\dots,k_m=1}^K \lambda^{(K)}_{k_1\dots k_m} v_{1k_1}\dots v_{Mk_M} \;=\; \langle T_M, v_1\otimes\dots\otimes v_M\rangle$$

for some deterministic tensor $T_M\in\mathbb{R}^{K^M}$, and where V_i^K's are zero-mean. By Lemma C.11,

$$\|U_M^{(K)}(V^K)\|_{L_2}^2 \;\ge\; \frac{1}{n^M}\big\|u_m^{(K)}(V_1^K,\dots,V_M^K)\big\|_{L_2}^2\,.$$

We defer to Lemma C.13 to show that for some absolute constant $C_*>0$,

$$\Big\|\partial_i U_M^{(K)}\Big(V_1^K,\dots,V_{i-1}^K,\mathbf{0},\xi_{i+1}^{(K)},\dots,\xi_n^{(K)}\Big)V_i^K\Big\|_{L_\nu}^\nu \le \frac{C_*^m\big\|u_m^{(K)}(V_1^K,\dots,V_M^K)\big\|_{L_\nu}^\nu}{n^{(M+1)\nu/2}}\,. \tag{C.24}$$

Let $Y_1',\dots,Y_m'$ be i.i.d. copies of Y_1. Now by the definition of V_i^K, the binomial theorem, the triangle inequality and Jensen's inequality, the truncation error satisfies

$$\begin{aligned}
&\Big\|U_M^{\mathrm{H}}(Y)-U_M^{(K)}(V^K)\Big\|_{L_\nu} \;\le\; \Big\|u_M^{\mathrm{H}}(Y_1,\dots,Y_M)-u_M^{(K)}(V_1^K,\dots,V_M^K)\Big\|_{L_\nu} \\
&=\Big\|\sum\nolimits_{r=0}^M(-1)^{M-r}\sum\nolimits_{1\le l_1<\dots<l_r\le M}\mathbb{E}\big[u(Y_{l_1},\dots,Y_{l_r},Y_1',\dots,Y_{m-r}')\big|Y_1,\dots,Y_M\big] \\
&\quad -\sum\nolimits_{k_1,\dots,k_m=1}^K\lambda^{(K)}_{k_1\dots k_m}(\phi_{k_1}^{(K)}(Y_i)-\mu_{k_1})\dots(\phi_{k_M}^{(K)}(Y_M)-\mu_{k_M})\mu_{k_{M+1}}\dots\mu_{k_m}\Big\|_{L_\nu} \\
&=\Big\|\sum\nolimits_{r=0}^M(-1)^{M-r}\sum\nolimits_{1\le l_1<\dots<l_r\le M}\mathbb{E}\big[u(Y_{l_1},\dots,Y_{l_r},Y_1',\dots,Y_{m-r}')\big|Y_1,\dots,Y_M\big] \\
&\quad -\sum\nolimits_{r=0}^M(-1)^{M-r}\sum\nolimits_{1\le l_1<\dots<l_r\le M}
\end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad \Big(\textstyle\sum_{k_1,\ldots,k_m=1}^{K} \lambda^{(K)}_{k_1\ldots k_m} \phi_{k_{l_1}}(Y_{l_1}) \ldots \phi_{k_{l_r}}(Y_{l_r}) \prod_{\substack{k'\in[M]\\ k'\neq l_s\,\forall s}} \mu_{k'} \Big) \Big\|_{L_\nu} \\
&\leq \textstyle\sum_{r=0}^{M} \binom{M}{r} \Big\| u(Y_{l_1},\ldots,Y_{l_r},Y'_1,\ldots,Y'_{m-r}) \\
&\qquad\qquad - \textstyle\sum_{k_1,\ldots,k_m=1}^{K} \Big(\lambda^{(K)}_{k_1\ldots k_m} \phi_{k_{l_1}}(Y_{l_1}) \ldots \phi_{k_{l_r}}(Y_{l_r}) \prod_{\substack{k'\in[M]\\ k'\neq l_s\,\forall s}} \phi_{k'}(Y'_{k'}) \Big) \Big\|_{L_\nu} \\
&\overset{(a)}{=} 2^M \epsilon_{K;\nu} \,.
\end{aligned}
$$

In (a) above, we have noted that for any permutation π on $\{1,\ldots,m\}$, the L_ν approximation error satisfies

$$
\begin{aligned}
\epsilon_{K;\nu} \;&=\; \big\| \textstyle\sum_{k_1,\ldots,k_m=1}^{K} \lambda^{(K)}_{k_1\ldots k_m} \phi^{(K)}_{k_1}(Y_1) \times \ldots \times \phi^{(K)}_{k_m}(Y_m) - u(Y_1,\ldots,Y_m) \big\|_{L_\nu} \\
&\overset{(b)}{=} \big\| \textstyle\sum_{k_1,\ldots,k_m=1}^{K} \lambda^{(K)}_{k_1\ldots k_m} \phi^{(K)}_{k_1}(Y_{\pi(1)}) \times \ldots \times \phi^{(K)}_{k_m}(Y_{\pi(m)}) - u(Y_{\pi(1)},\ldots,Y_{\pi(m)}) \big\|_{L_\nu} \\
&\overset{(c)}{=} \big\| \textstyle\sum_{k_1,\ldots,k_m=1}^{K} \lambda^{(K)}_{k_1\ldots k_m} \phi^{(K)}_{k_{\pi^{-1}(1)}}(Y_1) \times \ldots \times \phi^{(K)}_{k_{\pi^{-1}(m)}}(Y_m) - u(Y_{\pi(1)},\ldots,Y_{\pi(m)}) \big\|_{L_\nu} \\
&\overset{(d)}{=} \big\| \textstyle\sum_{k_1,\ldots,k_m=1}^{K} \lambda^{(K)}_{k_1\ldots k_m} \phi^{(K)}_{k_{\pi^{-1}(1)}}(Y_1) \times \ldots \times \phi^{(K)}_{k_{\pi^{-1}(m)}}(Y_m) - u(Y_1,\ldots,Y_m) \big\|_{L_\nu} \,,
\end{aligned}
$$

where we have used that $Y_1,\ldots,Y_m$ are i.i.d. in (b), the commutativity of scalar product in (c) and that u is symmetric in (d). Combining the bounds above, we get that there is some absolute constant $C'' > 0$ such that

$$
\begin{aligned}
R_1 \;&\leq\; C'' n^{-\frac{\nu-2}{2(\nu M+1)}} \left(\frac{\big\| u^{(K)}_M(V^K_1,\ldots,V^K_M) \big\|_{L_\nu}}{\big\| u^{(K)}_M(V^K_1,\ldots,V^K_M) \big\|_{L_2}} \right)^{\frac{\nu}{\nu M+1}} \\
&\leq\; C'' n^{-\frac{\nu-2}{2(\nu M+1)}} \left(\frac{\big\| u^{\mathrm{H}}_M(Y_1,\ldots,Y_M) \big\|_{L_\nu} + (2^M)\,\epsilon_{K;\nu}}{\big\| u^{\mathrm{H}}_M(Y_1,\ldots,Y_M) \big\|_{L_2} - (2^M)\,\epsilon_{K;\nu}} \right)^{\frac{\nu}{\nu M+1}} .
\end{aligned}
$$

Denote $\sigma_j^2 \coloneqq \operatorname{Var} \mathbb{E}[u(Y_1,\ldots,Y_m) \,|\, Y_1,\ldots,Y_j]$. By the truncation error bound again, we have

$$
\begin{aligned}
R_2 \;&\leq\; \left(\frac{2 \big\| \sum_{j\neq M} \binom{m}{j} U^{\mathrm{H}}_j(Y) \big\|^2_{L_2} + \binom{m}{M}^2 2^{M+1} \epsilon^2_{K;\nu}}{\big\| \binom{m}{M} U^{\mathrm{H}}_M(Y) \big\|^2_{L_2} - \binom{m}{M}^2 2^{M+1} \epsilon^2_{K;\nu}} \right)^{\frac{1}{2M+1}} \\
&\overset{(a)}{=} \left(\frac{2 \sum_{j\neq M} \binom{m}{j}^2 \| U^{\mathrm{H}}_j(Y) \|^2_{L_2} + \binom{m}{M}^2 2^{M+1} \epsilon^2_{K;\nu}}{\binom{m}{M}^2 \| U^{\mathrm{H}}_M(Y) \|^2_{L_2} - \binom{m}{M}^2 2^{M+1} \epsilon^2_{K;\nu}} \right)^{\frac{1}{2M+1}} \\
&\overset{(b)}{=} \left(\frac{2 \sum_{j\neq M} \binom{m}{j}^2 \binom{n}{j}^{-1} \sigma^2_j + \binom{m}{M}^2 2^{M+1} \epsilon^2_{K;\nu}}{\binom{m}{M}^2 \binom{n}{M}^{-1} \sigma^2_M - \binom{m}{M}^2 2^{M+1} \epsilon^2_{K;\nu}} \right)^{\frac{1}{2M+1}} \\
&\overset{(c)}{=} \left(\frac{2 \sum_{j\neq M} \sigma^2_{m,n;j} + \binom{m}{M}^2 2^{M+1} \epsilon^2_{K;\nu}}{\sigma^2_{m,n;M} - \binom{m}{M}^2 2^{M+1} \epsilon^2_{K;\nu}} \right)^{\frac{1}{2M+1}} .
\end{aligned}
$$

In (a), we have used the orthogonality of the degenerate U-statistics of different degrees by Lemma C.9; in (b), we have used the variance formula of $U^{\mathrm{H}}_j(Y)$ in Lemma C.11;

in (c), we have plugged in the definition of $\sigma^2_{m,n;j}$. Combining the bounds and taking $K \to \infty$, we get

$$\begin{aligned}
&\Big|\mathbb{P}\big(u_m(Y)-\bar{u} \le t\big) - \lim_{K\to\infty} \mathbb{P}\Big(\binom{m}{M} \tilde{U}^{(K)}_M(\Xi^{(K)}) \le t\Big)\Big| \\
&\qquad \le Cmn^{-\frac{\nu-2}{2(\nu M+1)}} \left(\frac{\big\|u^{\mathrm{H}}_M(Y_1,\dots,Y_M)\big\|_{L_\nu}}{\big\|u^{\mathrm{H}}_M(Y_1,\dots,Y_M)\big\|_{L_2}}\right)^{\frac{\nu}{\nu M+1}} + Cm\left(\frac{\sum_{j\neq M}\sigma^2_{m,n;j}}{\sigma^2_{m,n;M}}\right)^{\frac{1}{2M+1}} \\
&\qquad = Cm\Big(n^{-\frac{\nu-2}{2(\nu M+1)}}\tilde{\beta}^{\frac{\nu}{\nu M+1}}_{M,\nu} + \rho^{\frac{1}{2M+1}}_{m,n;M}\Big)
\end{aligned}$$

for some absolute constant $C > 0$. This proves the first bound. The Gaussian approximation bound is obtained by combining the above with Proposition 4.6. □

Lemma C.13. (C.24) *holds.*

Proof. Write $W_j = V^K_j$ for $j < i$, $W_i = 0$ and $W_j = \xi^{(K)}_j$ for $j > i$. Then we can express

$$\begin{aligned}
\big\|\partial_i U^{(K)}_M(V^K_1,\dots,V^K_{i-1},\mathbf{0},\xi^{(K)}_{i+1},\dots,\xi^{(K)}_n)X_i\big\|^\nu_{L_\nu} &= \big\|\partial_i U^{(K)}_M(W_1,\dots,W_n)V^K_i\big\|^\nu_{L_\nu} \\
&= \Big(\frac{(n-m)!}{n!}\Big)^\nu m \Big\| \sum_{\substack{j_1,\dots,j_{m-1}\\ \text{distinct and in } [n]\setminus\{i\}}} \langle T_m, V^K_i \otimes W_{j_1} \otimes \dots \otimes \dots \otimes W_{j_{m-1}}\rangle \Big\|^\nu_{L_\nu},
\end{aligned}$$

where we used the symmetry of T_m in the last equality. Now by the independence of W_j's and V^K_i and noting that $\mathbb{E}\, W_i = 0$ and $\nu \in (2,3]$, we can use Lemma C.4 repeatedly to get

$$\begin{aligned}
&\big\|\partial_i U^{(K)}_M(V^K_1,\dots,V^K_{i-1},\mathbf{0},\xi^{(K)}_{i+1},\dots,\xi^{(K)}_n)X_i\big\|^\nu_{L_\nu} \\
&\qquad \le (C')^m\Big(\frac{(n-m)!}{n!}\Big)^\nu m \Big\| \sum_{\substack{j_1,\dots,j_{m-1}\\ \text{distinct and in } [n]\setminus\{i\}}} \langle T_m, V^K_i \otimes V^K_{j_1} \otimes \dots \otimes V^K_{j_{m-1}}\rangle \Big\|^\nu_{L_\nu}
\end{aligned}$$

for some absolute constant $C' > 0$. Denote $\|\bullet\|_{L_\nu|i} := \mathbb{E}[\bullet\,|V^K_i]$, $V^{(-i)}_j := V^K_j$ for $j < i$, $V^{(-i)}_j := V^K_{j+1}$ for $j \ge i$ and $V^{(-i)} = (V^{(-i)}_j)_{j\le n-1}$. We can further express the above as

$$\begin{aligned}
&(C')^m \frac{m}{n^\nu}\mathbb{E}\left[\Big\| \frac{(n-m)!}{(n-1)!} \sum_{\substack{j_1,\dots,j_{m-1}\\ \text{distinct and in } [n]\setminus\{i\}}} \langle T_m, V^K_i \otimes V^K_{j_1} \otimes \dots \otimes V^K_{j_{m-1}}\rangle \Big\|^\nu_{L_\nu|i}\right] \\
&\quad =: (C')^m \frac{m}{n^\nu}\mathbb{E}\left[\Big\|u^{(i)}_{m-1}\big(V^{(-i)}\big)\Big\|^\nu_{L_\nu|i}\right].
\end{aligned}$$

Conditioning on V^K_i, $u^{(i)}_{m-1}(V^{(-i)})$ is now a degree $m-1$ degenerate U-statistic of $n-1$ i.i.d. zero-mean random vectors, $V^{(-i)}_1,\dots,V^{(-i)}_{n-1}$. By a standard moment bound on degenerate U-statistics (Theorem 4.1.1 of Ferger (1996)), there is some absolute constant $C'' > 0$ such that

$$\Big\|u^{(i)}_{m-1}\big(V^{(-i)}\big)\Big\|_{L_\nu|i}$$

$$
\begin{aligned}
&\le (C'')^{m-1} \binom{n}{m-1}^{-1} n^{\frac{m-1}{2}} \Big(\textstyle\prod_{l=1}^{m-1} \frac{2}{(l-1)\nu+2} \Big) \\
&\qquad\qquad \times \big\| \langle T_m, V_i^K \otimes V_1^{(-i)} \otimes \ldots \otimes V_{m-1}^{(-i)} \rangle \big\|_{L_\nu | i} \\
&\le (C'')^{m-1} \, n^{\frac{m-1}{2}} \frac{(m-1)^{m-1}}{n^{m-1}} \Big(\textstyle\prod_{l=1}^{m-1} \frac{\nu}{(l-1)\nu+\nu} \Big) \\
&\qquad\qquad \times \big\| \langle T_m, V_i^K \otimes V_1^{(-i)} \otimes \ldots \otimes V_{m-1}^{(-i)} \rangle \big\|_{L_\nu | i} \\
&= (C'')^{m-1} \, n^{-\frac{m-1}{2}} \frac{(m-1)^{m-1}}{(m-1)!} \big\| \langle T_m, V_i^K \otimes V_1^{(-i)} \otimes \ldots \otimes V_{m-1}^{(-i)} \rangle \big\|_{L_\nu | i} \\
&\le (C'')^{m-1} \, n^{-\frac{m-1}{2}} e^{m-1} \big\| \langle T_m, V_i^K \otimes V_1^{(-i)} \otimes \ldots \otimes V_{m-1}^{(-i)} \rangle \big\|_{L_\nu | i} \qquad \text{almost surely} \, ,
\end{aligned}
$$

where we have used Stirling's approximation of the factorial in the last line. Combining the two bounds finishes the proof. □

C.7.3. Proofs for Section 5.4

The results on vertex-level fluctuations concern a U-statistic of i.i.d. variables, which has already been studied in Proposition 5.6.

Proof of Corollary 5.7. The first bound follows directly from the proof of Proposition 5.6 in Appendix C.7.2 by replacing $u_m(Y)$ by $\kappa_1(Y)$. The variance computation follows from substituting u_1 into the definition of $\sigma^2_{m,n;j}$ in Proposition 5.6, and computing

$$
\binom{m}{r}^2 \binom{n}{m}^2 \binom{n}{r}^{-1} = \binom{m}{r}\binom{n}{m} \frac{(m!)(n!)(r!)(n-r)!}{(r!)((m-r)!)(m!)((n-m)!)(n!)} = \binom{m}{r}\binom{n}{m}\binom{n-r}{m-r} .
$$

□

Proof of Lemma 5.8. By the variance formula in Corollary 5.7 and the fact that w is bounded, $\sigma^2_{m,n;r} = O(n^{2m-r})$ for all $r \in [m]$, which implies (i) $\Leftrightarrow$ (ii). We also have (ii) $\Leftrightarrow$ (iii) in view of the formula in Corollary 5.7. To prove (iii) $\Leftrightarrow$ (iv), recall that $|\mathrm{Aut}(H)|$ is the number of automorphisms of H, and let P_m be the set of permutations of $\{1, \ldots, m\}$. Also denote $P_m^{i \to 1} \subseteq P_m$ as the set of permutations that sends $\{i\}$ to $\{1\}$. Then we can write

$$
\begin{aligned}
&\textstyle\sum_{H' \subseteq \mathcal{G}_H(\{1,\ldots,m\})} \mathbb{E}\Big[\prod_{(i_s,i_t) \in E(H')} w(U_{i_s}, U_{i_t}) \,\Big|\, U_1 = x \Big] \\
&= \frac{1}{|\mathrm{Aut}(H)|} \textstyle\sum_{\sigma \in P_m} \mathbb{E}\Big[\prod_{(i_s,i_t) \in E(H)} w\big(U_{\sigma(i_s)}, U_{\sigma(i_t)}\big) \,\Big|\, U_1 = x \Big] \\
&= \frac{1}{|\mathrm{Aut}(H)|} \textstyle\sum_{\sigma \in P_m} \mathbb{E}\Big[\prod_{(i_s,i_t) \in E(H)} w\big(U_{i_s}, U_{i_t}\big) \,\Big|\, U_{\sigma^{-1}(1)} = x \Big] \\
&= \frac{1}{|\mathrm{Aut}(H)|} \textstyle\sum_{i=1}^m \sum_{\sigma \in P_m^{i \to 1}} \mathbb{E}\Big[\prod_{(i_s,i_t) \in E(H)} w\big(U_{i_s}, U_{i_t}\big) \,\Big|\, U_i = x \Big] \\
&= \frac{(m-1)!}{|\mathrm{Aut}(H)|} \textstyle\sum_{i=1}^m \mathbb{E}\Big[\prod_{(i_s,i_t) \in E(H)} w\big(U_{i_s}, U_{i_t}\big) \,\Big|\, U_i = x \Big] \, .
\end{aligned}
$$

(iii) says the above is constant for almost every $x \in [0,1]$, whereas (iv) is equivalent to requiring that

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\Big[\prod_{(i_s,i_t)\in E(H)} w(U_{i_s}, U_{i_t}) \,\Big|\, U_i = x \Big]$$

is constant for almost every $x \in [0,1]$. This proves that (iii) $\Leftrightarrow$ (iv). □

To prove the results on the edge-level fluctuations, we first present a stronger lemma, which will imply Lemma 5.10 and greatly simplify the proof of Proposition 5.9. For $i \in [n_*]$, write $W_j^{(i)} := \bar{Y}_j$ for $j \le i$ and $W_j^{(i)} := Z_j$ for $j > i$. Denote e_j as the edge in K_n indexed by $j \in [n_*]$. Let I and E be edge sets. Recall that $\delta_H(I)$ is the indicator of whether the graph formed by I is isomorphic to H. Also define $n_H(k, I, E)$ as the number of subgraphs of K_n that are isomorphic to H and can be formed with k edges from I and all edges from E.

Given $i \in [n_*]$, $0 \le k' \le k$, the edge subsets $I \subseteq [n_*]$ with $|I| \ge k'$ and $E_{k-k'} \subseteq E(K_{n_*})$ with $|E_{k-k'}| = k - k'$, as well as some random variable W that are either zero-mean or constant almost surely and are independent of $\{W_j^{(i)} \mid j \in I\}$, we define

$$S^{(i)}(k', I, E_{k-k'}, W) := \sum_{\substack{j_1,\dots,j_{k'} \in I \\ j_1<\dots<j_{k'}}} \delta_H(\{e_{j_l}\}_{l\in[k']} \cup E_{k-k'})\, W \prod_{l\in[k']} W_{j_l}^{(i)} \quad \text{for } k' \ge 1,$$
$$S^{(i)}(0, I, E_k, W) := \delta_H(E_k)\, W\ .$$

Notice that $\kappa_2(\bar{Y}) = S^{(n_*)}(k, [n_*], \emptyset, 1)$. The next lemma controls the conditional moments of these quantities given U, by exploiting that $\bar{Y}_j$'s and Z_j's are conditionally independent and zero-mean given U:

Lemma C.14. *There are some absolute constants $c, C > 0$ and $\nu \in (2,3]$ such that almost surely*

$$\mathbb{E}\big[\big|S^{(i)}(k', I, E_{k-k'}, W)\big|^{\nu} \,\big|\, U\big] \le C^{k'} \left(n_H(k', I, E_{k-k'})\right)^{\nu/2} \mathbb{E}[|W|^{\nu} \,|\, U] \times \max_{\substack{i_1,\dots,i_{k'}\in I \\ \text{all distinct}}} \prod_{l=1}^{k'} \mathbb{E}\big[|X_{i_l}|^{\nu} \,\big|\, U\big]\ ,$$
$$\mathbb{E}\big[\big|S^{(n_*)}(k', I, E_{k-k'}, W)\big|^{2} \,\big|\, U\big] \ge c^{k'}\, n_H(k', I, E_{k-k'})\, \mathbb{E}[|W|^{2} \,|\, U] \times \max_{\substack{i_1,\dots,i_{k'}\in I \\ \text{all distinct}}} \prod_{l=1}^{k'} \mathbb{E}\big[|X_{i_l}|^{2} \,\big|\, U\big]\ .$$

Proof of Lemma C.14. The proof proceeds by induction on $k \ge 0$. We claim that the constant in the upper bound is specified by $C = C_1 C_2$, where C_1 is the maximum of the absolute constant C'_ν in the second upper bound of Lemma A.4 over $\nu \in [2,3]$, and C_2 is the absolute constant in Lemma C.4. The constant c in the lower bound is the absolute constant in the lower bound of Lemma A.4 with $\nu = 2$.

The base case $k' = 0$ is straightforward by noting that $\delta_H(E_k) = n_H(0, I, E_k)$ for all

$I \subseteq [n_*]$. Suppose that the inductive statement holds for $k'-1$. To prove the statement for k', we enumerate the elements of I as $i_1 < \ldots < i_{n'}$ and consider a martingale difference sequence $(S_{l'})_{l'=1}^{n'}$ conditioning on U: For $l' \in [n']$, define

$$\begin{aligned} S_{l'}^{(i)} \;&\coloneqq \mathbb{E}\big[S^{(i)}(k', I, E_{k-k'}, W) \,\big|\, U, W, W_{i_1}^{(i)}, \ldots, W_{i_{l'}}^{(i)}\big] \\ &\quad - \mathbb{E}\big[S^{(i)}(k', I, E_{k-k'}, W) \,\big|\, U, W, W_{i_1}^{(i)}, \ldots, W_{i_{l'-1}}^{(i)}\big] \;. \end{aligned}$$

Since $(W_{i_j}^{(i)})_{j \le l'}$ are zero-mean and $(W_{i_j}^{(i)})_{j \le l'} \cup \{W\}$ are independent conditioning on U, for $k' \ge 1$, we have

$$\mathbb{E}\big[S^{(i)}(k', I, E_{k-k'}, W) \,\big|\, U, W\big] \;=\; 0 \qquad \text{almost surely} \;.$$

This allows us to express, almost surely,

$$S^{(i)}(k', I, E_{k-k'}, W) \;=\; \textstyle\sum_{l'=1}^{n'} S_{l'}^{(i)} \;.$$

Now notice that for $l' < k'$, $S_{l'}^{(i)} = 0$ almost surely. Since the non-zero terms in each difference $S_{l'}^{(i)}$ must involve $W_{i_{l'}}^{(i)}$, for $l' \ge k' > 1$, we have that almost surely

$$S_{l'}^{(i)} \;=\; \sum_{\substack{j_1, \ldots, j_{k'-1} \in \{i_1, \ldots, i_{l'-1}\} \\ j_1 < \ldots < j_{k'-1}}} \delta_H(\{e_{j_l}\}_{l \in [k'-1]} \cup E_{k-k'} \cup \{e_{i_{l'}}\})\, W\, W_{i_{l'}}^{(i)} \textstyle\prod_{l \in [k'-1]} W_{j_l}^{(i)} \;,$$

and for $l' \ge k' = 1$, we have that almost surely

$$S_{l'}^{(i)} \;= \delta_H(E_{k-k'} \cup \{e_{i_{l'}}\})\, W\, W_{i_{l'}}^{(i)} \;.$$

This in particular implies that for $l' \ge k'$, almost surely

$$S_{l'}^{(i)} \;= S^{(i)}\big(k'-1\,,\, \{i_1, \ldots, i_{l'-1}\}\,,\, E_{k-k'} \cup \{e_{i_{l'}}\}\,,\, W\, W_{i_{l'}}^{(i)}\big) \;, \tag{C.25}$$

Now by the martingale moment bound in Lemma A.4 with $\nu \in [2, 3]$, we get the almost sure bound

$$\begin{aligned} c \textstyle\sum_{l'=k'}^{n'} \mathbb{E}[|S_{l'}^{(i)}|^\nu \,|\, U] \;&\le \mathbb{E}\big[\big|S^{(i)}(k', I, E_{k-k'}, W)\big|^\nu \,\big|\, U\big] \\ &\le C_1\, \mathbb{E}\Big[\big(\textstyle\sum_{l'=k'}^{n'} |S_{l'}^{(i)}|^2\big)^{\nu/2} \,\Big|\, U\Big] \;, \end{aligned} \tag{C.26}$$

and since $W \times W_{i_{l'}}^{(i)}$ is independent of $W_{i_1}^{(i)}, \ldots, W_{i_{l'-1}}^{(i)}$, we may apply the inductive statement to control the individual $S_{l'}^{(i)}$ term. Now notice that the moment bound on each $S_{l'}^{(i)}$ will introduce a constant $w_{l'} \coloneqq n_H(k'-1, \{i_1, \ldots, i_{l'-1}\}, E_{k-k'} \cup \{e_{i'_l}\})$. Denote their sum as $w_H \coloneqq \sum_{l'=k'}^{n'} w_{l'}$, which satisfies

$$w_H \;=\; n_H(k', I, E_{k-k'}) \;. \tag{C.27}$$

By reweighting the sum in the upper bound in (C.26) and the Jensen's inequality, almost surely

$$\mathbb{E}\big[\big|S^{(i)}(k', I, E_{k-k'}, W)\big|^\nu \,\big|\, U\big] \;\le C_1\, (w_H)^{\nu/2}\, \mathbb{E}\Big[\Big(\textstyle\sum_{l'=k'}^{n'} \frac{w_{l'}}{w_H} \frac{1}{w_{l'}} |S_{l'}^{(i)}|^2\Big)^{\nu/2} \,\Big|\, U\Big]$$

$$\leq C_1 (w_H)^{\nu/2} \sum_{l'=k'}^{n'} \frac{w_{l'}}{w_H} \mathbb{E}\Big[\Big(\frac{1}{w_{l'}} |S_{l'}^{(i)}|^2 \Big)^{\nu/2} \,\Big|\, U \Big]$$
$$= C_1 (w_H)^{\nu/2} \sum_{l'=k'}^{n'} \frac{w_{l'}}{w_H} \frac{\mathbb{E}[|S_{l'}^{(i)}|^\nu \,|\, U]}{w_{l'}^{\nu/2}} \,.$$

By applying the upper bound in the inductive statement to (C.25), we have that almost surely

$$\begin{aligned}
&\mathbb{E}\big[|S_{l'}^{(i)}|^\nu \,\big|\, U\big] \\
&\leq C^{k'-1} w_{l'}^{\nu/2} \mathbb{E}\big[|W \times W_{i_{l'}}^{(i)}|^\nu \,\big|\, U\big] \\
&\qquad \times \max_{j_1,\ldots,j_{k'-1} \in \{i_1,\ldots,i_{l'-1}\} \text{ distinct}} \prod_{l=1}^{k'-1} \mathbb{E}\big[|X_{j_l}|^\nu \,\big|\, U\big] \\
&\overset{(a)}{\leq} C_2 C^{k'-1} w_{l'}^{\nu/2} \mathbb{E}[|W|^\nu | U] \, \mathbb{E}[|X_{i_{l'}}|^\nu | U] \\
&\qquad \times \max_{j_1,\ldots,j_{k'-1} \in \{i_1,\ldots,i_{l'-1}\} \text{ distinct}} \prod_{l=1}^{k'-1} \mathbb{E}\big[|X_{j_l}|^\nu \,\big|\, U\big] \\
&\leq C_2 C^{k'-1} w_{l'}^{\nu/2} \mathbb{E}[|W|^\nu \,|\, U] \, \max_{i_1,\ldots,i_{k'} \in I \text{ distinct}} \prod_{l=1}^{k'} \mathbb{E}\big[|X_{i_l}|^\nu \,|\, U\big] \,,
\end{aligned}$$

where we have noted that W is independent of $W_{i_{l'}}^{(i)}$ and used Lemma C.4 in (a) to replace $W_{i_{l'}}^{(i)}$ by $X_{i_{l'}}$. By noting that $C_* = C_1 C_2$ and using (C.27), we get that almost surely

$$\begin{aligned}
&\mathbb{E}\big[|S^{(i)}(k', I, E_{k-k'}, W)|^\nu \,\big|\, U\big] \\
&\qquad \leq C^{k'} (w_H)^{\nu/2} \sum_{l'=k'}^{n'} \frac{w_{l'}}{w_H} \mathbb{E}[|W|^\nu \,|\, U] \max_{i_1,\ldots,i_{k'} \in I \text{ distinct}} \prod_{l=1}^{k'} \mathbb{E}[|X_{i_l}|^\nu \,|\, U] \\
&\qquad = C^{k'} (n_H(k', I, E_{k-k'}))^{\nu/2} \mathbb{E}[|W|^\nu \,|\, U] \, \max_{i_1,\ldots,i_{k'} \in I \text{ distinct}} \prod_{l=1}^{k'} \mathbb{E}[|X_{i_l}|^\nu] \,,
\end{aligned}$$

which proves the upper bound. Similarly by the lower bound in the inductive statement and noting that we do not need to use Lemma C.4, we get that

$$\begin{aligned}
&\mathbb{E}[|S^{(n_*)}(k', I, E_{k-k'}, W)|^2 \,|\, U] \\
&\qquad \geq c^{k'} n_H(k', I, E_{k-k'}) \, \mathbb{E}[|W|^2 \,|\, U] \, \min_{i_1,\ldots,i_{k'} \in I \text{ distinct}} \prod_{l=1}^{k'} \mathbb{E}[|X_{i_l}|^2 \,|\, U] \,.
\end{aligned}$$

which proves the lower bound. □

Proof of Lemma 5.10. Since $\mathbb{E}[\kappa_2(\bar{Y})|U] = 0$ almost surely, by the law of total variance,

$$\mathrm{Var}[\kappa_2(\bar{Y})] \;=\; \mathbb{E}\,\mathrm{Var}[\kappa_2(\bar{Y})|U] \;=\; \mathbb{E}\,\Big\| \sum_{\substack{j_1,\ldots,j_k \in [n_*] \\ j_1 < \ldots < j_k}} \delta_H\big(\{e_{j_l}\}_{l\in[k]}\big) \prod_{l\in[k]} \bar{Y}_l \Big\|_{L_2|U}^2 \,,$$

where we have denoted $\|\bullet\|_{L_2|U}^2 := \mathbb{E}[|\bullet|^2|\,U]$. Since $(\bar{Y}_{ij})_{(i,j)\in E(K_n)}$ is a collection of n_* random variables that are conditionally independent and zero-mean given U, the quantity inside the conditional norm can be identified with $S^{(n_*)}(k, [n_*], \emptyset, 1)$. Applying Lemma C.14 conditionally on U, we get that for some absolute constants $c, C > 0$, almost surely,

$$\mathrm{Var}[\kappa_2(\bar{Y})|U] \;\leq C^k \, n_H(k, [n_*], \emptyset) \, \max_{i_1,\ldots,i_k \in [n_*] \text{ distinct}} \prod_{l=1}^{k} \mathrm{Var}[\bar{Y}_{i_l} \,|\, U] \,,$$

$$\mathrm{Var}[\kappa_2(\bar{Y})|U] \;\geq c^k\, n_H(k,[n_*],\emptyset)\, \min_{i_1,\ldots,i_k\in[n_*] \text{ distinct}} \prod_{l=1}^{k} \mathrm{Var}[\bar{Y}_{i_l}\,|\,U]\;.$$

Noting that $n(k,[n_*],\emptyset) = |\mathcal{G}_H([n])|$ and taking an expectation yield the desired bounds. □

Proof of Proposition 5.9. Observe that $\kappa_2(\bar{Y})$ is multilinear in $(\bar{Y}_i)_{i\in[n_*]}$ and a degree-k polynomial. Moreover, as $\bar{Y}_{ij}$'s are conditionally centred given U, $\mathbb{E}[\kappa_2(\bar{Y})\,|\,U] = 0$. Since $\mathrm{Var}[\kappa_2(Z)|U] > 0$ almost surely by assumption, by applying Theorem 4.1 and noting that the bounding constant is absolute, we get that for some absolute constant $C' > 0$, almost surely

$$\begin{aligned}&\sup_{t\in\mathbb{R}} \Big|\mathbb{P}\big(\kappa_2(\bar{Y}) \leq\ t\,\big|\,U\big) - \mathbb{P}\big(\kappa_2(Z)\leq t\,\big|\,U\big)\Big| \\ &\qquad\qquad \leq C' m \left(\frac{\sum_{i=1}^{n_*}\mathbb{E}\big[\partial_i\kappa_2(\bar{Y}_1,\ldots,\bar{Y}_{i-1},0,Z_{i+1},\ldots,Z_{n_*})\,\bar{Y}_i\big|^\nu\,|\,U\big]}{\mathrm{Var}[\kappa_2(\bar{Y})\,|\,U]^{\nu/2}}\right)^{\frac{1}{\nu k+1}}.\end{aligned} \tag{C.28}$$

Lemma C.14 implies that, for some absolute constants $C_*, c_* > 0$, we have that almost surely

$$\begin{aligned}&\mathbb{E}\big[\partial_i\kappa_2(\bar{Y}_1,\ldots,\bar{Y}_{i-1},0,Z_{i+1},\ldots,Z_{n_*})\,\bar{Y}_i\big|^\nu\,\big|\,U\,\big] \\ &\quad= \mathbb{E}\Big[\Big|\sum_{\substack{j_1,\ldots,j_{k-1}\in[n_*]\setminus\{i\}\\ j_1<\ldots<j_{k-1}}} \delta_H\big(\{e_{j_l}\}_{l\in[k-1]}\cup\{e_i\}\big)\, W_i^{(i)}\prod_{l\in[k-1]} W_{j_l}^{(i)}\Big|^\nu\,\Big|\,U\Big] \\ &\quad= \mathbb{E}[\,|S^{(i)}(k-1,[n_*]\setminus\{i\},\{e_i\},\bar{Y}_i)|^\nu\,|\,U] \\ &\quad\leq (C_*)^{k-1}\left(n_H(k-1,[n_*]\setminus\{i\},\{e_i\})\right)^{\nu/2}\mathbb{E}\big[\,|\bar{Y}_i|^\nu\,\big|\,U\big] \\ &\qquad\qquad \times \max_{i_1,\ldots,i_{k-1}\in[n_*]\setminus\{i\} \text{ distinct}} \prod_{l=1}^{k-1}\mathbb{E}\big[|\bar{Y}_{i_l}|^\nu\,\big|\,U\big] \\ &\quad= (C_*)^{k-1}\left(n_H(k-1,[n_*]\setminus\{i\},\{e_i\})\right)^{\nu/2}\max_{i_1,\ldots,i_k\in[n_*] \text{ distinct}}\prod_{l=1}^{k}\mathbb{E}\big[|\bar{Y}_{i_l}|^\nu\,\big|\,U\big]\;,\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}[\kappa_2(\bar{Y})\,|\,U] \;&= \mathbb{E}\Big[\Big|\sum_{\substack{j_1,\ldots,j_k\in[n_*]\\ j_1<\ldots<j_k}}\delta_H\big(\{e_{j_l}\}_{l\in[k]}\big)\prod_{l\in[k]} W_{j_l}^{(n_*)}\Big|^2\,\Big|\,U\Big] \\ &= \mathbb{E}\big[\big|S^{(n_*)}(k,[n_*],\emptyset,1)\big|^2\big|\,U\,\big] \\ &\geq (c_*)^k\, n_H(k,[n_*],\emptyset)\,\min_{i_1,\ldots,i_{k'}\in[n_*] \text{ distinct}}\prod_{l=1}^{k'}\mathbb{E}[|\bar{Y}_{i_l}|^2\,|\,U]\;.\end{aligned}$$

Now recall that $[n_*]$ indexes all edges of an complete graph K_n, whereas $n_H(k-1,[n_*]\setminus\{i\},\{e_i\})$ counts the number of subgraphs in K_n that contains e_i and is isomorphic to H. By symmetry, $n_H(k-1,[n_*]\setminus\{i\},\{e_i\})$ is the same for all $i\in[n_*]$, and

$$\sum_{i=1}^{n_*} n_H(k-1,[n_*]\setminus\{i\},\{e_i\}) \;=\; k\,|\mathcal{G}_H([n])| \;=\; k\left(n_H(k,[n_*],\emptyset)\right),$$

since each subgraph of K_n that is isomorphic to H has been counted exactly $k = |H|$

times. Thus

$$\begin{aligned}\textstyle\sum_{i=1}^{n_*}\big(n_H(k-1,[n_*]\setminus\{i\},\{e_i\})\big)^{\nu/2} &= \textstyle\sum_{i=1}^{n_*}\big(n_H(k-1,[n_*]\setminus\{1\},\{e_1\})\big)^{\nu/2}\\ &= n_*\big(\frac{k}{n_*}\,|\mathcal{G}_H([n])|\big)^{\nu/2}\,.\end{aligned}$$

Combining the above, we get that for some absolute constants $C'', C > 0$, almost surely

$$\begin{aligned}&\sup_{t\in\mathbb{R}}\Big|\mathbb{P}\big(\kappa_2(\bar{Y})\le t\big)-\mathbb{P}\big(\kappa_2(Z)\le t\,\big|\,U\big)\Big|\\ &\le C''m\left(\frac{\sum_{i=1}^{n_*}\big(n_H(k-1,[n_*]\setminus\{i\},\{e_i\})\big)^{\nu/2}}{|\mathcal{G}_H([n])|^{\nu/2}}\,\frac{\max_{i_1,\ldots,i_k\in[n_*]\text{ distinct}}\prod_{l=1}^{k}\mathbb{E}[|\bar{Y}_{i_l}|^\nu\,|\,U]}{\min_{i_1,\ldots,i_{k'}\in[n_*]\text{ distinct}}\prod_{l=1}^{k'}\mathbb{E}[|\bar{Y}_{i_l}|^2\,|\,U]}\right)^{\frac{1}{\nu k+1}}\\ &\le C''m\left(\frac{\big(\sum_{i=1}^{n_*}n_H(k-1,[n_*]\setminus\{i\},\{e_i\})\big)^{\nu/2}}{|\mathcal{G}_H([n])|^{\nu/2}}\right)^{\frac{1}{\nu k+1}}\rho_{\bar{Y}}(U)^{\frac{1}{\nu k+1}}\\ &= C''m\,k^{\frac{\nu}{2\nu k+2}}\,n_*^{-\frac{\nu-2}{2\nu k+2}}\,\rho_{\bar{Y}}(U)^{\frac{1}{\nu k+1}}\;\le\; Cm\,n^{-\frac{\nu-2}{\nu k+1}}\,\rho_{\bar{Y}}(U)^{\frac{1}{\nu k+1}}\,.\end{aligned}$$

In the last line, we have noted that $n_* = \binom{n}{2} \le n^2$. This proves the first bound. The second bound follows by applying Proposition 4.6 while conditioning on U. □

C.8 Properties of univariate distributions in Theorem 4.7

C.8.1. Proof of Gaussian moment bound in Lemma C.6

Since $\mathbb{E}|Z|^\nu = \pi^{-1/2}2^{\nu/2}\Gamma(\frac{\nu+1}{2})$, the proof boils down to approximating the Gamma function: Alzer (2003) proves that for all $x \ge 0$,

$$\sqrt{\pi}\Big(\frac{x}{e}\Big)^x\Big(8x^3+4x^2+x+\frac{1}{100}\Big)^{1/6} < \Gamma(x+1) < \sqrt{\pi}\Big(\frac{x}{e}\Big)^x\Big(8x^3+4x^2+x+\frac{1}{30}\Big)^{1/6}\,.$$

Since $\nu \ge 1$, $\frac{\nu+1}{2} \ge 1$. As $8x^3+4x^2+x+\frac{1}{30} \le 14(1+x)^3$ for $x \ge 0$, we have

$$\begin{aligned}\Gamma\Big(\frac{\nu+1}{2}\Big) &\le 14^{1/6}\sqrt{\pi}\Big(\frac{\nu-1}{2e}\Big)^{(\nu-1)/2}\Big(\frac{\nu+1}{2}\Big)^{1/2}\\ &\le 14^{1/6}\sqrt{\pi}\Big(\frac{\nu-1}{2}\Big)^{\frac{\nu}{2}-1}\Big(\frac{\nu^2-1}{4}\Big)^{1/2} \;\le\; 14^{1/6}\sqrt{\pi}\Big(\frac{\nu}{2}\Big)^{\frac{\nu}{2}}\,,\end{aligned}$$

which implies the desired bound that, for some absolute constant $C > 0$,

$$\mathbb{E}|Z|^\nu \;\le\; C\nu^{\nu/2}\,.$$

For the second bound, we note that if m_1 and m_2 have different parities,

$$\mathrm{Cov}[Z^{m_1},Z^{m_2}] \;=\; \mathbb{E}[Z^{m_1+m_2}]-\mathbb{E}[Z^{m_1}]\mathbb{E}[Z^{m_2}] \;=\; 0$$

since odd moments of Z vanish. Now focus on the case when m_1 and m_2 have the same parity, and recall that $m_1 \ge m_2$ by assumption. For $x \ge 0$ and $\alpha \in [0,1]$, we define the

function $R(x;\alpha) \coloneqq 2^{x+\frac{1}{2}}\Big(\frac{x}{e}\Big)^x(8x^3+4x^2+x+\alpha)^{1/6}$, which implies that

$$R\Big(\frac{\nu-1}{2};\frac{1}{100}\Big) \;<\; \mathbb{E}|Z|^\nu \;<\; R\Big(\frac{\nu-1}{2};\frac{1}{30}\Big)\,.$$

Then we can bound

$$\begin{aligned}\mathrm{Cov}[Z^{m_1},Z^{m_2}] \;&\geq\; \mathbb{E}[Z^{m_1+m_2}] - \mathbb{E}[|Z|^{m_1}]\mathbb{E}[|Z|^{m_2}] \\ &\geq R\Big(\frac{m_1+m_2-1}{2};\frac{1}{100}\Big) - R\Big(\frac{m_1-1}{2};\frac{1}{30}\Big)R\Big(\frac{m_2-1}{2};\frac{1}{30}\Big)\,.\end{aligned}$$

First suppose $m_2 \geq 2$, and denote $\alpha = 1/100$ and $\beta = 1/30$. Note that for $x \geq y \geq 3/4$,

$$\begin{aligned}&\frac{R(x+y;\alpha)}{R(x-1/4;\beta)R(y-1/4;\beta)} \\ &=\; e^{-1/2}\frac{(x+y)^{x+y}}{(x-1/4)^{x-1/4}(y-1/4)^{y-1/4}} \\ &\quad\times \frac{\big(8(x+y)^3+4(x+y)^2+(x+y)+\alpha\big)^{1/6}}{\big(8(x-1/4)^3+4(x-1/4)^2+(x-1/4)+\beta\big)^{1/6}\big(8(y-1/4)^3+4(y-1/4)^2+(y-1/4)+\beta\big)^{\frac{1}{6}}} \\ &\geq\; e^{-1/2}\Big(\frac{x+y}{y-1/4}\Big)^{y-1/4} \\ &\quad\times\left(\frac{8(x+y)^6+4(x+y)^5+(x+y)^4+\alpha(x+y)^3}{\big(8(x-1/4)^3+4(x-1/4)^2+(x-1/4)+\beta\big)\big(8(y-1/4)^3+4(y-1/4)^2+(y-1/4)+\beta\big)}\right)^{\frac{1}{6}} \\ &=:\; a(x,y)\Big(\frac{b(x,y)}{c(x,y)}\Big)^{1/6}\,.\end{aligned}$$

Note that since $x \geq y \geq 3/4$, we have

$$a(x,y) \;=\; e^{-1/2}\Big(\frac{x+y}{y-1/4}\Big)^{y-1/4} \;\geq\; e^{-1/2}2^{y-1/4} \;\geq\; e^{-1/2}2^2 \;>\; 2\,.$$

On the other hand, a lengthy computation gives

$$\begin{aligned}c(x,y) \;&<\; 64x^3y^3+32x^3y^2+8x^3y+8\beta x^3+32x^2y^3+16x^2y^2+4x^2y+4\beta x^2 \\ &\qquad +8xy^3+4xy^2+xy+\beta x+8\beta y^3+4\beta y^2+\beta y+\beta^2 \\ &\leq b(x,y) - \big(96x^3y^3-4x^3y-(8\beta-\alpha)x^3-10x^2y^2-x^2y-4\beta x^2-4xy^3-4xy^2 \\ &\qquad -xy-\beta x-8\beta y^3-4\beta y^2-\beta y-\beta^2\big) \\ &\leq b(x,y)-46x^3y^3 \;<\; b(x,y)\,,\end{aligned}$$

where we have used that $x, y \geq 3/4$, $\alpha = 1/100$ and $\beta = 1/30$ in the last line. Therefore $b(x,y)/c(x,y) > 1$ and, for $x \geq y \geq 3/4$,

$$\frac{R(x+y;\alpha)}{R(x-1/4;\beta)R(y-1/4;\beta)} \;>\; 2\,.$$

By identifying $x = \frac{m_1}{2}-\frac{1}{4}$ and $y = \frac{m_2}{2}-\frac{1}{4}$, we get that for $m_1 \geq m_2 \geq 2$ such that m_1 and m_2 have the same parity,

$$\begin{aligned}\mathrm{Cov}[Z^{m_1},Z^{m_2}] \;&\geq R\Big(\frac{m_1+m_2-1}{2};\frac{1}{100}\Big) - R\Big(\frac{m_1-1}{2};\frac{1}{30}\Big)R\Big(\frac{m_2-1}{2};\frac{1}{30}\Big) \\ &> \frac{1}{2}R\Big(\frac{m_1+m_2-1}{2};\frac{1}{100}\Big)\,.\end{aligned}$$

Meanwhile for $m_2 = 1$ and m_1 odd, since $\mathbb{E}[Z] = 0$, we obtain directly that

$$\mathrm{Cov}[Z^{m_1}, Z^{m_2}] \;=\; \mathbb{E}[Z^{m_1+m_2-1}] \;\geq\; R\Big(\frac{m_1+m_2-1}{2}; \frac{1}{100}\Big) \;>\; \frac{1}{2}R\Big(\frac{m_1+m_2-1}{2}; \frac{1}{100}\Big)\,.$$

Therefore for any m_1 and m_2 with the same parity, we get the desired bound that

$$\begin{aligned}\mathrm{Cov}[\,Z^{m_1}, Z^{m_2}] \;&>\; 2^{\frac{m_1+m_2}{2}}\Big(\frac{m_1+m_2-1}{2e}\Big)^{\frac{m_1+m_2-1}{2}} \;=\; 2^{\frac{1}{2}}\Big(\frac{m_1+m_2-1}{e}\Big)^{\frac{m_1+m_2-1}{2}} \\ &\geq c_1^{m_1+m_2}(m_1+m_2-1)^{\frac{m_1+m_2}{2}} \;\geq\; c_1^{m_1+m_2}\Big(\frac{m_1+m_2-1}{m_1+m_2}\Big)^{\frac{m_1+m_2}{2}}(m_1+m_2)^{\frac{m_1+m_2}{2}} \\ &\geq c^{m_1+m_2}(m_1+m_2)^{\frac{m_1+m_2}{2}}\end{aligned}$$

for some sufficiently small absolute constants $c_1, c > 0$. □

C.8.2. Proofs for properties of the heavy-tailed distribution in Section 4.5

Proof of Lemma 4.8. The first two moments of V_1 can be obtained directly from construction. To control the ω-th absolute moment for $\omega \geq 1$, first note that

$$\mathbb{E}|U_1|^\omega \;= 2p|x_0|^\omega + p|2x_0|^\omega \;=\; (2+2^\omega)px_0^\omega \;=\; \frac{(2+2^\omega)\sigma^\omega}{6^{\omega/2}p^{\omega/2-1}} \;=\; \frac{(2+2^\omega)}{6^{\omega/2}\sigma^{2(\omega-\nu)/(\nu-2)}}\,. \tag{C.29}$$

Moreover, by Jensen's inequality, we have $|a+b|^\omega \leq 2^{\omega-1}(|a|^\omega + |b|^\omega)$ for $a, b \in \mathbb{R}$. Combining this with the upper bound on $\mathbb{E}|\sigma^{-1}Z_1|^\omega$ from Lemma C.6, we get that

$$\mathbb{E}|V_1|^\omega \;\leq \frac{2^{-\omega/2}\,2^{\omega-1}(2+2^\omega)}{6^{\omega/2}\sigma^{2(\omega-\nu)/(\nu-2)}} + 2^{-\omega/2}2^{\omega-1}\sigma^\omega\mathbb{E}|\sigma^{-1}Z|^\omega \;\leq\; c_1^\omega\sigma^{-\frac{2(\omega-\nu)}{\nu-2}} + c_2^\omega\sigma^\omega\omega^{\omega/2}$$

for some absolute constants $c_1, c_2 > 0$ as desired. □

Lemma 4.9 approximates an empirical average of V_i's by a Gaussian Z_1', and gives a finer control by considering an additional remainder term. The key idea is to perform a fourth-order Taylor expansion in the characteristic functions of both $n^{-1/2}\sum_{i=1}^n V_i$ and Z_1', before turning back to the distribution functions. Note that the smoothing by Z_1 in the construction of V_1 makes the distribution of V_1 continuous, which enables this approach.

Proof of Lemma 4.9. Write $S_n = n^{-1/2}\sum_{i=1}^n V_i$, $F_n(x) = \mathbb{P}(S_n < x)$ and $F_Z(x) = \mathbb{P}(Z_1' < x)$. By the inversion formula for continuous random variables (Theorem 4.2.3.1., Cuppens (1975)),

$$F_n(x) \;= \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{-e^{-itx}}{it}\chi_n(t)\,dt \quad\text{and}\quad F_Z(x) \;= \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{-e^{-itx}}{it}\chi_Z(t)\,dt \tag{C.30}$$

where χ_n and χ_Z are the characteristic functions of S_n and Z. We first compare the characteristic functions. Denoting $\theta := 2^{-1/2}$ for simplicity, the characteristic function of V satisfies

$$\chi_V(t) \;=\; \mathbb{E}[e^{it(\theta U+\theta\sigma Z)}] \;=\; e^{-\theta^2\sigma^2t^2/2}\big(2pe^{-i\theta x_0 t} + (1-3p) + pe^{2i\theta x_0 t}\big)$$

$$
\begin{aligned}
&= e^{-\theta^2\sigma^2t^2/2}\big(1-3p+2p\cos(\theta x_0 t)+p\cos(2\theta x_0 t)+i(-2p\sin(\theta x_0 t)+p\sin(2\theta x_0 t))\big)\\
&\overset{(a)}{=} e^{-\frac{\theta^2\sigma^2t^2}{2}}\Big(1-3p+2p\big(1-\frac{\theta^2x_0^2t^2}{2}+\frac{\theta^4x_0^4t^4\cos(t\theta x_1')}{24}\big)\\
&\qquad +p\big(1-2t^2\theta^2x_0^2+\frac{2\theta^4x_0^4t^4\cos(2t\theta x_1')}{3}\big)\\
&\qquad -i(2p)\big(\theta x_0 t-\frac{\theta^3x_0^3t^3}{6}+\frac{\theta^4x_0^4t^4\sin(t\theta x_2')}{24}\big)\\
&\qquad +ip\big(2\theta x_0 t-\frac{4\theta^3x_0^3t^3}{3}+\frac{2\theta^4x_0^4t^4\sin(2t\theta x_2')}{3}\big)\Big)\\
&= e^{-\theta^2\sigma^2t^2/2}\Big(1-3p\theta^2x_0^2t^2-ip\theta^3x_0^3t^3+\frac{p\theta^4x_0^4t^4}{12}\cos(t\theta x_1')+\frac{2p\theta^4x_0^4t^4}{3}\cos(2t\theta x_1')\\
&\qquad -\frac{ip\theta^4x_0^4t^4}{12}\sin(t\theta x_2')+\frac{2ip\theta^4x_0^4t^4}{3}\sin(2t\theta x_2')\Big)\,.
\end{aligned}
$$

In (a), we have performed Taylor expansions on the real and imaginary parts with some $x_1', x_2' \in [0, x_0]$. Since $x_0 = \sigma/\sqrt{6p}$ and $p = \sigma^{2\nu/(\nu-2)}$, we have that

$$
px_0^2 \;=\; \frac{\sigma^2}{6}\,, \qquad px_0^3 \;=\; \frac{1}{6^{3/2}\sigma^{(6-2\nu)/(\nu-2)}}\,, \qquad px_0^4 \;=\; \frac{1}{6^2\sigma^{(8-2\nu)/(\nu-2)}}\,.
$$

Then the characteristic function of S_n can be expressed as

$$
\begin{aligned}
\chi_n(t) \;&=\; (\chi_V(n^{-1/2}\,t))^n\\
&= e^{-\theta^2\sigma^2t^2/2}\Big(1-\frac{3p\theta^2x_0^2t^2}{n}-\frac{ip\theta^3x_0^3t^3}{n^{3/2}}+\frac{p\theta^4x_0^4t^4}{12n^2}\cos(t\theta x_1')+\frac{2p\theta^4x_0^4t^4}{3n^2}\cos(2t\theta x_1')\\
&\qquad -\frac{ip\theta^4x_0^4t^4}{12n^2}\sin(t\theta x_2')+\frac{2ip\theta^4x_0^4t^4}{3n^2}\sin(2t\theta x_2')\Big)^n\\
&= e^{-\theta^2\sigma^2t^2/2}\Big(1-\frac{\theta^2\sigma^2t^2}{2n}-\frac{i\theta^3t^3}{6^{3/2}n^{3/2}\sigma^{(6-2\nu)/(\nu-2)}}+\frac{\theta^4t^4\cos(t\theta x_1')}{432n^2\sigma^{(8-2\nu)/(\nu-2)}}\\
&\qquad +\frac{\theta^4t^4\cos(2t\theta x_1')}{54n^2\sigma^{(8-2\nu)/(\nu-2)}}-\frac{i\theta^4t^4\sin(t\theta x_2')}{432n^2\sigma^{(8-2\nu)/(\nu-2)}}+\frac{i\theta^4t^4\sin(2t\theta x_2')}{54n^2\sigma^{(8-2\nu)/(\nu-2)}}\Big)^n\\
&= \exp\Big(-\frac{\theta^2\sigma^2t^2}{2}+n\log\Big(1-\frac{\theta^2\sigma^2t^2}{2n}-i\frac{Q_n(t)}{n}+R_1(t)+iR_2(t)\Big)\Big)\,,
\end{aligned}
$$

where we have defined

$$
\begin{aligned}
Q_n(t) \;&\coloneqq \frac{\theta^3t^3}{6^{3/2}n^{1/2}\sigma^{(6-2\nu)/(\nu-2)}}\,,\\
R_1(t) \;&\coloneqq \frac{\theta^4t^4\cos(t\theta x_1')}{432n^2\sigma^{(8-2\nu)/(\nu-2)}}+\frac{\theta^4t^4\cos(2t\theta x_1')}{54n^2\sigma^{(8-2\nu)/(\nu-2)}}\,,\\
R_2(t) \;&\coloneqq -\frac{\theta^4t^4\sin(t\theta x_2')}{432n^2\sigma^{(8-2\nu)/(\nu-2)}}+\frac{\theta^4t^4\sin(2t\theta x_2')}{54n^2\sigma^{(8-2\nu)/(\nu-2)}}\,.
\end{aligned}
$$

Now define

$$
R_n(t) \;\coloneqq\; n\log\Big(1-\frac{\theta^2\sigma^2t^2}{2n}-i\frac{Q_n(t)}{n}+R_1(t)+iR_2(t)\Big)+\frac{\theta^2\sigma^2t^2}{2}+iQ_n(t)\,.
$$

By multiplying and dividing $e^{-\sigma^2t^2/4}$ and recalling that $\theta = 2^{-1/2}$, we get that

$$
\chi_n(t) \;= e^{-\frac{\sigma^2t^2}{2}-iQ_n(t)}e^{R_n(t)} \;=\; \chi_Z(t)\,e^{-iQ_n(t)}e^{R_n(t)}\,.
$$

Now define $q(t) \coloneqq -iQ_n(t)\chi_Z(t)$. Then

$$
|\chi_n(t)-\chi_Z(t)-q(t)| \;= \Big|\chi_Z(t)\Big(e^{-iQ_n(t)}e^{R_n(t)}-1+iQ_n(t)\Big)\Big|
$$

$$
\begin{aligned}
&\le |\chi_Z(t) e^{-iQ_n(t)}(e^{R_n(t)} - 1)| + |\chi_Z(t)(e^{-iQ_n(t)} - 1 + iQ_n(t))| \\
&\le e^{-\sigma^2 t^2/2}\big(|e^{R_n(t)} - 1| + |e^{-iQ_n(t)} - 1 + iQ_n(t)|\big) \\
&\le e^{-\sigma^2 t^2/2}\big(|e^{R_n(t)} - 1| + |Q_n(t)|^2\big) \ . \qquad \text{(C.31)}
\end{aligned}
$$

We seek to control $R_n(t)$ by a Taylor expansion of the complex logarithm around 1, which is only permitted outside the branch cut $\mathbb{R}^- \cup \{0\}$. Set $T_n := n^{1/2}\sigma^{(4-\nu)/(2\nu-4)}$. For $|t| \le T_n$,

$$
\begin{aligned}
&\Big| - \frac{\theta^2\sigma^2 t^2}{2n} - \frac{iQ_n(t)}{n} + R_1(t) + iR_2(t)\Big| \\
&\qquad \le \frac{\theta^2\sigma^2 t^2}{2n} + \frac{\theta^3 t^3}{6^{3/2} n^{3/2}\sigma^{(6-2\nu)/(\nu-2)}} + \frac{\theta^4 t^4}{24 n^2 \sigma^{(8-2\nu)/(\nu-2)}} \\
&\qquad \le \frac{1}{2} + \frac{1}{6^{3/2}} + \frac{1}{24} \ \le \ 1 \ .
\end{aligned}
$$

In this case, the quantity in the complex logarithm in $R_n(t)$ is outside the branch cut, so by a Taylor expansion,

$$
R_n(t) \ = nR_1(t) + inR_2(t) + nR_3(t)
$$

where $R_3(t)$ is a remainder term that satisfies, for $|t| \le T_n$,

$$
\begin{aligned}
|R_3(t)| \ &\le \ \Big| - \frac{\theta^2\sigma^2 t^2}{2n} - i\frac{Q_n(t)}{n} + R_1(t) + iR_2(t)\Big|^2 \\
&\overset{(a)}{\le} \ 3\Big(\frac{t^4\theta^4\sigma^4}{4n^2} + \frac{t^6\theta^6}{6^3 n^3 \sigma^{(12-4\nu)/(\nu-2)}} + \frac{t^8\theta^8}{24^2 n^4 \sigma^{(16-4\nu)/(\nu-2)}}\Big) \\
&\overset{(b)}{\le} \ 3\theta^4 t^4\Big(\frac{\sigma^4}{4n^2} + \frac{1}{6^3 n^2 \sigma^{(8-3\nu)/(\nu-2)}} + \frac{1}{24^2 n^2 \sigma^{(8-2\nu)/(\nu-2)}}\Big) \\
&\overset{(c)}{\le} \ \frac{t^4}{5n^2\sigma^{(8-2\nu)/(\nu-2)}} \ ,
\end{aligned}
$$

In (a), we have noted that $(A+B+C)^2 \le 3(A^2+B^2+C^2)$; in (b), we have used $|t| \le T_n$ and $\theta \le 1$; in (c), we have compared the powers of $\sigma \in (0,1]$ by noting that $\frac{2\nu-8}{\nu-2} \le \frac{3\nu-8}{\nu-2}$ and $\frac{2\nu-8}{\nu-2} \le 4$, and combined the constants while noting that $\theta^4 = 1/4$. By a further Taylor expansion of the complex exponential, we obtain that for $|t| \le T_n$,

$$
\begin{aligned}
\big|e^{R_n(t)} - 1\big| \ \le \ |R_n(t)| \ &\le n|R_1(t)| + n|R_2(t)| + n|R_3(t)| \\
&\le \frac{\theta^4 t^4}{24 n\sigma^{(8-2\nu)/(\nu-2)}} + \frac{t^4}{5n\sigma^{(8-2\nu)/(\nu-2)}} \ \le \ \frac{0.22 t^4}{n\sigma^{(8-2\nu)/(\nu-2)}} \ .
\end{aligned}
$$

By plugging this and $Q_n(t) = \frac{\theta^3 t^3}{6^{3/2} n^{1/2}\sigma^{(6-2\nu)/(\nu-2)}}$ into (C.31), we get that for $|t| \le T_n$,

$$
|\chi_n(t) - \chi_Z(t) - q(t)| \ \le \frac{1}{n} e^{-\sigma^2 t^2/2}\Big(\frac{0.22 t^4}{\sigma^{(8-2\nu)/(\nu-2)}} + \frac{0.03 t^6}{\sigma^{(12-4\nu)/(\nu-2)}}\Big) \ . \qquad \text{(C.32)}
$$

Now note that F_q is a function analogous to (C.30) for q:

$$
\begin{aligned}
\frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{-e^{-itx}}{it} q(t)\, dt \ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-itx}}{t} Q_n(t)\chi_Z(t)\, dt \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-itx}}{t} \frac{\theta^3 t^3}{6^{3/2} n^{1/2}\sigma^{(6-2\nu)/(\nu-2)}} e^{-\sigma^2 t^2/2}\, dt
\end{aligned}
$$

$$= \frac{\theta^3}{(2^5 2^{3/2}\pi)\, n^{1/2}\sigma^{(6-2\nu)/(\nu-2)}} \int_{-\infty}^{\infty} t^2 e^{-\sigma^2 t^2/2 - itx} dt$$
$$= \frac{\theta^3}{(2^2 3^{3/2}\pi^{1/2})\, n^{1/2}\sigma^{\nu/(\nu-2)}} \Big(1 - \frac{x^2}{\sigma^2}\Big) e^{-x^2/(2\sigma^2)} \;=\; F_q(x)\;,$$

where we have recalled that the constant A in the definition of F_q satisfies $A = \frac{\theta^3}{2^2 3^{3/2}\pi^{1/2}}$ since $\theta = 2^{-1/2}$. Therefore by (C.30), a bound on the distribution functions can be given as

$$|\, F_n(x) - F_Z(x) - F_q(x)| \;\le\; \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{|\chi_n(t) - \chi_Z(t) - q(t)|}{|t|}\, dt$$
$$= \frac{1}{2\pi}\int_{|t|\le T_n} \frac{|\chi_n(t) - \chi_Z(t) - q(t)|}{|t|}\, dt + \frac{1}{2\pi}\int_{|t|>T_n} \frac{|\chi_n(t) - \chi_Z(t) - q(t)|}{|t|}\, dt \;=:\; I_1 + I_2\;.$$

(C.32) allows the first integral to be controlled as

$$I_1 \;\le \frac{1}{2\pi n}\int_{|t|\le T_n} e^{-\sigma^2 t^2/2}\Big(\frac{0.22|t|^3}{\sigma^{(8-2\nu)/(\nu-2)}} + \frac{0.03|t|^5}{\sigma^{(12-4\nu)/(\nu-2)}}\Big)\, dt$$
$$= \frac{1}{2\pi n}\Big(\frac{0.44\big(2 - (\sigma^2 T_n^2 + 2)e^{-\sigma^2 T_n^2/2}\big)}{\sigma^{2\nu/(\nu-2)}} + \frac{0.06\big(8 - (\sigma^4 T_n^4 + 4\sigma^2 T_n^2 + 8)e^{-\sigma^2 T_n^2/2}\big)}{\sigma^{2\nu/(\nu-2)}}\Big)$$
$$\le \frac{0.22}{n\sigma^{2\nu/(\nu-2)}}\;.$$

To deal with the case $|t| > T_n$, we first let $\{(U_i, V_i)\}_{i=1}^n$ be i.i.d. copies of (U, V) while applying independence and Jensen's inequality to obtain

$$|\chi_n(t)| \;= \Big|\mathbb{E}\big[e^{i\theta^2 t n^{-1/2}\sum_{i=1}^n U_i}\big]\mathbb{E}\big[e^{i\theta^2 t n^{-1/2}\sum_{i=1}^n V_i}\big]\Big|$$
$$\le \Big|\mathbb{E}\big[e^{i\theta^2 t n^{-1/2}\sum_{i=1}^n V_i}\big]\Big| \;=\; e^{-\frac{\theta^2\sigma^2 t^2}{2}}\;.$$

By noting $\chi_Z(t) = e^{-\sigma^2 t^2/2}$ and $|q(t)| = |Q_n(t)|e^{-\sigma^2 t^2/2} \le \frac{\theta^3 t^3}{6^{3/2} n^{1/2}\sigma^{(6-2\nu)/(\nu-2)}} e^{-\sigma^2 t^2/2}$, we can bound I_2 via the triangle inequality:

$$I_2 \;\le \frac{1}{2\pi}\int_{|t|>T_n} \frac{|\chi_n(t)|}{|t|} + \frac{|\chi_Z(t)|}{|t|} + \frac{|q(t)|}{|t|} dt$$
$$\le \frac{1}{2\pi}\int_{|t|>T_n} \frac{e^{-\theta^2\sigma^2 t^2/2}}{|t|} + \frac{e^{-\sigma^2 t^2/2}}{|t|} + \frac{\theta^3 t^2 e^{-\sigma^2 t^2/2}}{6^{3/2} n^{1/2}\sigma^{(6-2\nu)/(\nu-2)}} dt$$
$$\le \frac{1}{\pi T_n^2}\Big(\int_{t>T_n} t e^{-\theta^2\sigma^2 t^2/2} dt\Big) + \frac{1}{\pi T_n^2}\Big(\int_{t>T_n} t e^{-\sigma^2 t^2/2} dt\Big)$$
$$+ \frac{2}{\pi (12)^{3/2} n^{1/2}\sigma^{(6-2\nu)/(\nu-2)} T_n}\Big(\int_0^{\infty} t^3 e^{-\sigma^2 t^2/2} dt\Big)$$
$$= \frac{2e^{-\sigma^2 T_n^2/4}}{\pi T_n^2\sigma^2} + \frac{e^{-\sigma^2 T_n^2/2}}{\pi T_n^2\sigma^2} + \frac{4}{\pi (12)^{3/2} n^{1/2}\sigma^{(6-2\nu)/(\nu-2)+4}\, T_n}$$
$$\le \Big(\frac{3}{\pi} + \frac{4}{\pi\,(12)^{3/2}}\Big)\max\Big\{\frac{1}{n\sigma^{\nu/(\nu-2)}}\,,\, \frac{1}{n\sigma^{3\nu/(2\nu-4)}}\Big\} \;\le\; \frac{1}{n\sigma^{2\nu/(\nu-2)}}\;.$$

In the last line, we have recalled that $T_n = n^{1/2}\sigma^{(4-\nu)/(2\nu-4)}$ and noted that $\sigma \in (0, 1]$. Combining the bounds for I_1 and I_2, we get that

$$\Big|\mathbb{P}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n V_i < x\Big) - \mathbb{P}(Z < x) - F_q(x)\Big| \;= |F_n(x) - F_Z(x) - F_q(x)|$$
$$\le \frac{2}{n\sigma^{2\nu/(\nu-2)}}$$

as desired. □

Lemma 4.10 provides a normal approximation error with tighter x-dependence than a typical non-uniform Berry-Esseen bound. The proof is tedious, but the key ideas are the following:

(i) In the interval $|x| \le \frac{2\sigma}{\sqrt{3}\,M}$, i.e. x is within some multiples of the standard deviation of the Gaussian from the mean, Berry-Esseen bound is sufficiently tight so we can apply it directly. The interesting case is when $x > \frac{2\sigma}{\sqrt{3}\,M}$ ($x < \frac{2\sigma}{\sqrt{3}\,M}$ can be handled by symmetry).

(ii) To give a tighter control in the region $x > \frac{2\sigma}{\sqrt{3}\,M}$, the first part of the proof expands a telescoping sum of $n^{-1/2}\sum_{i=1}^n V_i$ and successively replacing the non-Gaussian part of each data point, i.e. U_i in $V_i = 2^{-1/2}U_i + 2^{-1/2}Z_i$, by a Gaussian.

(iii) Since the support of U_i has size $\Theta(\sigma^{-\frac{2}{\nu-2}})$, which is made to grow slower than $n^{1/2}$ by assumption of the lemma, the support of $n^{-1/2}U_i$ shrinks. Therefore on most part of the real line, the Gaussian approximation error is determined solely by the mass of the Gaussian counterpart of $n^{-1/2}U_i$. By carefully choosing a region moderately far from the origin, we can make use of the exponential tail of $n^{-1/2}U_i$. This is the argument leading up to (C.33), and gives rise to the first exponential term in Lemma 4.10.

(iv) At places where we cannot exploit the tail of $n^{-1/2}U_i$, we make use of the approximate Gaussianity of the remaining sum $n^{-1/2}\sum_{i'=1}^n V_{i'} - n^{-1/2}U_i$, which is the proof starting from (C.34). The idea is that when $n^{-1/2}U_i$ is close to the origin, the remaining sum $n^{-1/2}\sum_{i'=1}^n V_{i'} - n^{-1/2}U_i$ is allowed to be far from the origin, and we can obtain a sharper tail. This is achieved by symmetrising the sum $n^{-1/2}\sum_{i'=1}^n V_{i'} - n^{-1/2}U_i$ in (C.35) and then exploiting the improved $n^{-3/2}$ rate of normal approximation of a symmetric sum via the Lindeberg method in (C.36). This yields the second $n^{-3/2}x^{-4}$ term in Lemma 4.10; as suggested by the x^{-4} term, this error is only well-controlled at regions far from the origin.

The proof below is an elaborate version of the arguments in Example 9.1.3 of Senatov (1998) and with their final step replaced by the Lindeberg method.

Proof of Lemma 4.10. Recall that $V := 2^{-1/2}U + 2^{-1/2}Z$. Let $\{U_i, Z_i\}_{i=1}^n$ be i.i.d. copies of (U, Z) and let $\{Z_i'\}_{i=1}^n$ be an independent copy of $\{Z_i\}_{i=1}^n$. Write

$$W_i \;:=\; \frac{1}{\sqrt{2n}}\Big(\sum_{j=1}^n Z_j + \sum_{j=1}^{i-1} U_i + \sum_{j=i+1}^n Z_i'\Big)$$

and denote its probability measure by μ_{W_i}. By expressing the quantity to be bounded as a telescoping sum and noting the independence of (W_i, U_i, V_i) across different i's, we have

$$|F_n(x) - F_Z(x)| \;=\; \big|\mathbb{P}\big(W_n + (2n)^{-1/2}U_n < x\big) - \mathbb{P}\big(W_1 + (2n)^{-1/2}Z_1 < x\big)\big|$$

$$
\begin{aligned}
&\leq \textstyle\sum_{i=1}^n \big|\mathbb{P}\big(W_i + (2n)^{-1/2}U_i < x\big) - \mathbb{P}\big(W_i + (2n)^{-1/2}Z_i < x\big)\big| \\
&\leq \textstyle\sum_{i=1}^n \int_{-\infty}^{\infty} \big|\mathbb{P}\big((2n)^{-1/2}U_i < x - w\big) - \mathbb{P}\big((2n)^{-1/2}Z_i < x - w\big)\big| d\mu_{W_i}(w) \\
&= \textstyle\sum_{i=1}^n \int_{-\infty}^{\infty} \big|\mathbb{P}\big((2n)^{-1/2}U_i < x - w\big) - \mathbb{P}\big((2n)^{-1/2}Z_i < x - w\big)\big|\, d\mu_{W_i}(w) \\
&\coloneqq \textstyle\sum_{i=1}^n \int_{-\infty}^{\infty} J_i(w)\, d\mu_{W_i}(w) \;=\; \sum_{i=1}^n \Big(\int_{-\infty}^{x/2} J_i(w)\, d\mu_{W_i}(w) + \int_{x/2}^{\infty} J_i(w)\, d\mu_{W_i}(w) \Big) \,.
\end{aligned}
$$

We first focus on the case $x > \frac{2\sigma}{\sqrt{3M}}$, where M is the constant in the assumption $\sigma^{\nu/(\nu-2)} \geq Mn^{-1/2}$. Since $(2n)^{-1/2}U_i$ is bounded in norm by $(2n)^{-1/2}(2x_0) = \frac{\sigma}{\sqrt{3n}\,\sigma^{\nu/(\nu-2)}} \leq \frac{\sigma}{\sqrt{3M}} < \frac{x}{2}$ almost surely, the probability $\mathbb{P}\big(n^{-1/2}U_i \geq x - w\big)$ for $w < x/2$ is zero. Therefore

$$
\begin{aligned}
&\int_{-\infty}^{x/2} J_i(w) d\mu_{W_i}(w) \\
&\qquad = \int_{-\infty}^{x/2} \big|\mathbb{P}\big((2n)^{-1/2}U_i \geq x - w\big) - \mathbb{P}\big((2n)^{-1/2}Z_i \geq x - w\big)\big| d\mu_{W_i}(w) \\
&\qquad = \int_{-\infty}^{x/2} \mathbb{P}\big((2n)^{-1/2}Z_i \geq x - w\big) d\mu_{W_i}(w) \quad \leq \quad \mathbb{P}\Big((2n)^{-1/2}Z_i \geq \frac{x}{2}\Big) \\
&\qquad \overset{(a)}{\leq} \frac{\sigma}{\sqrt{\pi n}\, x} e^{-\frac{nx^2}{4\sigma^2}} \overset{(b)}{\leq} \frac{\sigma}{\sqrt{\pi n}\, x} \frac{8\sigma^2}{nx^2} e^{-\frac{nx^2}{8\sigma^2}} \overset{(c)}{\leq} \frac{3^{3/2}M^3}{\sqrt{\pi}\, n^{3/2}} e^{-\frac{nx^2}{8\sigma^2}} \leq \frac{3M^3}{n^{3/2}} e^{-\frac{nx^2}{8\sigma^2}} \,.
\end{aligned}
\tag{C.33}
$$

In (a), we used a standard bound for the complementary error function. In (b), we have noted that $y \coloneqq \frac{nx^2}{8\sigma^2} > \frac{n}{6M^2} \geq 1$ since $n \geq 6M^2$ and applied the bound $e^{-y} \leq \frac{1}{y}$ for $y \geq 1$. In (c), we have used $x > \frac{2\sigma}{\sqrt{3M}}$ again. Now for the other integral, by the standard Berry-Esseen bound,

$$
\int_{x/2}^{\infty} J_i(w) d\mu_{W_i}(w) \;\leq\; \frac{\mathbb{E}|U_i|^3}{n^{3/2}\sigma^3} \int_{x/2}^{\infty} d\mu_{W_i}(w) \;=\; \frac{10}{6^{3/2} n^{3/2} \sigma^{\nu/(\nu-2)}} \mathbb{P}(W_i \geq x/2) \,,
\tag{C.34}
$$

where we have used the moment formula from (C.29) to get that

$$
\mathbb{E}|U_i|^3 \;=\; \frac{10}{6^{3/2}\sigma^{(6-2\nu)/(\nu-2)}} \,.
$$

To handle the probability term, we first consider rewriting the sum $W_i = \xi_i + S_i$ with a Gaussian component ξ_i and an independent non-Gaussian component S_i given by

$$
\xi_i \;\coloneqq\; \frac{1}{\sqrt{2n}}\big(\textstyle\sum_{j=1}^n Z_i + \sum_{j=i+1}^n Z_i'\big) \sim \mathcal{N}\Big(0, \frac{(2n-i)}{2n}\sigma^2\Big) \quad \text{and} \quad S_i \;\coloneqq\; \frac{1}{\sqrt{2n}} \textstyle\sum_{j=1}^{i-1} U_i \,.
$$

We also write $W_i' = \xi_i' + S_i'$, where ξ_i' and S_i' are i.i.d. copies of ξ_i and S_i, and denote the symmetrisation $\bar{W}_i = W_i - W_i'$. Let μ_{W_i} be the measure associated with W_i. Then

$$
\begin{aligned}
\mathbb{P}(\bar{W}_i \geq x/2) \;&=\; \int_{-\infty}^{\infty} \mathbb{P}(-W_i' \geq x/2 - t)\, d\mu_{W_i}(t) \;\geq\; \int_{x/2}^{\infty} \mathbb{P}(-W_i' \geq x/2 - t)\, d\mu_{W_i}(t) \\
&\geq \big(\inf_{t \in [\frac{x}{2}, \infty)} \mathbb{P}(-W_i' \geq x/2 - t)\big) \mathbb{P}(W_i \geq x/2) \;=\; \mathbb{P}(-W_i' \geq 0)\, \mathbb{P}(W_i \geq x/2) \,.
\end{aligned}
$$

This rearranges to give

$$\mathbb{P}(W_i \geq x/2) \;\leq\; \mathbb{P}(-W_i' \geq 0)^{-1}\, \mathbb{P}(\bar{W}_i \geq x/2)\;, \tag{C.35}$$

The first probability can be lower bounded by

$$\mathbb{P}(-W_i' \geq 0) \;= \mathbb{P}(\xi_i + S_i \leq 0) \;\geq\; \mathbb{P}(\xi_i \leq 0)\, \mathbb{P}(S_i \leq 0)\;.$$

Note that $\mathbb{P}(\xi_i \leq 0) = 1/2$ since ξ_i is symmetric and $\mathbb{P}(S_1 \leq 0) = 1$. For $i > 1$, by a Berry-Esseen bound and the assumption that $\sigma^{\nu/(\nu-2)} \geq M n^{-1/2}$ and $M \geq 10$,

$$\mathbb{P}(S_i \leq 0) \;\geq\; \frac{1}{2} - \frac{2^{3/2}\mathbb{E}|U_i|^3}{n^{1/2}\sigma^3} \;=\; \frac{1}{2} - \frac{10}{3^{3/2} n^{1/2} \sigma^{\nu/(\nu-2)}} \;\geq\; \frac{1}{2} - \frac{10}{3^{3/2} M} \;\geq\; \frac{1}{4}\;.$$

This implies that $\mathbb{P}(-W_i' \geq 0) \geq \frac{1}{8}$ and therefore

$$\mathbb{P}(W_i \geq x/2) \;\leq\; 8\, \mathbb{P}(\bar{W}_i \geq x/2)\;.$$

In other words, we have shown that $\mathbb{P}(W_i \geq x/2)$ in the bound (C.34) can now be controlled in terms of the symmetric variable $\bar{W}_i$.

The final step is to compare $\bar{W}_i$ to $\bar{Z} \sim \mathcal{N}(0, \frac{\theta^2(2n-1)}{n}\sigma^2)$. We apply the standard Lindeberg's argument with a smooth test function h. By Lemma A.10 with $\delta = x/4$ and $m = 3$, there is a three-times differentiable function h such that for some absolute constant $C'' > 0$,

$$\mathbb{I}_{\{t \geq x/2\}} \;\leq\; h(t) \;\leq\; \mathbb{I}_{\{t \geq x/4\}} \qquad \text{and} \qquad |h'''(t) - h'''(s)| \;\leq\; C'' x^{-4} |t - s|\;.$$

Consider the symmetric variable $\bar{U}_i = \frac{\theta}{\sqrt{n}}(U_i - U_i')$, where $\{U_i'\}_{i=1}^n$ is an independent copy of $\{U_i\}_{i=1}^n$. Write $Y_k = \sum_{j=1}^{k-1} \bar{U}_i + \bar{\xi}_i$, where $\bar{\xi}_i \sim \mathcal{N}\big(0, \frac{\theta^2(4n-2k)}{n}\sigma^2\big)$ is independent of $\bar{U}_i$. Also denote $\zeta_i \sim \mathcal{N}\big(0, \frac{2\theta^2}{n}\sigma^2\big)$. Then

$$\begin{aligned} \mathbb{P}\big(\bar{W}_i \geq x/2\big) \;&\leq\; \mathbb{P}(\bar{Z} \geq x/4) + |\mathbb{E}[h(\bar{W}_i) - h(\bar{Z})]| \\ &\leq \mathbb{P}(\bar{Z} \geq x/4) + \textstyle\sum_{k=1}^{i} |\mathbb{E}[h(Y_k + \bar{U}_k) - h(Y_k + \zeta_k)]| =: \mathbb{P}(\bar{Z} \geq x/4) + \sum_{k=1}^{i} A_i. \end{aligned} \tag{C.36}$$

Since $\bar{Z} \sim \mathcal{N}(0, \frac{\theta^2(2n-1)}{n}\sigma^2)$, by noting $\frac{\theta^2(2n-1)}{n}\sigma^2 \leq \sigma^2$ and $x > \frac{2\sigma}{\sqrt{3}\,M}$, we have

$$\mathbb{P}(\bar{Z} \geq x/4) \;\leq\; \frac{\sigma}{\sqrt{2\pi}\,(x/4)} e^{-\frac{x^2}{2\sigma^2}} \;\leq\; \frac{\sqrt{6}\,M}{\sqrt{\pi}} e^{-\frac{x^2}{2\sigma^2}} \;\leq\; 2M e^{-\frac{x^2}{2\sigma^2}}\;.$$

By performing two third-order Taylor expansions around Y_k, each A_i satisfies

$$A_i \;=\; \big|\mathbb{E}\big[h'(Y_k)(\bar{U}_k - \zeta_k) + \frac{1}{2}h''(Y_k)(\bar{U}_k^2 - \zeta_k^2) + \frac{1}{6}h'''(Y_k + \tilde{U}_k)\bar{U}_k^3 - \frac{1}{6}h'''(Y_k + \tilde{\zeta}_k)\zeta_k^3\big]\big|$$

for some $\tilde{U}_k \in [0, U_k]$ and $\tilde{\zeta}_k \in [0, \zeta_k]$ that exist almost surely. The first two terms vanish by independence of Y_k and $(\bar{U}_k, \zeta_k)$ as well as the fact that $\bar{U}_k$ and ζ_k match in mean and variance. Since $\bar{U}_k$ and ζ_k are both symmetric, they both have zero third moments. By adding and subtracting a third moment term and applying the Lipschitz property of h''',

we obtain

$$\begin{aligned} A_i &= \frac{1}{6}\big|\mathbb{E}\big[\big(h'''(Y_k+\tilde{U}_k)-h'''(Y_k)\big)\bar{U}_k^3-\big(h'''(Y_k+\tilde{\zeta}_k)-h'''(\zeta_k)\big)\zeta_k^3\big] \\ &\le \frac{C''}{6x^4}\big(\mathbb{E}\,\bar{U}_k^4+\mathbb{E}\,\zeta_k^4\big) \overset{(a)}{\le} \frac{C'}{2n^2x^4}\Big(\frac{1}{\sigma^{(8-2\nu)/\nu-2}}+\sigma^4\Big) \le \frac{C'}{n^2x^4\sigma^{(8-2\nu)/(\nu-2)}}\,, \end{aligned}$$

where C' is an absolute constant; in (a), we have noted that $|U_i-U_i'|^4 \le (|U_i|+|U_i'|)^4 \le 8(|U_i|^4+|U_i'|^4)$ and applied the moment bound from (C.29) in the proof of Lemma 4.8. Combining the bounds gives

$$\mathbb{P}(W_i \ge x/2) \le 8\mathbb{P}\big(\bar{W}_i \ge x/2\big) \le 2Me^{-\frac{x^2}{2\sigma^2}}+\frac{C'}{nx^4\sigma^{(8-2\nu)/(\nu-2)}}\,,$$

and therefore

$$\int_{x/2}^{\infty} J_i(w)d\mu_{W_i}(w) \le \frac{10\,\mathbb{P}(W_i\ge x/2)}{6^{3/2}n^{3/2}\sigma^{\nu/(\nu-2)}} \le \frac{20Me^{-\frac{x^2}{2\sigma^2}}}{6^{3/2}n^{3/2}\sigma^{\nu/(\nu-2)}}+\frac{10C'}{6^{3/2}n^{5/2}x^4\sigma^{(8-\nu)/(\nu-2)}}\,.$$

Combining this with the bound for $\int_{-\infty}^{x/2} J_i(w)d\mu_{W_i}(w)$ in (C.33), we get that for $x > \frac{2\sigma}{\sqrt{3}M}$,

$$\begin{aligned} |F_n(x)-F_Z(x)| &\le \sum_{i=1}^n\Big(\int_{-\infty}^{x/2} J_i(w)d\mu_{W_i}(w)+\int_{x/2}^{\infty} J_i(w)d\mu_{W_i}(w)\Big) \\ &\le \frac{3M^3e^{-\frac{nx^2}{8\sigma^2}}}{n^{1/2}}+\frac{20Me^{-\frac{x^2}{2\sigma^2}}}{6^{3/2}n^{1/2}\sigma^{\nu/(\nu-2)}}+\frac{10C'}{6^{3/2}n^{3/2}x^4\sigma^{(8-\nu)/(\nu-2)}} \\ &\le C_M'\Big(\frac{1}{n^{1/2}\sigma^{\nu/(\nu-2)}}e^{-\frac{x^2}{8\sigma^2}}+\frac{1}{n^{3/2}x^4\sigma^{(8-\nu)/(\nu-2)}}\Big)\,, \end{aligned}$$

where $C_M' > 0$ is a constant that depends only on M. Now for $|x| \le \frac{2\sigma}{\sqrt{3}\,M}$, we can use the standard Berry-Esseen bound directly and the moment bound from Lemma 4.8 to get that

$$|F_n(x)-F_Z(x)| \le \frac{\mathbb{E}|X|^3}{n^{1/2}\sigma^3} \le \frac{a}{n^{1/2}\sigma^3}\big(\sigma^{-\frac{6-2\nu}{\nu-2}}+1\big) \le \frac{a\max\{1,\sigma^3\}}{n^{1/2}}\sigma^{-\frac{\nu}{\nu-2}}$$

for some absolute constant $a > 0$. Since in this case, $e^{-x^2/(8\sigma^2)} \ge e^{-1/6M}$, we again obtain the desired bound that for some constant $C_M'' > 0$ depending only on M,

$$|F_n(x)-F_Z(x)| \le \frac{C_M'\max\{1,\sigma^3\}e^{-\frac{x^2}{8\sigma^2}}}{n^{1/2}\sigma^{\nu/(\nu-2)}} \le C_M''\Big(\frac{\max\{1,\sigma^3\}e^{-\frac{x^2}{8\sigma^2}}}{n^{1/2}\sigma^{\nu/(\nu-2)}}+\frac{1}{n^{3/2}x^4\sigma^{(8-\nu)/(\nu-2)}}\Big)\,.$$

The proof for $x < -\frac{2\sigma}{\sqrt{3}M}$ is exactly analogous to the proof for $x > \frac{2\sigma}{\sqrt{3}M}$. □

Appendix D

Proofs for Chapter 6

This appendix concerns additional results and proofs that concern data augmentation in Chapter 6. The appendix is organised as follows:

Appendix D.1 states several generalisations and additional corollaries of the main result.

Appendix D.2 states additional results for the toy statistic in Section 6.4.3 and the ridgeless regressor in Section 6.5.

Appendix D.3 states and proves auxiliary tools used in subsequent proofs.

Appendix D.4 proves our main theorem. A proof overview is given in Appendix D.4.1.

Appendix D.5 presents the proofs of the results in Appendix D.1.

Appendix D.6 proves all results in Section 6.4 and Appendix D.2.1, all of which concern the asymptotic distribution and variance of the estimator.

Appendix D.7 proves all results in Section 6.5 and Appendix D.2.2, which concern the limiting risk of an estimator.

Notation. Throughout the appendix, we shorten $\alpha_{r;m}(f)$ to $\alpha_{r;m}$ whenever f is clear from the context, and write $\mathcal{Z}^\delta := \{\mathbf{Z}_1^\delta, \ldots, \mathbf{Z}_n^\delta\} \in \mathcal{D}^{nk}$.

D.1 Variants and corollaries of the main result

This section provides some additional results. Theorem D.1 below generalises Theorem 6.1 such that (i) transformed data $\phi(x)$ and x are allowed to live in different domains, and (ii) an additional parameter δ trades off between a tighter bound and lower variance. Corresponding generalisations of the corollaries in Section 6.3 follow. We also provide a formal statement for the convergence of estimates of the form $g(\text{empirical average})$ discussed in Section 6.4.3 (see Lemma D.7), and a variant of Theorem 6.1for non-smooth statistics in high dimensions.

Throughout this section, the big-O and small-o notations are stated under the asymptotic that $n \to \infty$, and dimensions d and q are treated as variables that may depend on n, which manifest through $\nu_{r;m}$ terms. The leading constants are always absolute constants and independent of n.

D.1.1. Generalisations of results in Section 6.3

We first allow the domain and range of elements of $\mathcal{T}$, i.e. augmentations to differ: Let $\mathcal{T}'$ be a family of measurable transformations $\mathcal{D}' \to \mathcal{D}$, and the data $\mathbf{X}_1, \dots, \mathbf{X}_n$ be i.i.d. random elements of $\mathcal{D}' \subseteq \mathbb{R}^{d'}$. An example where this formulation is useful is the empirical risk, where we study the empirical average of the following quantities

$$l(\tau_{11}\mathbf{X}_1), \dots, l(\tau_{nk}\mathbf{X}_n) \,, \qquad \text{for some loss function } l : \mathcal{D}' \to \mathbb{R} \,.$$

Note that Theorem D.1 remains applicable by setting $\phi_{ij}(\mathbf{X}_1) \coloneqq l(\tau_{ij}\mathbf{X}_1)$, with the augmentations used on data are determined through τ_{ij}. This is used in the softmax ensemble example in Proposition 6.14.

Next, we introduce a deterministic parameter $\delta \in [0, 1]$, and redefine the moment and mixed smoothness conditions. Recall $\Sigma_{11} \coloneqq \mathrm{Var}[\phi_{11}\mathbf{X}_1]$ and $\Sigma_{12} \coloneqq \mathrm{Cov}[\phi_{11}\mathbf{X}_1, \phi_{12}\mathbf{X}_1]$, the $d \times d$ matrices defined in (6.4) in the main text. Consider the following alternative requirements on moments of surrogates $\{\mathbf{Z}_i^\delta\}_{i \le n}$:

$$\mathbb{E}\mathbf{Z}_i^\delta = \mathbf{1}_{k\times 1} \otimes \mathbb{E}[\phi_{11}\mathbf{X}_1], \ \mathrm{Var}\mathbf{Z}_i^\delta = \mathbf{I}_k \otimes \big((1-\delta)\Sigma_{11} + \delta\Sigma_{12}\big) + (\mathbf{1}_{k\times k} - \mathbf{I}_k) \otimes \Sigma_{12}. \tag{D.1}$$

Note that when $\delta = 0$, this recovers (6.1). Write $\mathbf{Z}_1^\delta = (\mathbf{Z}_{1j}^\delta)_{j \le k}$ where $\mathbf{Z}_{ij}^\delta \in \mathcal{D}$. In lieu of the moment terms defined in (6.4) , we consider the moment terms defined by

$$c_1 \coloneqq \frac{1}{2}\big\|\mathbb{E}\mathrm{Var}[\phi_{11}\mathbf{X}_1|\mathbf{X}_1]\big\| \,, \quad c_X \coloneqq \frac{1}{6}\sqrt{\mathbb{E}\|\phi_{11}\mathbf{X}_1\|^6} \,, \qquad c_{Z^\delta} \coloneqq \frac{1}{6}\sqrt{\mathbb{E}\big[\|\mathbf{Z}_{11}^\delta\|^6\big]} \,.$$

Again when $\delta = 0$, the last two moment terms are exactly those defined in (6.4). Finally, we also use a tighter moment control on noise stability. Denote $\mathbf{W}_i^\delta$ as the analogue of $\mathbf{W}_i$ with $\{\mathbf{Z}_{i'}\}_{i'>i}$ replaced by $\{\mathbf{Z}_{i'}^\delta\}_{i'>i}$, and define

$$\alpha_{r;m}(f) \coloneqq \sum_{s\le q} \max_{i\le n} \max\Big\{ \Big\| \sup_{\mathbf{w}\in[\mathbf{0},\Phi_i\mathbf{X}_i]} \|D_i^r f_s(\mathbf{W}_i^\delta(\mathbf{w}))\| \Big\|_{L_m} , \Big\| \sup_{\mathbf{w}\in[\mathbf{0},\mathbf{Z}_i^\delta]} \|D_i^r f_s(\mathbf{W}_i^\delta(\mathbf{w}))\| \Big\|_{L_m} \Big\}.$$

$\alpha_{r;m}(f)$ is related to $\alpha_r(f)$ defined in (6.2) by $\alpha_r(f) = \alpha_{r;6}(f)$ in the case $\delta = 0$. The mixed smoothness terms of interest are in turn defined by

$$\begin{aligned} \lambda_1(n,k) &\coloneqq \gamma_2(h)\alpha_{1;2}(f)^2 + \gamma_1(h)\alpha_{2;1}(f) \,, \\ \text{and } \lambda_2(n,k) &\coloneqq \gamma_3(h)\alpha_{1;6}(f)^3 + 3\gamma_2(h)\alpha_{1;4}(f)\alpha_{2;4}(f) + \gamma_1(h)\alpha_{3;2}(f) \,. \end{aligned} \tag{D.2}$$

The choice of L_6 norm in Theorem 6.1 is out of simplicity rather than necessity.

Theorem D.1. *(Main result, generalised) Consider i.i.d. random elements $\mathbf{X}_1, \dots, \mathbf{X}_n$ of $\mathcal{D}'$, and two functions $f \in \mathcal{F}_3(\mathcal{D}'^{nk}, \mathbb{R}^q)$ and $h \in \mathcal{F}_3(\mathbb{R}^q, \mathbb{R})$. Let $\phi_{11}, \dots, \phi_{nk}$ be i.i.d.*

random elements of $\mathcal{T}'$, independent of $\mathcal{X}$. Then for any i.i.d. variables $\mathbf{Z}_1^\delta, \ldots, \mathbf{Z}_n^\delta$ in $\mathcal{D}^k$ satisfying (D.1),

$$\left|\mathbb{E}h(f(\Phi\mathcal{X})) - \mathbb{E}h(f(\mathbf{Z}_1^\delta, \ldots, \mathbf{Z}_n^\delta))\right| \;\leq\; \delta n k^{1/2}\lambda_1(n,k)c_1 + nk^{3/2}\lambda_2(n,k)(c_X + c_{Z^\delta}) \,.$$

The proof of Theorem D.1 is delayed to Appendix D.4. By observing the bound in Theorem D.1 and the moment condition (6.4),we see that δ is a parameter that trades off between a tighter bound at the price of higher variances $\mathrm{Var}[\mathbf{Z}_i]$ (for $\delta = 0$), versus an additional term in the bound and smaller variance ($\delta = 1$). In particular, setting $\delta = 0$ recovers Theorem 6.1:

Proof of Theorem 6.1. In Theorem D.1, setting $\mathcal{D}' = \mathcal{D}$ recovers $\mathcal{T}$ from $\mathcal{T}'$, and setting $\delta = 0$ recovers $\{\mathbf{Z}_i\}_{i\leq n}, c_Z$ from $\{\mathbf{Z}_i^\delta\}_{i\leq n}, c_{Z^\delta}$. Moreover, only the second term remains in the RHS bound. Since for $m \leq 12$ and $\delta = 0$, each $\alpha_{r;m}(f)$ is bounded by $\alpha_r(f)$, we have that $\lambda_2(n,k)$ is bounded from above by $\lambda_h(n,k)$, which recovers the result of Theorem 6.1. □

Next, we present generalisations of the corollaries in Section 6.3. Corollary 6.2 concerns convergence of variance, which can be proved by taking h to be the identity function on $\mathbb{R}$, replacing f with coordinates of f, $f_r(\bullet)$ and $f_r(\bullet)f_s(\bullet)$ for $r, s \leq q$, and multiplying across by the scale n. We again present a more general result in terms of $\mathcal{Z}^\delta$ and noise stability terms $\alpha_{r;m}$ defined in Theorem D.1, of which Corollary 6.2 is then an immediate consequence:

Lemma D.2 (Variance result, generalised)**.** *Assume the conditions of Theorem D.1, then*

$$\begin{aligned} n\left\|\mathrm{Var}[f(\Phi\mathcal{X})] - \mathrm{Var}[f(\mathcal{Z}^\delta)]\right\| \;&\leq 4\delta n^2 k^{1/2}(\alpha_{0;4}\alpha_{2;4} + \alpha_{1;4}^2)c_1 \\ &\quad + 6n^2k^{3/2}(\alpha_{0;4}\alpha_{3;4} + \alpha_{1;4}\alpha_{2;4})(c_X + c_{Z^\delta}) \,. \end{aligned}$$

Proof of Corollary 6.2. Since $\alpha_{r;m}(f) \leq \alpha_r(f)$ for $m \leq 12$ and $\delta = 0$, the second term in the bound in Lemma D.2 can be further bounded from above by the desired quantity

$$6n^2k^{3/2}(\alpha_0\alpha_3 + \alpha_1\alpha_2)(c_X + c_Z) \,.$$

Setting $\delta = 0$ recovers $\{\mathbf{Z}_i\}_{i=1}^n$ from $\mathcal{Z}^\delta$ and causes the first term to vanish, which recovers Corollary 6.2. □

Corollary 6.4 concerns convergence in d_H. We present a tighter bound below:

Lemma D.3 ($d_{\mathcal{H}}$ result, generalised)**.** *Assume the conditions of Theorem 6.1, then*

$$\begin{aligned} &d_{\mathcal{H}}(\sqrt{n}f(\Phi\mathcal{X}), \sqrt{n}f(\mathcal{Z}^\delta)) \\ &\leq\; \delta n^{3/2}k^{1/2}c_1\left(n^{1/2}\alpha_{1;2}^2 + \alpha_{2;1}\right) + (nk)^{3/2}(n\alpha_{1;6}^3 + 3n^{1/2}\alpha_{1;4}\alpha_{2;4} + \alpha_{3;2})(c_X + c_{Z^\delta}) \,. \end{aligned}$$

Proof of Corollary 6.4. We again note that setting $\delta = 0$ recovers $\{\mathbf{Z}_i\}_{i=1}^n$ from $\mathcal{Z}^\delta$ and c_Z from c_{Z^δ}. The required bound is obtained by setting $\delta = 0$ and bounding each $\alpha_{r;m}$ term by α_r in the result in Lemma D.3:

$$(nk)^{3/2}(n(\alpha_{1;6})^3 + 3n^{1/2}\alpha_{1;4}\alpha_{2;4} + \alpha_{3;2}) \;\leq\; (nk)^{3/2}(n\alpha_1^3 + 3n^{1/2}\alpha_1\alpha_2 + \alpha_3)(c_X + c_Z)\;.$$

□

As discussed in the main text, the result for no augmentations in (6.7) is immediate from setting the augmentations ϕ_{ij} to identity almost surely in Theorem 6.1. Equivalent versions of Lemma D.2 and Lemma D.3 for no augmentation can be obtained similarly, and the statements are omitted here. This means that to compare the case with augmentation versus the case without, we only need to check the conditions of Lemma D.2 and Lemma D.3 once.

D.1.2. Results corresponding to Remark 6.1

As mentioned in Remark 6.1(ii), one may allow q to grow with n and k. While Corollary 6.2 still applies if q grows sufficiently slowly, Lemma 6.3 does not apply unless q is fixed. The following lemma is a substitute. As is typical in high-dimensional settings, we focus on studying the convergence of f_s, a fixed s-th coordinate of f for $s \leq q$. The lemma gives a sufficient condition on f for convergence of variance for f and convergence in $d_{\mathcal{H}}$ for f_s to hold when q grows with n and k.

Lemma D.4. *Assume the conditions of Theorem 6.1 and fix $s \leq q$. Assume that coordinates of $\phi_{11}\mathbf{X}_1$ and $\mathbf{Z}_1$ are $O(1)$ a.s., $\alpha_1 = o(n^{-5/6}(kd)^{-1/2})$, $\alpha_3 = o((nkd)^{-3/2})$ and $\alpha_0\alpha_3, \alpha_1\alpha_2 = o(n^{-2}(kd)^{-3/2})$, either as n, d, q grow with k fixed or as n, d, q, k all grow. Then under the same limit,*

$$d_{\mathcal{H}}(\sqrt{n}f_s(\Phi\mathcal{X}), \sqrt{n}f_s(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)) \xrightarrow{d} \mathbf{0}\;, \quad n\|\mathrm{Var}[f(\Phi\mathcal{X})] - \mathrm{Var}[f(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)]\| \to 0\;.$$

The proof is a straightforward result from Corollary 6.2, Corollary 6.4 and Lemma 6.3. In practice, one may want to use Lemma D.2 and Lemma D.3 directly for tighter controls on moments and noise stability, which is the method we choose for the derivation of examples in Appendix D.6.

Remark 6.1(iii) discusses the setting where data is distributionally invariant to augmentations. In this case, Theorem D.1 becomes:

Corollary D.5 ($\mathcal{T}$-invariant data source)**.** *Assume the conditions of Theorem D.1 and $\phi\mathbf{X} \overset{d}{=} \mathbf{X}$ for every $\phi \in \mathcal{T}$. Then*

$$\left|\mathbb{E}h(f(\Phi\mathcal{X})) - \mathbb{E}h(f(\mathbf{Z}_1^\delta, \ldots, \mathbf{Z}_n^\delta))\right| \;\leq\; \delta nk^{1/2}\lambda_1(n,k)c_1 + nk^{3/2}\lambda_2(n,k)(c_X + c_{Z^\delta})\;,$$

where $\mathbf{Z}_1^\delta, \ldots, \mathbf{Z}_n^\delta$ *are i.i.d. variables satisfying*

$$\mathbb{E}\mathbf{Z}_i^\delta = \mathbf{1}_{k\times 1} \otimes \mathbb{E}[\phi_{11}\mathbf{X}_1], \quad \mathrm{Var}\mathbf{Z}_i^\delta = \mathbf{I}_k \otimes \left((1-\delta)\tilde{\Sigma}_{11} + \delta\Sigma_{12}\right) + (\mathbf{1}_{k\times k} - \mathbf{I}_k) \otimes \Sigma_{12},$$

where we have denoted

$$\begin{aligned}\tilde{\Sigma}_{11} &:= \mathbb{E}\mathrm{Var}[\phi_{11}\mathbf{X}_1|\phi_{11}] \,, \\ \Sigma_{12} &:= \mathrm{Cov}[\phi_{11}\mathbf{X}_1, \phi_{12}\mathbf{X}_1] = \mathbb{E}\mathrm{Cov}[\phi_{11}\mathbf{X}_1, \phi_{12}\mathbf{X}_1|\phi_{11}, \phi_{12}] \,.\end{aligned}$$

This result is connected to results on central limit theorem under group invariance (Austern and Orbanz, 2022), by observing that when $\mathcal{T}$ is a group, the distribution of $\mathbf{Z}$ is described exactly by group averages. We also note that since $\tilde{\Sigma}_{11} \preceq \Sigma_{11}$, the invariance assumption leads to a reduction in *data* variance, although this does not imply reduction in variance in the estimate f. Finally, the invariance assumption implies $\mathbb{E}[\phi_{11}\mathbf{X}_1] = \mathbb{E}[\mathbf{X}_1]$, in which case the augmented estimate $f(\Phi\mathcal{X})$ is a consistent estimate of the unaugmented estimate $f(\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_n)$.

Remark 6.1(iii) says that a stricter condition on f that typically requires k to grow recovers a variance structure resembling that observed in Chen et al. (2020): variance of an conditional average taken over the distribution of augmentations. This is obtained directly by setting $\delta = 1$ in Theorem D.1 and noting that, by Lemma D.19, $\mathrm{Cov}[\phi_{11}\mathbf{X}_1, \phi_{12}\mathbf{X}_1] = \mathrm{Var}\mathbb{E}[\phi_{11}\mathbf{X}_1|\mathbf{X}_1]$:

Corollary D.6 (Smaller data variance)**.** *Assume the conditions of Theorem D.1 with* $\delta = 1$*. Then*

$$\left|\mathbb{E}h(f(\Phi\mathcal{X})) - \mathbb{E}h(f(\mathbf{Z}_1, \ldots, \mathbf{Z}_n))\right| \;\leq\; nk^{1/2}\lambda_1(n,k)c_1 + nk^{3/2}\lambda_2(n,k)(c_X + c_Z) \,,$$

where $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ *are i.i.d. variables satisfying*

$$\mathbb{E}\mathbf{Z}_i = \mathbf{1}_{k\times 1} \otimes \mathbb{E}[\phi_{11}\mathbf{X}_1], \qquad \mathrm{Var}\mathbf{Z}_i = \mathbf{1}_{k\times k} \otimes \mathrm{Var}\mathbb{E}[\phi_{11}\mathbf{X}_1|\mathbf{X}_1] \,.$$

Note that the data variance is smaller than that in Theorem D.1 in the following sense: By Lemma D.19, $\mathrm{Var}[\phi_{11}\mathbf{X}_1] \succeq \mathrm{Cov}[\phi_{11}\mathbf{X}_1, \phi_{12}\mathbf{X}_1] = \mathrm{Var}\mathbb{E}[\phi_{11}\mathbf{X}_1|\mathbf{X}_1]$ and therefore we have $\mathbf{I}_k \otimes (\mathrm{Var}[\phi_{11}\mathbf{X}_1] - \mathrm{Cov}[\phi_{11}\mathbf{X}_1, \phi_{12}\mathbf{X}_1]) \succeq \mathbf{0}$. This implies $\mathrm{Var}\mathbf{Z}_i$ in Corollary D.6 can be compared to that in Theorem D.1 by

$$\begin{aligned}\mathbf{1}_{k\times k} \otimes \mathrm{Var}\mathbb{E}[\phi_{11}\mathbf{X}_1|\mathbf{X}_1] &= \mathbf{1}_{k\times k} \otimes \mathrm{Cov}[\phi_{11}\mathbf{X}_1, \phi_{12}\mathbf{X}_1] \\ &\preceq \mathbf{I}_k \otimes \mathrm{Var}[\phi_{11}\mathbf{X}_1] + (\mathbf{1}_{k\times k} - \mathbf{I}_k) \otimes \mathrm{Cov}[\phi_{11}\mathbf{X}_1, \phi_{12}\mathbf{X}_1] \,.\end{aligned}$$

The stricter condition on f comes from the fact that, for the bound to decay to zero, on top of requiring $\lambda_2(n,k)$ to be $o(n^{-1}k^{-3/2})$, we also require $\lambda_1(n,k)$ to be $o(nk^{-1/2})$. In the case of empirical average in Proposition 6.7, one may compute that $\lambda_1(n,k) = \gamma^2(h)n^{-1}k^{-1}$, so a smaller *data* variance is only obtained when we require $k \to \infty$.

D.1.3. Plug-in estimates g(empirical average)

We present convergence results that compare $f(\Phi\mathcal{X}) \coloneqq g(\text{empirical average})$ to two other statistics. One of them is $f(\mathcal{Z}^\delta)$, which is already discussed in Theorem D.1, and the other one is the limit discussed in (6.12), which is the following truncated first-order Taylor expansion:

$$f^T(x_{11},\dots,x_{nk}) \coloneqq g(\mathbb{E}[\phi_{11}\mathbf{X}_1]) + \partial g(\mathbb{E}[\phi_{11}\mathbf{X}_1])\big(\tfrac{1}{nk}\textstyle\sum_{i=1}^n\sum_{j=1}^k \mathbf{x}_{ij} - \mathbb{E}[\phi_{11}\mathbf{X}_1]\big)\ .$$

Since we need to study the convergence towards a first-order Taylor expansion of g, we need to define variants of noise stability terms in terms of g. Given $\{\phi_{ij}\mathbf{X}_i\}_{i\le n,j\le k}$ and $\{\mathbf{Z}_i^\delta\}_{i\le n} \coloneqq \{\mathbf{Z}_{ij}^\delta\}_{i\le n,j\le k}$, denote the mean and centred sums

$$\mu \coloneqq \mathbb{E}[\phi_{11}\mathbf{X}_1]\ , \qquad \bar{\mathbf{X}} \coloneqq \frac{1}{nk}\textstyle\sum_{i,j}\phi_{ij}\mathbf{X}_i - \mu\ , \qquad \bar{\mathbf{Z}}^\delta \coloneqq \frac{1}{nk}\textstyle\sum_{i,j}\mathbf{Z}_{ij}^\delta - \mu\ .$$

For a function $g:\mathcal{D}\to\mathbb{R}^q$ and $s\le q$, we denote the s^{th} coordinate of $g(\,\bullet\,)$ as $g_s(\,\bullet\,)$ as before, and define a new noise stability term controlling the noise from first-order Taylor expansion around μ:

$$\kappa_{r;m}(g) \coloneqq \textstyle\sum_{s\le q}\big\|\sup_{\mathbf{w}\in[\mathbf{0},\bar{\mathbf{X}}]}\big\|\partial^r g_s\big(\mu+\mathbf{w}\big)\big\|\big\|_{L_m}\ .$$

The first-order Taylor expansion also introduces additional moment terms, which is controlled by Rosenthal's inequality from Corollary D.22 and bounded in terms of:

$$\bar{c}_m \coloneqq \Big(\textstyle\sum_{s=1}^d \max\Big\{n^{\frac{2}{m}-1}\big\|\frac{1}{k}\sum_{j=1}^k(\phi_{1j}\mathbf{X}_1-\mu)_s\big\|_{L_m}^2,\ \big\|\frac{1}{k}\sum_{j=1}^k(\phi_{1j}\mathbf{X}_1-\mu)_s\big\|_{L_2}^2\Big\}\Big)^{1/2}\ .$$

Finally, since we will compare $f(\Phi\mathcal{X})$ to $f(\mathcal{Z}^\delta)$, we consider noise stability terms that resemble $\alpha_{r;m}$ from Theorem D.1 but expressed in terms of g:

$$\begin{aligned}\nu_{r;m}(g) \coloneqq &\sum_{s\le q}\max_{i\le n}\max\Big\{\Big\|\sup_{\mathbf{w}\in[\mathbf{0},\Phi_i\mathbf{X}_i]}\|\partial^r g_s(\overline{\mathbf{W}}_i^\delta(\mathbf{w}))\|\Big\|_{L_m}\ ,\\ &\qquad\qquad\Big\|\sup_{\mathbf{w}\in[\mathbf{0},\mathbf{Z}_i^\delta]}\|\partial^r g_s(\overline{\mathbf{W}}_i^\delta(\mathbf{w}))\|\Big\|_{L_m}\Big\}\\ =&\textstyle\sum_{s\le q}\max_{i\le n}\zeta_{i;m}\big(\big\|\partial^r g_s\big(\overline{\mathbf{W}}_i^\delta(\,\bullet\,)\big)\big\|\big) \ge \max_{i\le n}\zeta_{i;m}\big(\big\|\partial^r g\big(\overline{\mathbf{W}}_i^\delta(\,\bullet\,)\big)\big\|\big)\ ,\end{aligned} \tag{D.3}$$

where

$$\overline{\mathbf{W}}_i^\delta(\mathbf{w}) \coloneqq \frac{1}{nk}\big(\textstyle\sum_{i'=1}^{i-1}\sum_{j=1}^k\phi_{i'j}\mathbf{X}_{i'} + \sum_{j=1}^k\mathbf{w}_j + \sum_{i'=i+1}^n\sum_{j=1}^k\mathbf{Z}_{i'j}^\delta\big)\ .$$

We omit g-dependence in $\kappa_{r;m}$ and $\nu_{r;m}$ whenever the choice of g is obvious.

Lemma D.7. *(Plug-in estimates) Assume the conditions of Theorem D.1. For $g\in\mathcal{F}_3(\mathcal{D},\mathbb{R}^q)$, define the plug-in estimate $f(\mathbf{x}_{11:nk}) = g\big(\frac{1}{nk}\sum_{i\le n,j\le k}\mathbf{x}_{ij}\big)$ and its Taylor expansion $f^T(\mathbf{x}_{11:nk})$ as in (6.11). Then, for any $\mathcal{Z}^\delta$ satisfying the conditions of Theorem D.1,*

(i) the following bounds hold concerning the approximation by $f^T(\mathcal{Z}^\delta)$:

$$\begin{aligned} &d_{\mathcal{H}}\big(\sqrt{n}f(\Phi\mathcal{X}), \sqrt{n}\, f^T(\mathcal{Z}^\delta)\big) \\ &\qquad = O\big(n^{-1/2}\kappa_{2;3}\,\bar{c}_3^2 + \delta k^{-1/2}\kappa_{1;1}^2 c_1 + n^{-1/2}\kappa_{1;1}^3(c_X + c_{Z^\delta})\big)\ , \\ &n\big\|\mathrm{Var}[f(\Phi\mathcal{X})] - \mathrm{Var}\big[f^T(\mathcal{Z}^\delta)\big]\big\| \\ &\qquad = O\big(\delta k^{-1}\|\partial g(\mu)\|_2^2\, c_1^2 + n^{-1/2}\kappa_{1;1}\kappa_{2;4}\bar{c}_4^3 + n^{-1}\kappa_{2;6}^2\bar{c}_6^4\big)\ . \end{aligned}$$

(ii) the following bounds hold concerning the approximation by $f(\mathcal{Z}^\delta)$:

$$\begin{aligned} d_{\mathcal{H}}(\sqrt{n}f(\Phi\mathcal{X}), \sqrt{n}f(\mathcal{Z}^\delta)) \ &= O\big(\delta\big(k^{-1/2}\nu_{1;2}^2 + n^{-1/2}k^{-1/2}\nu_{2;1}\big)c_1 \\ &\quad + \big(n^{-1/2}\nu_{1;6}^3 + 3n^{-1}\nu_{1;4}\nu_{2;4} + n^{-3/2}\nu_{3;2}\big) \\ &\quad \times (c_X + c_{Z^\delta})\big)\ , \\ n\|\mathrm{Var}[f(\Phi\mathcal{X})] - \mathrm{Var}[f(\mathcal{Z}^\delta)]\| \ &= O\big(\delta k^{-1/2}(\nu_{0;4}\nu_{2;4} + \nu_{1;4}^2)c_1 \\ &\quad + n^{-1}(\nu_{0;4}\nu_{3;4} + \nu_{1;4}\nu_{2;4})(c_X + c_{Z^\delta})\big)\ . \end{aligned}$$

Remark D.1. The statement in (6.12) in the main text is obtained from Lemma D.7(i) by fixing q, setting $\delta = 0$ and requiring the bounds to go to 0, which is a noise stability assumption on g and a constraint on how fast d is allowed to grow. Weak convergence can again be obtained from convergence in $d_{\mathcal{H}}$ by Lemma 6.3.

D.1.4. Non-smooth statistics in high dimensions

Our results can be extended to non-smooth statistics that are well approximated by a sequence of smooth ones. To deal with high dimensions, instead of using moment terms that involve a vector-2 norm of a high-dimensional vector, we use moment terms involving a vector-∞ norm. This gives the following variant of Theorem D.1:

Theorem D.8. *(Main result adapted for non-smooth statistics in high dimensions) For a function $f : \mathcal{D}^{nk} \to \mathbb{R}^q$, suppose exists a sequence of functions $f^{(t)} \in \mathcal{F}_3(\mathcal{D}^{nk}, \mathbb{R}^q)$ that satisfy*

$$\varepsilon(t) \ := \ \max\big\{\|\,\|f(\Phi\mathcal{X}) - f^{(t)}(\Phi\mathcal{X})\|\,\|_{L_1}, \|\,\|f(\mathcal{Z}) - f^{(t)}(\mathcal{Z})\|\,\|_{L_1}\big\} \xrightarrow{t\to\infty} 0\ .$$

Let $\Phi\mathcal{X}$ and $\mathcal{Z}$ be defined as in Theorem D.1 with $\delta = 0$, and let $\|\bullet\|_1$ be the vector-1 norm. Then,

$$\begin{aligned} \big|\mathbb{E}h(f(\Phi\mathcal{X})) - \mathbb{E}h(f(\mathcal{Z}))\big| \ &\le nk^{3/2}(\tilde{c}_X + \tilde{c}_Z) \\ &\quad \times \big(\gamma_3(h)(\tilde{\alpha}_1^{(t)})^3 + 3\gamma_2(h)\tilde{\alpha}_1^{(t)}\tilde{\alpha}_2^{(t)} + \gamma_1(h)\tilde{\alpha}_3^{(t)}\big) \\ &\quad + 2\gamma_1(h)\varepsilon(t)\ . \end{aligned}$$

Here, we have defined the new noise stability terms in terms of vector-1 norm $\|\bullet\|_1$ *as*

$$\tilde{\alpha}_r^{(t)} := \sum_{s=1}^{q} \max_{i\le n} \max \Bigg\{ \Bigg\| \sup_{\mathbf{w}\in[\mathbf{0},\Phi_i\mathbf{X}_i]} \Bigg(\sum_{j_1,\ldots,j_r=1}^{k} \Big\| \frac{\partial^r}{\partial \mathbf{x}_{ij_1}\ldots\partial \mathbf{x}_{ij_r}} f_s^{(t)}(\mathbf{W}_i(\mathbf{w})) \Big\|_1^2 \Bigg)^{1/2} \Bigg\|_{L_6}, \\ \Bigg\| \sup_{\mathbf{w}\in[\mathbf{0},\mathbf{Z}_i]} \Bigg(\sum_{j_1,\ldots,j_r=1}^{k} \Big\| \frac{\partial^r}{\partial \mathbf{x}_{ij_1}\ldots\partial \mathbf{x}_{ij_r}} f_s^{(t)}(\mathbf{W}_i(\mathbf{w})) \Big\|_1^2 \Bigg)^{1/2} \Bigg\|_{L_6} \Bigg\},$$

where $f_s^{(t)}$ *is the* s*th coordinate of* $f^{(t)}$*, and the new moment terms as*

$$\tilde{c}_X := \frac{1}{6}\sqrt{\mathbb{E}\big[\max_{l\le d}|(\phi_{11}\mathbf{X}_i)_l|^6\big]}, \quad \tilde{c}_Z := \frac{1}{6}\sqrt{\mathbb{E}\Big[\frac{1}{k}\sum_{j=1}^{k}\max_{l\le d}|(\mathbf{Z}_{1j})_l|^6\Big]}.$$

A similar argument to Lemma D.7(ii) then yields an analogue of the result for plugin-estimates, which is useful for the derivation of maximum of exponentially many averages in Appendix D.6.5:

Corollary D.9. *For a function* $g : \mathcal{D} \to \mathbb{R}^q$*, suppose there exists a sequence of functions* $g^{(t)} \in \mathcal{F}_3(\mathcal{D}^{nk}, \mathbb{R}^q)$ *that satisfy* $\varepsilon(t) \xrightarrow{t\to\infty} 0$ *for*

$$\varepsilon(t) := \max\Big\{ \Big\| \Big\| g\big(\frac{1}{nk}\sum_{i,j}\phi_{ij}\mathbf{X}_i\big) - g^{(t)}\big(\frac{1}{nk}\sum_{i,j}\phi_{ij}\mathbf{X}_i\big)\Big\| \Big\|_{L_2}, \\ \Big\| \Big\| g\big(\frac{1}{nk}\sum_{i,j}\mathbf{Z}_{ij}\big) - g^{(t)}\big(\frac{1}{nk}\sum_{i,j}\mathbf{Z}_{ij}\big)\Big\| \Big\|_{L_2} \Big\}.$$

Define the plug-in estimate $f(\mathbf{x}_{11:nk}) = g\big(\frac{1}{nk}\sum_{i\le n, j\le k}\mathbf{x}_{ij}\big)$*. Then, the following convergences hold:*

$$\begin{aligned} d_{\mathcal{H}}(\sqrt{n}f(\Phi\mathcal{X}), \sqrt{n}f(\mathcal{Z})) &= O\big((n^{-1/2}(\tilde{\nu}_1^{(t)})^3 + 3n^{-1}\tilde{\nu}_1^{(t)}\tilde{\nu}_2^{(t)} + n^{-3/2}\tilde{\nu}_3^{(t)})(\tilde{c}_X + \tilde{c}_Z) \\ &\quad + \sqrt{n}\varepsilon(t)\big), \\ n\|\mathrm{Var}[f(\Phi\mathcal{X})] - \mathrm{Var}[f(\mathcal{Z})]\| &= O\big(n^{-1}(\tilde{\nu}_0^{(t)}\tilde{\nu}_3^{(t)} + \tilde{\nu}_1^{(t)}\tilde{\nu}_2^{(t)})(\tilde{c}_X + \tilde{c}_Z) \\ &\quad + n(\|\,\|f(\Phi\mathcal{X})\|\,\|_{L_2} + \|\,\|f(\mathcal{Z})\|\,\|_{L_2})\varepsilon(t) + n\varepsilon(t)^2\big). \end{aligned}$$

Here, we have used the moment terms $\tilde{c}_X$ *and* $\tilde{c}_Z$ *from Theorem D.8 and the modified noise stability terms as*

$$\tilde{\nu}_r^{(t)} := \sum_{s=1}^{q} \max_{i\le n} \max \Big\{ \Big\| \sup_{\mathbf{w}\in[\mathbf{0},\Phi_i\mathbf{X}_i]} \|\partial^r g_s^{(t)}(\overline{\mathbf{W}}_i(\mathbf{w}))\|_1 \Big\|_{L_6}, \\ \Big\| \sup_{\mathbf{w}\in[\mathbf{0},\mathbf{Z}_i]} \|\partial^r g_s^{(t)}(\overline{\mathbf{W}}_i(\mathbf{w}))\|_1 \Big\|_{L_6} \Big\},$$

where $g_s^{(t)}$ *is the* s*th coordinate of* $g^{(t)}$*.*

For smooth statistics, a similar argument to Lemma D.7(i) also yields an analogue of the result for plugin-estimates compared to first-order Taylor expansion in high dimensions. The following result is useful for the derivation of softmax ensemble in Appendix D.6.6:

Corollary D.10. *For a function $g \in \mathcal{F}_3(\mathcal{D}, \mathbb{R}^q)$, define $f(\mathbf{x}_{11:nk}) = g\big(\frac{1}{nk}\sum_{i\le n, j\le k}\mathbf{x}_{ij}\big)$ as the plug-in estimate. Let $\Phi\mathcal{X}$ and $\mathcal{Z}$ be defined as in Theorem D.1 with $\delta = 0$, and the first-order Taylor expansion f^T be defined by (6.11). Let $\tilde{c}_X$ and $\tilde{c}_Z$ be defined as in Theorem D.8. Then if $\log d = o(n^\alpha)$ for some $\alpha \ge 0$, the following bounds hold:*

$$\begin{aligned} d_{\mathcal{H}}(\sqrt{n}f(\Phi\mathcal{X}), \sqrt{n}f^T(\mathcal{Z})) &= o\big(n^{-1/2+\alpha}\tilde{\kappa}_2 \max\{1, (\tilde{c}_X)^{4/3}\}\big) \\ &\quad + O\big(n^{-1/2}\tilde{\kappa}_1^3(\tilde{c}_X + \tilde{c}_Z)\big)\ , \\ n\big\|\mathrm{Var}[f(\Phi\mathcal{X})] - \mathrm{Var}\big[f^T(\mathcal{Z})\big]\big\| &= o\big((n^{-1/2+3\alpha/2}\tilde{\kappa}_1\tilde{\kappa}_2 + n^{-1+2\alpha}\tilde{\kappa}_2^2)\max\{1, (\tilde{c}_X)^{8/3}\}\big)\ , \end{aligned}$$

We have defined the modified noise stability terms in terms of the expectation $\mu = \mathbb{E}[\phi_{11}\mathbf{X}_1]$ and the centred average $\bar{\mathbf{X}} := \frac{1}{nk}\sum_{i,j}\phi_{ij}\mathbf{X}_i - \mu$ as

$$\tilde{\kappa}_r := \sum_{s\le q}\big\|\sup_{\mathbf{w}\in[\mathbf{0},\bar{\mathbf{X}}]}\big\|\partial^r g_s(\mu + \mathbf{w})\big\|_1\big\|_{L_6}\ .$$

D.1.5. Repeated augmentation

In Theorem 6.1, each transformation is used once and then discarded. A different strategy is to generate only k transformations i.i.d., and apply each to all n observations. That introduces additional dependence: In the notation of Section 6.2, $\Phi_i\mathbf{X}_i$ and $\Phi_j\mathbf{X}_j$ are no longer independent if $i \ne j$. The next result adapts Theorem 6.1 to this case. We require that f satisfies

$$f(\mathbf{x}_{11}, \ldots, \mathbf{x}_{1k}, \ldots, \mathbf{x}_{n1}, \ldots, \mathbf{x}_{nk}) = f(\mathbf{x}_{1\pi_1(1)}, \ldots, \mathbf{x}_{1\pi_1(k)}, \ldots, \mathbf{x}_{n\pi_n(1)}, \ldots, \mathbf{x}_{n\pi_n(k)}) \tag{D.4}$$

for any permutations $\pi_1, \ldots, \pi_n$ of k elements. That holds for most statistics of practical interest, including empirical averages and M-estimators.

Theorem D.11. (Repeated Augmentation) *Assume the conditions in Theorem 6.1 with $\mathcal{D} = \mathbb{R}^d$ and that f satisfies (D.4). Define $\tilde{\Phi} := (\phi_{ij} | i \le n, j \le k)$, where $\phi_{1j} = \ldots = \phi_{nj} =: \phi_j$ and $\phi_1, \ldots, \phi_k$ are i.i.d. random elements of $\mathcal{T}$. Then there are random variables $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ in $\mathbb{R}^{kd}$ such that*

$$\begin{aligned} &\big|\mathbb{E}h(f(\tilde{\Phi}\mathcal{X})) - \mathbb{E}h(f(\mathbf{Y}_1, \ldots, \mathbf{Y}_n))\big| \\ &\qquad \le n\gamma_1(h)\alpha_1 m_1 + n\omega_2(n,k)(\gamma_2(h)\alpha_1^2 + \gamma_1(h)\alpha_2) + nk^{3/2}\lambda_h(n,k)(c_X + c_Y)\ . \end{aligned}$$

Here, λ, c_X and α_r are defined as in Theorem 6.1, and c_Y is defined in a way analogous to c_Z:

$$c_Y := \frac{1}{6}\sqrt{\mathbb{E}\Big[\Big(\frac{|Y_{111}|^2 + \ldots + |Y_{1kd}|^2}{k}\Big)^3\Big]}$$

The additional constant moment terms are defined by $m_1 \coloneqq \sqrt{2\mathrm{Tr}\mathrm{Var}\mathbb{E}[\phi_1\mathbf{X}_1|\phi_1]}$, *and*

$$m_2 \coloneqq \sqrt{\sum_{r,s\leq d} \frac{\mathrm{Var}\mathbb{E}\left[(\phi_1\mathbf{X}_1)_r(\phi_1\mathbf{X}_1)_s\middle|\phi_1\right]}{2}},\ m_3 \coloneqq \sqrt{\sum_{r,s\leq d} 12\mathrm{Var}\mathbb{E}\left[(\phi_1\mathbf{X}_1)_r(\phi_2\mathbf{X}_1)_s\middle|\phi_1,\phi_2\right]}\ .$$

The variables $\mathbf{Y}_i$ *are conditionally i.i.d. Gaussian vectors with mean* $\mathbb{E}[\Psi\mathbf{X}_1|\Psi_1]$ *and covariance matrix* $\mathrm{Var}[\Psi\mathbf{X}_1|\Psi_1]$, *conditioning on* $\Psi \coloneqq \{\psi_1,\dots,\psi_k\}$ *i.i.d. distributed as* $\{\phi_1,\dots,\phi_k\}$.

The result shows that the additional dependence introduced by using transformations repeatedly does not vanish as n and k grow. Unlike the Gaussian limit in Theorem 6.1 (when $\mathcal{D}$ is taken as $\mathbb{R}^d$), the limit here is characterized by variables $\mathbf{Y}_i$ that are only *conditionally* Gaussian, given an i.i.d. copy of the augmentations. That further complicates the effects of augmentation. Indeed, there exist statistics f for which i.i.d. augmentation as in Theorem 6.1 does not affect the variance, but repeated augmentation either increases or decreases it. Lemma D.12 gives such an example: Even when distributional invariance holds, augmentation may increase variance for one statistic and decrease variance for the other.

Lemma D.12. *Consider i.i.d. random vectors* $\mathbf{X}_1, \mathbf{X}_2$ *in* $\mathbb{R}^d$ *with mean* μ *and* $\phi_1, \phi_2 \in \mathbb{R}^{d\times d}$ *be i.i.d. random matrices such that* $\phi_1\mathbf{X}_1 \overset{d}{=} \mathbf{X}_1$. *Then for* $f_1(\mathbf{x}_1,\mathbf{x}_2) \coloneqq \mathbf{x}_1 + \mathbf{x}_2$ *and* $f_2(\mathbf{x}_1,\mathbf{x}_2) \coloneqq \mathbf{x}_1 - \mathbf{x}_2$,

(i) $\mathrm{Var} f_1(\mathbf{X}_1,\mathbf{X}_2) = \mathrm{Var} f_1(\phi_1\mathbf{X}_1,\phi_2\mathbf{X}_2) \preceq \mathrm{Var} f_1(\phi_1\mathbf{X}_1,\phi_1\mathbf{X}_2)$, *and*

(ii) $\mathrm{Var} f_2(\mathbf{X}_1,\mathbf{X}_2) = \mathrm{Var} f_2(\phi_1\mathbf{X}_1,\phi_2\mathbf{X}_2) \succeq \mathrm{Var} f_2(\phi_1\mathbf{X}_1,\phi_1\mathbf{X}_2)$.

D.2 Additional results for the examples

D.2.1. Results for the toy statistic

In this section, we present results concerning the toy statistic defined in (6.13). For convenience, we write $f \equiv f_{\mathrm{toy}}$. To express variances concisely, we define the function $V(s) \coloneqq (1+4s^2)^{-1/2} - (1+2s^2)^{-1}$, and write

$$\tilde{\sigma} \coloneqq \sqrt{\mathrm{Var}[\mathbf{X}_1]} \qquad \text{and} \qquad \sigma \coloneqq \left(\frac{1}{k}\mathrm{Var}[\phi_{11}\mathbf{X}_1] + \frac{k-1}{k}\mathrm{Cov}[\phi_{11}\mathbf{X}_1,\phi_{12}\mathbf{X}_1]\right)^{1/2}.$$

The next result applies Theorem 6.1 to derive closed-form formula for the quantities plotted in Figure 6.3:

Proposition D.13. *Require that* $\mathbb{E}[\mathbf{X}_1] = \mathbb{E}[\phi_{11}\mathbf{X}_1] = 0$, *and that* $\mathbb{E}[|\mathbf{X}_1|^{12}]$ *and* $\mathbb{E}[|\phi_{11}\mathbf{X}_1|^{12}]$

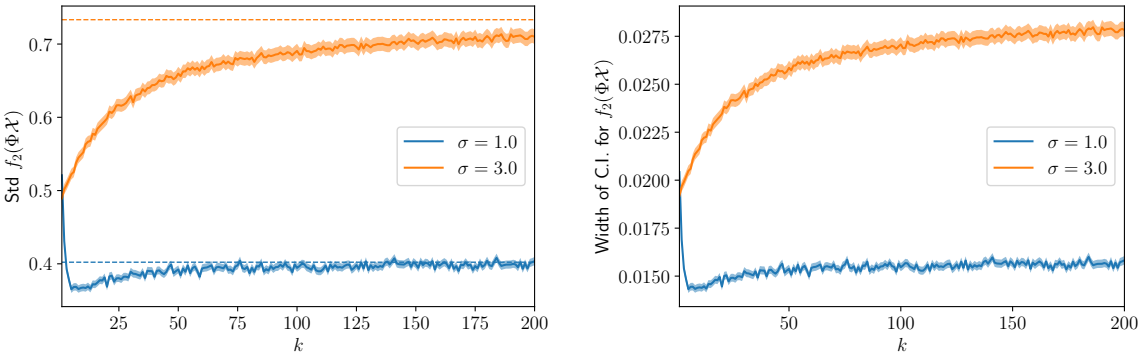


Figure D.1: Simulation for f_2 with $n = 100$ and varying k. *Left*: The standard deviation $\mathrm{Std}\, f_2(\Phi\mathcal{X})$. The dotted lines indicate the theoretical value of $\mathrm{Std}\, f_2(\mathcal{Z})$ computed in Lemma D.14, in which we also verify the convergence of $f_2(\Phi\mathcal{X})$ to $f_2(\mathcal{Z})$ in $d_{\mathcal{H}}$. *Right*: Difference between 0.025-th and 0.975-th quantiles for $f_2(\Phi\mathcal{X})$. In all figures, shaded regions denote 95% confidence intervals for simulated quantities.

are finite. Let $\mathcal{Z}, \mathcal{Z}'$ be Gaussian. Then $f \equiv f_{\text{toy}}$ defined in (6.13) satisfies

$$d_{\mathcal{H}}(f(\Phi\mathcal{X}), f(\mathcal{Z})) \to 0 \quad \textit{and} \quad \mathrm{Var}[f(\Phi\mathcal{X})] - \mathrm{Var}[f(\mathcal{Z})] \to 0 \quad \textit{as } n \to \infty$$

and the same holds in the unaugmented case where $\Phi\mathcal{X}$ and $\mathcal{Z}$ are replaced by $\tilde{\mathcal{X}}$ and $\tilde{\mathcal{Z}}$. The asymptotic variances are

$$\mathrm{Var} f(\mathcal{Z}) = V(\sigma) \quad \textit{and} \quad \mathrm{Var} f(\tilde{\mathcal{Z}}) = V(\tilde{\sigma}) \quad \textit{and hence} \quad \vartheta(f) = \sqrt{V(\tilde{\sigma})/V(\sigma)}\,.$$

For any $\alpha \in [0, 1]$, the lower and upper $\alpha/2$-th quantiles for $f(\mathcal{Z})$ and $f(\tilde{\mathcal{Z}})$ are given by

$$\big(\exp\big(-\sigma^2\pi_u\big), \exp\big(-\sigma^2\pi_l\big)\big) \qquad \textit{and} \qquad \big(\exp(-\tilde{\sigma}^2\pi_u), \exp(-\tilde{\sigma}^2\pi_l)\big)\,,$$

where π_u and π_l are the upper and lower $\alpha/2$-th quantiles of a χ^2_1 random variable.

As discussed in the main text, the behavior of f under augmentation is more complicated than that of averages as both $V(s)$ and $D(s) := \exp(-s^2\pi_l) - \exp(-s^2\pi_u)$ are not monotonic. This phenomenon persists if we extends f to two dimensions, by defining

$$f_2(\mathbf{x}_{11}, \ldots, \mathbf{x}_{nk}) \;:=\; f(x_{111}, \ldots, x_{nk1}) + f(x_{112}, \ldots, x_{nk2})\,. \tag{D.5}$$

Figure D.1 shows results for

$$\mathbf{X}_i \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2 \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right))\,, -1 < \rho < 1\,, \quad \text{and} \quad \phi_{ij} \overset{i.i.d.}{\sim} \mathrm{Uniform}\{\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right), \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)\} \tag{D.6}$$

under $\rho = 0.5$. In this case, the data distribution is invariant under both possible transformations. Thus, invariance does not guarantee augmentation to be well-behaved.

For completeness, we also include Lemma D.14, a result that confirms the applicability of Theorem 6.1 to f_2. We also compute an explicit formula for the variances of $f(\mathcal{Z})$ and $f(\tilde{\mathcal{Z}})$ under (D.6) for a general ρ.

Lemma D.14. *Under the setting (D.6), the statistic f_2 defined in (D.5) satisfies*

(i) as $n \to \infty$, $f_2(\Phi\mathcal{X}) - f_2(\mathcal{Z}) \xrightarrow{d} 0$ and $\|\mathrm{Var}[f_2(\Phi\mathcal{X})] - \mathrm{Var}[f_2(\mathcal{Z})]\| \to 0$, and the same holds with $(\Phi\mathcal{X}, \mathbf{Z})$ replaced by the unaugmented data and surrogates $(\tilde{\mathcal{X}}, \tilde{\mathcal{Z}})$;

(ii) $\mathbf{Z}_i$ has zero mean and covariance matrix

$$\sigma^2 \mathbf{I}_k \otimes \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right) + \frac{(1+\rho)\sigma^2}{2}(\mathbf{1}_{k\times k} - \mathbf{I}_k) \otimes \mathbf{1}_{2\times 2} \,,$$

while $\tilde{\mathbf{Z}}_i$ has zero mean and covariance matrix $\sigma^2 \mathbf{1}_{k\times k} \otimes \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$;

(iii) write $\sigma_-^2 := \frac{(1-\rho)\sigma^2}{2}$ and $\sigma_+^2 := \frac{(1+\rho)\sigma^2}{2}$, then variance of the augmented data is given by

$$\begin{aligned}\mathrm{Var}[f_2(\mathcal{Z})] \; &= 2\Big(1 + \frac{4\sigma_-^2}{k} + 4\sigma_+^2\Big)^{-1/2} + 2\Big(1 + \frac{4\sigma_-^2}{k}\Big)^{-1/2}(1 + 4\sigma_+^2)^{-1/2} \\ &\quad - 4(1 + \frac{2\sigma_-^2}{k} + 2\sigma_+^2)^{-1} \,.\end{aligned}$$

In particular, at $\rho = 0.5$, $\lim_{k\to\infty} \mathrm{Var}[f_2(\mathcal{Z})] = 4(1+3\sigma^2)^{-1/2} - 4\big(1 + \frac{3}{2}\sigma^2\big)^{-1}$.

Remark D.2. Note that (i) above only verifies the convergence under $n \to \infty$ with k fixed. Nevertheless, one may easily check that f_2 satisfies the stronger Corollary D.6 corresponding to a smaller variance of $\mathbf{Z}_i$ given by $\frac{(1+\rho)\sigma^2}{2}\mathbf{1}_{2k\times 2k}$ as $n, k \to \infty$. In that case, the asymptotic variance of the statistic is given exactly by the formula $\lim_{k\to\infty} \mathrm{Var}[f_2(\mathcal{Z})]$ in (iii) above.

D.2.2. Additional results for ridgeless regressor

This section complements Section 6.5 and provides tools for simplifying the risk of ridgeless regressors.

Notation. For $A, B \in \mathbb{R}^{d\times d}$ symmetric and $\lambda \geq 0$, we denote

$$f_\lambda^{(1)}(A) := \begin{cases} \lambda^2 \beta^\top \big(A + \lambda \mathbf{I}_d\big)^{-2} \beta & \text{for } \lambda > 0 \\ \big\| \big(A^\dagger A - \mathbf{I}_d\big)\beta \big\|^2 & \text{for } \lambda = 0 \end{cases},$$

$$f_\lambda^{(2)}(A, B) := \frac{\sigma_\epsilon^2}{n} \mathrm{Tr}\big(\big(A + \lambda \mathbf{I}_d\big)^{-2} B\big) \,, \qquad f_\lambda(A, B) := f_\lambda^{(1)}(A) + f_\lambda^{(2)}(A, B) \,,$$

where $(\bullet)^{-2}$ is a shorthand for the square of the pseudoinverse $(\bullet)^\dagger$. Observe that by a standard bias-variance decomposition as in Hastie et al. (2022), the risk under the oracle augmentations can be expressed as, for both the case $\lambda > 0$ and the case $\lambda = 0$,

$$\begin{aligned}\hat{L}_\lambda^{(\mathrm{ora})} \; &= \big\| \mathbb{E}\big[\hat{\beta}_\lambda^{(\mathrm{ora})}(\mathcal{X}) \big| \mathcal{X}\big] - \beta \big\|^2 + \mathrm{Tr}\big[\mathrm{Cov}\big[\hat{\beta}_\lambda^{(\mathrm{ora})}(\mathcal{X}) \big| \mathcal{X}\big]\big] \\ &= \big\| \big(\big(\bar{\mathbf{X}}_1 + \lambda \mathbf{I}_d\big)^\dagger \bar{\mathbf{X}}_1 - \mathbf{I}_d \big)\beta \big\|^2 + \frac{\sigma_\epsilon^2}{n} \mathrm{Tr}\big(\big(\bar{\mathbf{X}}_1 + \lambda \mathbf{I}_d\big)^\dagger \bar{\mathbf{X}}_2 \big(\bar{\mathbf{X}}_1 + \lambda \mathbf{I}_d\big)^\dagger \big)\end{aligned}$$

$$= f_\lambda^{(1)}(\bar{\mathbf{X}}_1) + f_\lambda^{(2)}(\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2) \;=\; f_\lambda(\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2)\;.$$

Throughout, we write $\mathbf{e}_l$ as the l-th standard basis vector of $\mathbb{R}^d$ and denote X_{ijl} as the l-th coordinate of $\pi_{ij}V_i$.

The general case. The next lemma approximates $f_\lambda(\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2)$ by $f_0(\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2)$ in the Lévy–Prokhorov metric d_P defined in (D.8). The proof exploits the assumption below on the distribution of the extreme eigenvalues of $\bar{\mathbf{X}}_1$, $\bar{\mathbf{X}}_2$, $\bar{\mathbf{Z}}_1$ and $\bar{\mathbf{Z}}_2$, as well as the alignment of their zero eigenspace.

Lemma D.15. *Under Assumption 6.2, if $d = O(n)$ and $\lambda > 0$, then*

$$\begin{aligned}
d_P\big(f_\lambda^{(1)}(\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2)\,,\; f_0^{(1)}(\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2)\big) &= O_{\gamma'}(\lambda^2)\;,\\
d_P\big(f_\lambda^{(2)}(\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2)\,,\; f_0^{(2)}(\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2)\big) &= O_{\gamma'}\Big(\lambda + \frac{1}{n\lambda^2}\Big)\;,\\
d_P\Big(f_\lambda(\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2)\,,\; f_0(\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2)\Big) &= O_{\gamma'}\Big(\lambda^2 + \lambda + \frac{1}{n\lambda^2}\Big)\;,\\
d_P\Big(f_\lambda(\bar{\mathbf{Z}}_1, \bar{\mathbf{Z}}_1 + E_{12})\,,\; f_0(\bar{\mathbf{Z}}_1, \bar{\mathbf{Z}}_1 + E_{12})\Big) &= O_{\gamma'}\Big(\lambda^2 + \lambda + \frac{1}{n\lambda^2}\Big)\;,
\end{aligned}$$

where $O_{\gamma'}$ indicates that the bounding constant is allowed to depend on γ.

The isotropic case. In the isotropic case, one may exploit the property of Gaussians to express $\bar{\mathbf{Z}}_1$ and $\bar{\mathbf{Z}}_2$ explicitly in terms of the same rectangular Gaussian matrix. This allows the risk to be completely characterized by moments and Stieltjes transforms of the Marchenko-Pastur law under appropriate transformations, and simplifies how the two strongly correlated matrices affects the risk. The risk formula then extends to the non-Gaussian case by our universality results. The alternative expression for $\bar{\mathbf{Z}}_1$ below also formally justifies (6.23) in the discussion in the main text.

Lemma D.16. *Assume (6.22). Fix any mutually orthogonal unit vectors $\mathbf{v}_1, \dots, \mathbf{v}_{k-1} \in \mathbb{R}^k$ such that the sum of coordinates of each $\mathbf{v}_i$ equals zero. Consider the orthogonal matrix $Q_k \in \mathbb{R}^{k\times k}$ and the diagonal matrix $D_k \in \mathbb{R}^{k\times k}$, defined as*

$$Q_k \;\coloneqq\; \begin{pmatrix} k^{-1/2} & \dots & k^{-1/2} \\ \leftarrow & \mathbf{v}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{v}_{k-1}^\top & \rightarrow \end{pmatrix} \qquad \textit{and} \qquad D_k \;\coloneqq\; \begin{pmatrix} (k+\sigma_A^2)/k & & & \\ & \sigma_A^2/k & & \\ & & \ddots & \\ & & & \sigma_A^2/k \end{pmatrix}.$$

Also define the $\mathbb{R}^{nk\times n}$ matrix

$$K \;\coloneqq\; \frac{1}{\sqrt{k}}\mathbf{I}_n \otimes \mathbf{1}_k \;=\; \frac{1}{\sqrt{k}} \begin{pmatrix} 1 & \dots & 1 & & & \\ & & & 1 & \dots & 1 & & & \\ & & & & & & \ddots & & \\ & & & & & & & 1 & \dots & 1 \end{pmatrix}^\top.$$

Then almost surely,

$$\bar{\mathbf{Z}}_1 \;= \frac{1}{n}\mathbf{H}\,(\mathbf{I}_n \otimes D_k)\mathbf{H}^\top \quad \textit{and} \quad \bar{\mathbf{Z}}_2 \;= \frac{1}{n}\mathbf{H}\,(\mathbf{I}_n \otimes D_k^{1/2}Q_k)KK^\top(\mathbf{I}_n \otimes Q_k^\top D_k^{1/2})\mathbf{H}^\top\;,$$

for some $\mathbf{H}$ *that is an* $\mathbb{R}^{d\times nk}$ *matrix with i.i.d. standard Gaussian entries. As a consequence, we have*

$$\bar{\mathbf{Z}}_1 \;=\; \frac{1}{n}\sum_{i=1}^{n}\left(\eta_{i1}\eta_{i1}^\top + \frac{\sigma_A^2}{k}\sum_{j=1}^{k}\eta_{ij}\eta_{ij}^\top\right) \;=\; \bar{\mathbf{Z}}_2 + \frac{\sigma_A^2}{nk}\sum_{i=1}^{n}\sum_{j=2}^{k}\eta_{ij}\eta_{ij}^\top$$

almost surely for some i.i.d. standard Gaussian vectors η_{ij} *in* $\mathbb{R}^d$.

The next result verifies Assumptions 6.1 and 6.2 for isotropic Gaussian data.

Lemma D.17. *Suppose* $\mathbf{X}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ *and* $\xi_{ij} \sim \mathcal{N}(\mathbf{0}, \sigma_A^2\mathbf{I}_d)$*, and consider the asymptotic* (6.20) *with* $\gamma' = \lim d/(kn) \neq 1$. *Then Assumptions 6.1 and 6.2 hold.*

D.3 Auxiliary results

In this section, we include a collection of results useful for various parts of our proof.

D.3.1. Convergence in $d_{\mathcal{H}}$

The weak convergence lemma Lemma 6.3 shows that convergence in $d_{\mathcal{H}}$ implies weak convergence. The gist of the proof is as follows. Assuming dimension to be one, in Step 1, we construct a thrice-differentiable function in $\mathcal{H}$ to approximate indicator functions in $\mathbb{R}$. This allows us to bound the difference in probabilities of two random variables X and Y lying in nearby regions by their distance in $d_{\mathcal{H}}$. In Step 2, we consider a sequence of random variables Y_n converging to Y in $d_{\mathcal{H}}$, and use Step 1 to bound the probability of Y_n lying in a given region by the probability of Y lying in a nearby region plus $d_{\mathcal{H}}(Y_n, Y)$, which converges to zero. This allows us to show convergence of the distribution function of Y_n to that of Y. Finally, we make use of Cramer-Wold and Slutsky's Lemma to generalise our result to $q \geq 1$ dimensions.

Proof of Lemma 6.3. Step 1. Assume $q = 1$. Let $A \subset \mathbb{R}$ be a Borel set, and $\epsilon \in (0, 1)$ a constant. We will first show that

$$\mathbb{P}(Y \in A_{8\epsilon}) \geq \mathbb{P}(X \in A) - d_{\mathcal{H}}(X, Y)/\epsilon^4 \,. \tag{D.7}$$

where $A_\epsilon \coloneqq \{x \in \mathbb{R} \mid \exists y \in A \text{ s.t. } |x - y| \leq \epsilon\}$. To this end, define a smoothed approximation of the indicator function of A as

$$h_\epsilon(x) \coloneqq \frac{1}{\epsilon^4}\int_{x-\epsilon}^{x}\int_{s-\epsilon}^{s}\int_{t-\epsilon}^{t}\int_{y-\epsilon}^{y}\mathbb{I}\{z \in A_{4\epsilon}\}\,dz\,dy\,dt\,ds \,.$$

Then h_ϵ is three times differentiable everywhere on $\mathbb{R}$, and its first three derivatives are bounded in absolute value by $1/\epsilon^4$. It follows that $\epsilon^4 h_\epsilon \in \mathcal{H}$, and hence that

$$|\mathbb{E}h_\epsilon(X) - \mathbb{E}h_\epsilon(Y)| \leq d_{\mathcal{H}}(X, Y)/\epsilon^4 \, .$$

Since $h_\epsilon = 0$ outside $A_{8\epsilon}$ and $h_\epsilon = 1$ on A, we have $\mathbb{P}(Z \in A) \leq \mathbb{E}[h_\epsilon(Z)] \leq \mathbb{P}(Z \in A_{8\epsilon})$ for any random variable Z. It follows that

$$\mathbb{E}h_\epsilon(X) - \mathbb{E}h_\epsilon(Y) \geq \mathbb{P}(X \in A) - \mathbb{P}(Y \in A_{8\epsilon}) \, ,$$

which implies (D.7).

Step 2. To establish weak convergence for $q = 1$, denote by F the c.d.f of Y. To show $Y_n \xrightarrow{\text{d}} Y$, it suffices to show that $\mathbb{P}(Y_n \leq b) \to F(b)$ at every point $b \in \mathbb{R}$ at which F is continuous. For any $\epsilon \in (0, 1)$, we have

$$\mathbb{P}(Y \leq b + 8\epsilon) \;\geq\; \mathbb{P}(Y_n \leq b) - d_{\mathcal{H}}(Y_n, Y)/\epsilon^4 \;\geq\; \limsup_n \mathbb{P}(Y_n \leq b) \, ,$$

where the first inequality uses (D.7) and the second $d_{\mathcal{H}}(Y_n, Y) \to 0$. Set $a = b - 8\epsilon$. Then

$$\begin{aligned} \mathbb{P}(Y_n \leq b) \;&= \mathbb{P}(Y_n \leq a + 8\epsilon) \\ &\geq \mathbb{P}(Y \leq a) - d_{\mathcal{H}}(Y_n, Y)/\epsilon^4 \;=\; \mathbb{P}(Y \leq b - 8\epsilon) - d_{\mathcal{H}}(Y_n, Y)/\epsilon^4 \, , \end{aligned}$$

and hence $\liminf_n \mathbb{P}(Y_n \leq b) \geq \mathbb{P}(Y \leq b - 8\epsilon)$. To summarize, we have

$$\mathbb{P}(Y \leq b - 8\epsilon) \;\leq\; \liminf_n \mathbb{P}(Y_n \leq b) \;\leq\; \limsup_n \mathbb{P}(Y_n \leq b) \;\leq\; \mathbb{P}(Y \leq b + 8\epsilon)$$

for any $\epsilon \in (0, 1)$. Since F is continuous at b, we can choose ϵ arbitrary small, which shows $\lim \mathbb{P}(Y_n \leq b) = \mathbb{P}(Y \leq b)$. Thus, weak convergence holds in $\mathbb{R}$.

Step 3. Finally, consider any $q \in \mathbb{N}$. In this case, it is helpful to write $\mathcal{H}(q)$ for the class $\mathcal{H}$ of functions with domain $\mathbb{R}^q$. Recall the Cramer-Wold theorem (Kallenberg, 2001, Corollary 5.5): Weak convergence $Y_n \xrightarrow{\text{d}} Y$ in $\mathbb{R}^q$ holds if, for every vector $v \in \mathbb{R}^q$, the scalar products $v^\top Y_n$ converge weakly to $v^\top Y$. By Slutsky's lemma, it is sufficient to consider only vectors v with $\|v\| = 1$. Now observe that, if $h \in \mathcal{H}(1)$ and $\|v\| = 1$, the function $y \mapsto h(v^\top y)$ is in $\mathcal{H}(q)$, for every $v \in \mathbb{R}^q$. It follows that $d_{\mathcal{H}(q)}(Y_n, Y) \to 0$ implies $d_{\mathcal{H}(1)}(v^\top Y_n, v^\top Y) \to 0$ for every vector v, which by Step 2 implies $v^\top Y_n \xrightarrow{\text{d}} v^\top Y$, and weak convergence in $\mathbb{R}^q$ holds by Cramer-Wold. □

Comparison of $d_{\mathcal{H}}$ with known probability metrics In this section, let X, Y be random variables taking values in $\mathbb{R}$, and define A^ϵ as in the proof of Lemma 6.3. We present a result that helps to build intuitions of $d_{\mathcal{H}}$ by bounding it with known metrics. Specifically, we consider the Lévy–Prokhorov metric d_P and Kantorovich metric d_K, defined

respectively as

$$d_P(X,Y) \;=\; \inf_{\epsilon>0}\{\epsilon \mid \mathbb{P}(X\in A)\le \mathbb{P}(Y\in A_\epsilon)+\epsilon, \\ \mathbb{P}(Y\in A)\le \mathbb{P}(X\in A_\epsilon)+\epsilon \text{ for all Borel set } A\subseteq\mathbb{R}\}\;, \tag{D.8}$$

$$d_K(X,Y) \;=\; \sup\{\mathbb{E}[h(X)]-\mathbb{E}[h(Y)] \mid h:\mathbb{R}\to\mathbb{R} \text{ has Lipschitz constant} \le 1\}\;.$$

The Kantorovich metric is equivalent to the Wasserstein-1 metric when the distributions of X and Y have bounded support. We can compare $d_{\mathcal{H}}$ to d_P and d_K as follows:

Lemma D.18. *$d_P(X,Y)\le 8^{4/5}d_{\mathcal{H}}(X,Y)^{1/5}$ and $d_{\mathcal{H}}(X,Y)\le d_K(X,Y)$.*

Proof. For the first inequality, recall from (D.7) in the proof of Lemma 6.3 that for $\delta>0$ and any Borel set $A\subset\mathbb{R}$, $\mathbb{P}(Y\in A_{8\delta})\ge\mathbb{P}(X\in A)-d_{\mathcal{H}}(X,Y)/\delta^4$. Setting $\delta=\big(d_{\mathcal{H}}(\mathbf{X},\mathbf{Y})/8\big)^{1/5}$ gives

$$\mathbb{P}(X\in A)\;\le\;\mathbb{P}\big(Y\in A_{8^{4/5}d_{\mathcal{H}}(X,Y)^{1/5}}\big)+8^{4/5}d_{\mathcal{H}}(X,Y)^{1/5}\;.$$

By the definition of d_P, this implies that $d_P(X,Y)\le 8^{4/5}d_{\mathcal{H}}(X,Y)^{1/5}$. The second inequality $d_{\mathcal{H}}(X,Y)\le d_K(X,Y)$ directly follows from the fact that every $h\in\mathcal{H}$ has its first derivative uniformly bounded above by 1. □

Remark D.3. The proof for $d_P(X,Y)\le 8^{4/5}d_{\mathcal{H}}(X,Y)^{1/5}$ in Lemma D.19 can be generalised to $\mathbb{R}^q$ so long as q is fixed. Since the inequality says convergence in $d_{\mathcal{H}}$ implies convergence in d_P and d_P metrizes weak convergence, this gives an alternative proof for Lemma 6.3.

Convergence in $d_{\mathcal{H}}$ implies convergence of mean Lemma 6.6 is useful for translating the convergence in $d_{\mathcal{H}}$ of uncentred quantities to centred versions, and we present the proof below.

Proof of Lemma 6.6. The first bound can be proved by noting that each coordinate function that maps an $\mathbb{R}^d$ vector to one of its coordinate in $\mathbb{R}$ belongs to $\mathcal{H}$:

$$\|\mathbb{E}\mathbf{X}-\mathbb{E}\mathbf{Y}\| \;=\; \big(\textstyle\sum_{l=1}^q|\mathbb{E}[X_l]-\mathbb{E}[Y_l]|^2\big)^{1/2} \;\le\; \big(q\,d_{\mathcal{H}}(\mathbf{X},\mathbf{Y})^2\big)^{1/2} \;\le\; q^{1/2}\epsilon\;.$$

To prove the second bound, notice that the class of functions $\mathcal{H}$ is invariant under a constant shift in the argument of the function, which implies $d_{\mathcal{H}}(\mathbf{X}-\mathbb{E}\mathbf{X},\mathbf{Y}-\mathbb{E}\mathbf{X})\le\epsilon$. By the triangle inequality, we have

$$\begin{aligned} d_{\mathcal{H}}(\mathbf{X}-\mathbb{E}\mathbf{X},\mathbf{Y}-\mathbb{E}\mathbf{Y}) &\le \epsilon + d_{\mathcal{H}}(\mathbf{Y}-\mathbb{E}\mathbf{X},\mathbf{Y}-\mathbb{E}\mathbf{Y}) \\ &\le \epsilon+\sup_{h\in\mathcal{H}}\big|\mathbb{E}\big[h(\mathbf{Y}-\mathbb{E}\mathbf{X})-h(\mathbf{Y}-\mathbb{E}\mathbf{Y})\big]\big| \\ &\overset{(a)}{\le} \epsilon+\|\mathbb{E}\mathbf{X}-\mathbb{E}\mathbf{Y}\| \;\le\; (1+q^{1/2})\epsilon\;. \end{aligned}$$

In (a), we have applied the mean value theorem to h on the interval $[\mathbf{Y}-\mathbb{E}\mathbf{X}, \mathbf{Y}-\mathbb{E}\mathbf{Y}]$ and used $\|\partial h\| \leq 1$. This finishes the proof. □

D.3.2. Additional tools

The following lemma establishes identities for comparing different variances obtained in Theorem 6.1 (main result with augmentation), (6.7) (no augmentation) and other variants of the main theorem in Appendix D.1.2.

Lemma D.19. *Consider independent random elements ϕ, ψ of $\mathcal{T}$ and $\mathbf{X}$ of $\mathcal{D} \subseteq \mathbb{R}^d$, where $\phi \overset{\mathrm{d}}{=} \psi$. Then*

(i) $\mathrm{Cov}[\phi\mathbf{X}, \psi\mathbf{X}] = \mathbb{E}\mathrm{Cov}[\phi\mathbf{X}, \psi\mathbf{X}|\phi, \psi] = \mathrm{Var}\mathbb{E}[\phi\mathbf{X}|\mathbf{X}]$,

(ii) $\mathrm{Var}[\phi\mathbf{X}] \succeq \mathbb{E}\mathrm{Var}[\phi\mathbf{X}|\phi] \succeq \mathrm{Cov}[\phi\mathbf{X}, \psi\mathbf{X}]$, *where $\succeq$ denotes Löwner's partial order.*

Proof. (i) By independence of ϕ and ψ, $\mathrm{Cov}\big[\mathbb{E}[\phi\mathbf{X}|\phi], \mathbb{E}[\psi\mathbf{X}|\psi]\big] = \mathbf{0}$. By combining this with the law of total covariance, we obtain that

$$\mathrm{Cov}[\phi\mathbf{X}, \psi\mathbf{X}] = \mathbb{E}[\mathrm{Cov}[\phi\mathbf{X}, \psi\mathbf{X}|\phi, \psi]] + \mathrm{Cov}\big[\mathbb{E}[\phi\mathbf{X}|\phi], \mathbb{E}[\psi\mathbf{X}|\psi]\big] = \mathbb{E}[\mathrm{Cov}[\phi\mathbf{X}, \psi\mathbf{X}|\phi, \psi]].$$

Moreover, independence of ϕ and ψ also gives $\mathrm{Cov}[\phi\mathbf{X}, \psi\mathbf{X}|\mathbf{X}] = 0$ almost surely. Therefore by law of total covariance with conditioning performed on $\mathbf{X}$, we get

$$\mathrm{Cov}[\phi\mathbf{X}, \psi\mathbf{X}] = \mathrm{Cov}\big[\mathbb{E}[\phi\mathbf{X}|\mathbf{X}], \mathbb{E}[\psi\mathbf{X}|\mathbf{X}]\big] \overset{(a)}{=} \mathrm{Var}\mathbb{E}[\phi\mathbf{X}|\mathbf{X}]\ ,$$

where to obtain (a) we used the fact that as $\phi \overset{d}{=} \psi$ we have $\mathbb{E}[\phi\mathbf{X}|\mathbf{X}] \overset{a.s}{=} \mathbb{E}[\psi\mathbf{X}|\mathbf{X}]$.

(ii) By the law of total variance we have:

$$\mathrm{Var}[\phi\mathbf{X}] = \mathbb{E}[\mathrm{Var}[\phi\mathbf{X}|\phi]] + \mathrm{Var}[\mathbb{E}[\phi\mathbf{X}|\phi]]. \tag{D.9}$$

We know that $\mathrm{Var}[\mathbb{E}[\phi\mathbf{X}|\phi]] \succeq 0$ almost surely, which implies that $\mathrm{Var}[\phi\mathbf{X}] \succeq \mathbb{E}[\mathrm{Var}[\phi\mathbf{X}|\phi]]$. For the second inequality, note that for all deterministic vector $\mathbf{v} \in \mathbb{R}^d$ we have

$$\mathbf{v}^\top\big(\mathbb{E}\mathrm{Var}[\phi\mathbf{X}|\phi] - \mathbb{E}\mathrm{Cov}[\phi\mathbf{X}, \psi\mathbf{X}|\phi, \psi]\big)\mathbf{v} \overset{(b)}{=} \mathbb{E}\big[\mathrm{Var}[\mathbf{v}^\top(\phi\mathbf{X})|\phi] - \mathrm{Cov}[\mathbf{v}^\top(\phi\mathbf{X}), \mathbf{v}^\top(\psi\mathbf{X})|\phi, \psi]\big]$$

where (b) is obtained by bilinearity of covariance. By Cauchy-Schwarz, almost surely,

$$\mathrm{Cov}[\mathbf{v}^\top(\phi\mathbf{X}), \mathbf{v}^\top(\psi\mathbf{X})|\phi, \psi] \leq \sqrt{\mathrm{Var}[\mathbf{v}^\top(\phi\mathbf{X})|\phi]}\sqrt{\mathrm{Var}[\mathbf{v}^\top(\psi\mathbf{X})|\psi]}.$$

This implies that

$$\mathbf{v}^\top\big(\mathbb{E}\mathrm{Var}[\phi\mathbf{X}|\phi] - \mathbb{E}\mathrm{Cov}[\phi\mathbf{X}, \psi\mathbf{X}|\phi, \psi]\big)\mathbf{v} \geq 0\ .$$

Therefore we conclude that

$$\mathbb{E}\mathrm{Var}[\phi\mathbf{X}|\phi] \succeq \mathbb{E}\mathrm{Cov}[\phi\mathbf{X}, \psi\mathbf{X}|\phi, \psi] = \mathrm{Cov}[\phi\mathbf{X}, \psi\mathbf{X}],$$

where the last inequality is given by (i). This gives the second inequality as desired. □

The function $\zeta_{i;m}$ defined in the following lemma enters the bound in Theorem 6.1 and its variants through the noise stability terms α_r defined in (6.2) and $\alpha_{r;m}$ defined in Theorem D.1, and will recur throughout the proofs for different examples. We collect useful properties of $\zeta_{i;m}$ into Lemma D.20 for convenience.

Lemma D.20. *For $1 \le i \le n$, let $\Phi_1\mathbf{X}_i$, $\mathbf{Z}_i$ be random quantities in $\mathcal{D}$. For a random function $\mathbf{T} : \mathcal{D}^k \to \mathbb{R}_0^+$, where $\mathbb{R}_0^+$ is the set of non-negative reals, and for $m \in \mathbb{N}$, define*

$$\zeta_{i;m}(\mathbf{T}) \;:=\; \max\left\{ \left\| \sup_{\mathbf{w}\in[\mathbf{0},\Phi_1\mathbf{X}_i]} \mathbf{T}(\mathbf{w}) \right\|_{L_m}, \left\| \sup_{\mathbf{w}\in[\mathbf{0},\mathbf{Z}_i]} \mathbf{T}(\mathbf{w}) \right\|_{L_m} \right\}.$$

Then for any deterministic $\alpha \in \mathbb{R}_0^+$, random functions $\mathbf{T}_j : \mathcal{D}^k \to \mathbb{R}_0^+$, and $s \in \mathbb{N}$,

(i) ***(triangle inequality)*** $\zeta_{i;m}(\mathbf{T}_1 + \mathbf{T}_2) \;\le\; \zeta_{i;m}(\mathbf{T}_1) + \zeta_{i;m}(\mathbf{T}_2)$,

(ii) ***(positive homogeneity)*** $\zeta_{i;m}(\alpha\mathbf{T}_1) \;=\; \alpha\zeta_{i;m}(\mathbf{T}_1)$,

(iii) ***(order preservation)*** *if for all $\mathbf{w} \in \mathbb{R}^{dk}$, $\mathbf{T}_1(\mathbf{w}) \;\le\; \mathbf{T}_2(\mathbf{w})$ almost surely, then $\zeta_{i;m}(\mathbf{T}_1) \;\le\; \zeta_{i;m}(\mathbf{T}_2)$,*

(iv) ***(Hölder's inequality)*** $\zeta_{i;m}(\prod_{j=1}^s \mathbf{T}_j) \;\le\; \prod_{j=1}^s \zeta_{i;ms}(\mathbf{T}_j)$, *and*

(v) ***(coordinate decomposition)*** *if $g : \mathcal{D}^k \to \mathbb{R}^q$ is a r-times differentiable function and $g_s : \mathcal{D}^k \to \mathbb{R}$ denotes the s-th coordinate of g, then $\zeta_{i;m}(\|\partial^r g(\bullet)\|) \le \sum_{s\le q} \zeta_{i;m}(\|\partial^r g_s(\bullet)\|)$.*

Proof. (i), (ii) and (iii) are straightforward by properties of $\sup$ and $\max$ and the triangle inequality. To prove (iv), we note that

$$\left\| \sup_{\mathbf{w}\in[\mathbf{0},\Phi_1\mathbf{X}_i]} \prod_{j=1}^s \mathbf{T}(\mathbf{w}) \right\|_{L_m} \le \left\| \prod_{j=1}^s \sup_{\mathbf{w}\in[\mathbf{0},\Phi_1\mathbf{X}_i]} \mathbf{T}_j(\mathbf{w}) \right\|_{L_m}.$$

By the generalised Hölder's inequality we also have

$$\left\| \prod_{j=1}^s \sup_{\mathbf{w}\in[\mathbf{0},\Phi_1\mathbf{X}_i]} \mathbf{T}_j(\mathbf{w}) \right\|_{L_m} \le \prod_{j=1}^s \left\| \sup_{\mathbf{w}\in[\mathbf{0},\Phi_1\mathbf{X}_i]} \mathbf{T}_j(\mathbf{w}) \right\|_{L_{ms}} \le \prod_{j=1}^s \zeta_{i;ms}(\mathbf{T}_j).$$

Similarly we can prove that $\left\| \sup_{\mathbf{w}\in[\mathbf{0},\mathbf{Z}_i]} \prod_{j=1}^s \mathbf{T}_j(\mathbf{w}) \right\|_{L_m} \le \prod_{j=1}^s \zeta_{i;ms}(\mathbf{T}_j)$. This directly implies that $\zeta_{i;m}(\prod_{j=1}^s \mathbf{T}_j) \le \prod_{j=1}^s \zeta_{i;ms}(\mathbf{T}_j)$. Finally to show (v), note that

$$\|\partial^r g(\mathbf{v})\| \;=\; \sqrt{\textstyle\sum_{s\le q} \|\partial^r g_s(\mathbf{v})\|^2} \;\le\; \textstyle\sum_{s\le q} \|\partial^r g_s(\mathbf{v})\|$$

for every $\mathbf{v} \in \mathcal{D}^k$. By (iii), this implies $\zeta_{i;m}(\|\partial^r g(\bullet)\|) \le \sum_{s\le q} \zeta_{i;m}(\|\partial^r g_s(\bullet)\|)$ as required. □

The following result from Rosenthal (1970) is useful for controlling moment terms, and is used throughout the proofs for different examples. We also prove a corollary that extends the result to vectors since we deal with data in $\mathcal{D} \subseteq \mathbb{R}^d$.

Lemma D.21 (Theorem 3 of Rosenthal (1970)). *Let $2 \le m < \infty$, and $\mathbf{X}_1, \dots, \mathbf{X}_n$ be independent centred random variables in $\mathbb{R}$ admitting a finite m-th moment. Then there exists a constant K_m depending only on m such that*

$$\big\| \textstyle\sum_{i=1}^n \mathbf{X}_i \big\|_{L_m} \;\le\; K_m \max\big\{ \big(\sum_{i=1}^n \|\mathbf{X}_i\|_{L_m}^m \big)^{1/m}, \big(\sum_{i=1}^n \|\mathbf{X}_i\|_{L_2}^2 \big)^{1/2} \big\} \,.$$

Corollary D.22. *Let $2 \le m < \infty$, and $\mathbf{X}_1, \dots, \mathbf{X}_n$ be independent centred random vectors in $\mathbb{R}^d$ such that for all i, $\|\mathbf{X}_i\|$ admits a finite m-th moment. Denote the s-th coordinate of $\mathbf{X}_i$ by X_{is}. Then, there exists a constant K_m depending only on m such that*

$$\Big\| \big\| \textstyle\sum_{i=1}^n \mathbf{X}_i \big\| \Big\|_{L_m} \le K_m \Big(\sum_{s=1}^d \max\big\{ \big(\sum_{i=1}^n \|X_{is}\|_{L_m}^m \big)^{2/m}, \sum_{i=1}^n \|X_{is}\|_{L_2}^2 \big\} \Big)^{1/2} \,.$$

Proof. By the triangle inequality followed by Lemma D.21 applied to $\| \sum_{i=1}^n X_{is} \|_{L_m}$, there exists a constant K_m depending only on m such that

$$\begin{aligned} \Big\| \big\| \textstyle\sum_{i=1}^n \mathbf{X}_i \big\| \Big\|_{L_m} \;&=\; \Big(\Big\| \sum_{s=1}^d \big(\sum_{i=1}^n X_{is} \big)^2 \Big\|_{L_{m/2}} \Big)^{1/2} \\ &\le\; \Big(\sum_{s=1}^d \big\| \big(\sum_{i=1}^n X_{is} \big)^2 \big\|_{L_{m/2}} \Big)^{1/2} \;=\; \Big(\sum_{s=1}^d \big\| \sum_{i=1}^n X_{is} \big\|_{L_m}^2 \Big)^{1/2} \\ &\le\; K_m \Big(\sum_{s=1}^d \max\big\{ \big(\sum_{i=1}^n \|X_{is}\|_{L_m}^m \big)^{2/m}, \sum_{i=1}^n \|X_{is}\|_{L_2}^2 \big\} \Big)^{1/2} \,. \end{aligned} \tag{D.10}$$

□

The following lemma bounds the moments of vector norms of a Gaussian random vector in terms of its first two moments, which is useful throughout the proofs.

Lemma D.23. *Consider a random vector $\mathbf{X}$ in $\mathbb{R}^d$ with bounded mean and variance. Let ξ be a Gaussian vector in $\mathbb{R}^d$ with its mean and variance matching those $\mathbf{X}$, and write $\|\bullet\|_\infty$ as the vector-infinity norm. Then for every integer $m \in \mathbb{N}$,*

$$\big\| \|\xi\|_\infty \big\|_{L_m} \;\le\; C_m \big\| \|\mathbf{X}\|_\infty \big\|_{L_2} \sqrt{1 + \log d} \,.$$

Proof. Denote $\Sigma \coloneqq \mathrm{Var}[\mathbf{X}]$, and write $\xi = \mathbb{E}[\mathbf{X}] + \Sigma^{1/2}\mathbf{Z}$ where $\mathbf{Z}$ is a standard Gaussian vector in $\mathbb{R}^d$. First note that by triangle inequality and Jensen's inequality,

$$\big\| \|\xi\|_\infty \big\|_{L_m} \;\le\; \|\mathbb{E}[\mathbf{X}]\|_\infty + \big\| \|\Sigma^{1/2}\mathbf{Z}\|_\infty \big\|_{L_m} \;\le\; \big\| \|\mathbf{X}\|_\infty \big\|_{L_1} + \big\| \|\Sigma^{1/2}\mathbf{Z}\|_\infty \big\|_{L_m} \,.$$

Write $\sigma_l \coloneqq \sqrt{\Sigma_{l,l}}$, the square root of the (l,l)-th coordinate of Σ. If $\sigma_l = 0$ for some $l \le d$, then the l-th coordinate of $\Sigma^{1/2}\mathbf{Z}$ is zero almost surely and does not play a role in $|\Sigma^{1/2}\mathbf{Z}\|_\infty$. We can then remove the l-th row and column of Σ and consider a lower-dimensional Gaussian vector such that its covariance matrix has strictly positive diagonal entries. If all σ_l's are zero, we get the following bound

$$\big\| \|\xi\|_\infty \big\|_{L_m} \;\le\; \big\| \|\mathbf{X}\|_\infty \big\|_{L_1} \,,$$

which implies that $\| \|\xi\|_\infty \|_{L_m}$ satisfies the statement in the lemma. Therefore WLOG we consider the case where $\sigma_l > 0$ for every $l \le d$. By splitting the integral and applying a union bound, we have that for any $c > 0$,

$$\begin{aligned}
&\| \|\Sigma^{1/2}\mathbf{Z}\|_\infty \|_{L_m}^m \;=\; \mathbb{E}[\max_{l\le d} |(\Sigma^{1/2}\mathbf{Z})_l|^m] \;=\; \int_0^\infty \mathbb{P}\big(\max_{l\le d} |(\Sigma^{1/2}\mathbf{Z})_l| > x^{1/m}\big)\, dx \\
&\le c + d\int_c^\infty \mathbb{P}\big(|(\Sigma^{1/2}\mathbf{Z})_l| > x^{1/m}\big)\, dx \;=\; c + d\int_c^\infty \mathbb{P}\big(\frac{1}{\sigma_l}|(\Sigma^{1/2}\mathbf{Z})_l| > \frac{1}{\sigma_l}x^{1/m}\big)\, dx \\
&\overset{(a)}{\le} c + d\int_c^\infty \frac{1}{\sqrt{2\pi}\frac{1}{\sigma_l}x^{1/m}} \exp\big(-\frac{x^{2/m}}{2\sigma_l^2}\big)\, dx \\
&\le c + \frac{d\sigma_l}{\sqrt{2\pi}c^{1/m}} \int_c^\infty \exp\big(-\frac{x^{2/m}}{2\sigma_l^2}\big)\, dx\,. \qquad \text{(D.11)}
\end{aligned}$$

In (a) we have noted that $\frac{1}{\sigma_l}(\Sigma^{1/2}\mathbf{Z})_l \sim \mathcal{N}(0,1)$, and used the standard lower bound for the c.d.f. of a standard normal random variable Z to obtain

$$\mathbb{P}(|Z| > u) \;=\; 2\mathbb{P}(Z > u) \;\ge\; \frac{1}{\sqrt{2\pi}\,x}\exp\big(-\frac{x^2}{2}\big)\,.$$

Choose $c = (2\sigma_l^2(1+\log d))^{\frac{m}{2}}$. Then by a change of variable, the integral in (D.11) becomes

$$\begin{aligned}
\int_c^\infty \exp\big(-\frac{x^{2/m}}{2\sigma_l^2}\big)\, dx \;&= (2\sigma_l)^{m/2}\int_{1+\log d}^\infty e^{-y}\, y^{\frac{m}{2}-1} dy \\
&\le (2\sigma_l)^{m/2}\int_{1+\log d}^\infty y^{\lfloor\frac{m}{2}\rfloor} e^{-y}\, dy \;=:\; (2\sigma_l)^{m/2} I_{\lfloor\frac{m}{2}\rfloor}\,.
\end{aligned}$$

We have denoted $I_k \coloneqq \int_{1+\log d}^\infty y^k e^{-y} dy$. By integration by parts, we get the following recurrence for $k \ge 1$,

$$I_k = (1+\log d)^k e^{-1-\log d} + k\, I_{k-1} = (1+\log d)^k (ed)^{-1} + k\, I_{k-1}\,,$$

and also $I_0 = (ed)^{-1}$. This implies that there exists some constant A_m depending only on m such that

$$\begin{aligned}
&\int_c^\infty \exp\big(-\frac{x^{2/m}}{2\sigma_l^2}\big)\, dx \;\le\; (2\sigma_l^2)^{m/2} I_{\lfloor\frac{m}{2}\rfloor} \\
&\le (2\sigma_l^2)^{m/2}(ed)^{-1}\lfloor\frac{m}{2}\rfloor! + (2\sigma_l^2)^{m/2}\sum_{k=1}^{\lfloor\frac{m}{2}\rfloor}(ed)^{-(\lfloor\frac{m}{2}\rfloor+1-k)}\frac{\lfloor\frac{m}{2}\rfloor!}{k!}(1+\log d)^k \\
&\le A_m d^{-1}\sigma_l^m (1+\log d)^{\lfloor\frac{m}{2}\rfloor}\,.
\end{aligned}$$

Substituting this and our choice of c into (D.11), while noting that $\sigma_l = \sqrt{\Sigma_{l,l}} \le \|\Sigma\|_\infty^{1/2}$, we get that

$$\begin{aligned}
\| \|\Sigma^{1/2}\mathbf{Z}\|_\infty \|_{L_m}^m \;&\le\; (2\sigma_l^2(1+\log d))^{\frac{m}{2}} + \frac{d\sigma_l}{\sqrt{2\pi}(2\sigma_l^2(1+\log d))^{\frac{1}{2}}} A_m d^{-1}\sigma_l^m (1+\log d)^{\lfloor\frac{m}{2}\rfloor} \\
&\le\; B_m(\|\Sigma\|_\infty(1+\log d))^{m/2}\,,
\end{aligned}$$

for some constant B_m depending only on m. Finally, by the property of a covariance

matrix and Jensen's inequality, we get that

$$\|\Sigma\|_\infty^{1/2} \le \max_{l\le d} \mathrm{Var}[X_l]^{1/2} \le \max_{l\le d} \mathbb{E}[X_l^2]^{1/2} \le \|\mathbb{E}[\mathbf{X}\mathbf{X}^\top]\|_\infty^{1/2} \le \| \|\mathbf{X}\|_\infty \|_{L_2} .$$

These two bounds on $\| \|\Sigma^{1/2}\mathbf{Z}\|_\infty \|_{L_m}^m$ and $\|\Sigma\|_\infty^{1/2}$ imply that, for $C_m := B_m + 1$,

$$\begin{aligned}
\| \|\xi\|_\infty \|_{L_m} &\le \| \|\mathbf{X}\|_\infty \|_{L_1} + \| \|\Sigma^{1/2}\mathbf{Z}\|_\infty \|_{L_m} \\
&\le \| \|\mathbf{X}\|_\infty \|_{L_1} + B_m \| \|\mathbf{X}\|_\infty \|_{L_2} (1+\log d)^{1/2} \\
&\le C_m \| \|\mathbf{X}\|_\infty \|_{L_2} (1+\log d)^{1/2} .
\end{aligned}$$

□

The next result controls the norm of the largest eigenvalue of a sum of i.i.d. zero-mean (not necessarily symmetric) matrices.

Lemma D.24. *Let $(\mathbf{A}_i)_{i\le n}$ be i.i.d. zero-mean random matrices in $\mathbb{R}^{d\times d}$ and $m \ge 1$. There exists some absolute constant $C > 0$ such that*

$$\begin{aligned}
&\left\| \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{A}_i \right\|_{op} \right\|_{L_m} \\
&\quad \le \frac{C\sqrt{m+\log d}}{\sqrt{n}} \left(\left\| \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{A}_i \mathbf{A}_i^\top \right\|_{op}^{1/2} \right\|_{L_m} + \left\| \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{A}_i^\top \mathbf{A}_i \right\|_{op}^{1/2} \right\|_{L_m} \right)
\end{aligned}$$

Proof of Lemma D.24. As $\mathbf{A}_i$'s are not symmetric, we consider the symmetric matrices

$$\mathbf{H}_i := \begin{pmatrix} \mathbf{0} & \mathbf{A}_i \\ \mathbf{A}_i^\top & \mathbf{0} \end{pmatrix} \in \mathbb{R}^{2d\times 2d} ,$$

which satisfies the identities

$$\mathbf{H}_i^2 = \begin{pmatrix} \mathbf{A}_i\mathbf{A}_i^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_i^\top\mathbf{A}_i \end{pmatrix} \qquad \text{and} \qquad \|\mathbf{H}_i\|_{op} = \|\mathbf{A}_i\|_{op} .$$

This allows us to express the quantity of interest in terms of a sum of symmetric matrices

$$\left\| \frac{1}{n} \sum_{i=1}^n \mathbf{A}_i \right\|_{op} = \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{H}_i \right\|_{op} .$$

Let $\varepsilon_1, \dots, \varepsilon_n$ be i.i.d. Rademacher variables. By the symmetrization lemma for random vectors (see e.g. Exercise 6.4.5 of Vershynin (2018)), we have that for $m \ge 1$,

$$\begin{aligned}
\left\| \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{H}_i \right\|_{op} \right\|_{L_m} &\le 2 \left\| \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \mathbf{H}_i \right\|_{op} \right\|_{L_m} \\
&= 2 \left(\mathbb{E}\left[\mathbb{E}\left[\left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \mathbf{H}_i \right\|_{op}^m \,\middle|\, (\mathbf{H}_i)_{i\le n} \right] \right] \right)^{1/m} ,
\end{aligned}$$

and by the matrix Khintchine's inequality (see e.g. Exercise 5.4.13(b) of Vershynin (2018)), there exists some absolute constant $C > 0$ such that almost surely

$$\mathbb{E}\left[\left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \mathbf{H}_i \right\|_{op}^m \,\middle|\, (\mathbf{H}_i)_{i\le n} \right] \le \left(\frac{C}{2} \sqrt{m+\log d} \left\| \frac{1}{n^2} \sum_{i=1}^n \mathbf{H}_i^2 \right\|_{op}^{1/2} \right)^m .$$

Combining the bounds yields

$$\begin{aligned}
&\left\| \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{A}_i \right\|_{op} \right\|_{L_m} \leq C\sqrt{m+\log d} \left\| \left\| \frac{1}{n^2} \sum_{i=1}^n \mathbf{H}_i^2 \right\|_{op}^{1/2} \right\|_{L_m} \\
&= C\sqrt{m+\log d} \left\| \left\| \frac{1}{n^2} \sum_{i=1}^n \begin{pmatrix} \mathbf{A}_i \mathbf{A}_i^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_i^\top \mathbf{A}_i \end{pmatrix} \right\|_{op}^{1/2} \right\|_{L_m} \\
&\leq \frac{C\sqrt{m+\log d}}{\sqrt{n}} \left(\left\| \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{A}_i \mathbf{A}_i^\top \right\|_{op}^{1/2} \right\|_{L_m} + \left\| \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{A}_i^\top \mathbf{A}_i \right\|_{op}^{1/2} \right\|_{L_m} \right) .
\end{aligned}$$

□

D.4 Proof of the main result

In this section, we prove Theorem D.1. Theorem 6.1 then follows as a special case. We begin with an outline of the proof technique.

D.4.1. Proof overview

We quickly recall the Lindeberg method from Chapter 2, and point out the adaptations in the block dependent case. The original Lindeberg method is as follows: The goal is to bound the difference $|\mathbb{E}[g(\xi_1, \ldots, \xi_n)] - \mathbb{E}[g(\zeta_1, \ldots, \zeta_n)]|$, for independent collections $\xi_1, \ldots, \xi_n$ and $\zeta_1, \ldots, \zeta_n$ of i.i.d. variables and a function g. To this end, abbreviate $V_i(\bullet) = (\xi_1, \ldots, \xi_{i-1}, \bullet, \zeta_{i+1}, \ldots, \zeta_n)$, and expand into a telescopic sum:

$$\begin{aligned}
&\mathbb{E}[g(\xi_1, \ldots, \xi_n)] - \mathbb{E}[g(\zeta_1, \ldots, \zeta_n)] = \sum_{i \leq n} \mathbb{E}[g(V_i(\xi_i)) - g(V_i(\zeta_i))] \\
&= \sum_{i \leq n} \big(\mathbb{E}[g(V_i(\xi_i)) - g(V_i(0))] - \mathbb{E}[g(V_i(\zeta_i)) - g(V_i(0))] \big) .
\end{aligned}$$

By Taylor-expanding the function $g_i(\bullet) \coloneqq g(V_i(\bullet))$ to third order around 0, each summand can be represented as

$$\mathbb{E}[\partial g_i(0)(\xi_i - \zeta_i)] + \mathbb{E}[\partial^2 g_i(0)(\xi_i - \zeta_i)^2] + \mathbb{E}[\partial^3 g_i(\tilde{\xi}_i)\xi_i^3 + \partial^3 g_i(\tilde{\zeta}_i)\zeta_i^3] ,$$

for some $\tilde{\xi}_i \in [0, \xi_i]$ and $\tilde{\zeta}_i \in [0, \zeta_i]$. Since each (ξ_i, ζ_i) is independent of all other pairs $\{(\xi_j, \zeta_j)\}_{j \neq i}$, expectations factorize, and the expression above becomes

$$\mathbb{E}[\partial g_i(0)]\mathbb{E}[\xi_i - \zeta_i] + \mathbb{E}[\partial^2 g_i(0)]\mathbb{E}[(\xi_i - \zeta_i)^2] + \mathbb{E}[\partial^3 g_i(\tilde{\xi}_i)\xi_i^3 + \partial^3 g_i(\tilde{\zeta}_i)\zeta_i^3] . \quad \text{(D.12)}$$

The first two terms can then be controlled by matching expectations and variances of ξ_i and ζ_i. To control the third term, one imposes boundedness assumptions on $\partial^3 g_i$ and the moments of ξ_i^3 and ζ_i^3.

Proving our result requires some modifications: Since augmentation induces dependence, the i.i.d. assumption above does not hold. On the other hand, the function g in our

problems is of a more specific form. In broad strokes, our proof proceeds as follows:

- We choose $g := h \circ f$, where h belongs to the class of thrice-differentiable functions with the first three derivatives bounded above by 1. Since the statistic f has (by assumption) three derivatives, so does g.
- We group the augmented data into n independent blocks $\Phi_i \mathbf{X}_i := \{\phi_{i1}\mathbf{X}_i, \ldots, \phi_{ik}\mathbf{X}_i\}$, for $i \leq n$. We can then sidestep dependence by applying the technique above to each block.
- To do so, we to take derivates of $g = h \circ f$ with respect to blocks of variables. The relevant block-wise version of the chain rule is a version of the Faà di Bruno formula. It yields a sum of terms of the form in (D.12).
- The first two terms in (D.12) contribute a term of order k to the bound: The first expectation vanishes by construction. The second also vanishes under the conditions of Theorem 6.1, and more generally if $\delta = 0$. If $\delta > 0$, the matrices $\mathrm{Var}[\mathbf{Z}_i^\delta]$ and $\mathrm{Var}[\Phi_i \mathbf{X}_i]$ may differ in their k diagonal entries.
- The third term in (D.12) contributes a term of order k^3: Here, we use noise stability, which lets us control terms involving $\partial^3 g_i$ on the line segments $[0, \Phi_i \mathbf{X}_i]$ and $[0, \mathbf{Z}_i]$, and moments of $(\Phi_i \mathbf{X}_i)^{\otimes 3}$ and $(\mathbf{Z}_i)^{\otimes 3}$. The moments have dimension of order k^3.
- Summing over n quantities of the form (D.12) then leads to the bound of the form

  $$nk \times (\text{second derivative terms}) + nk^3 \times (\text{third derivative terms}) \quad .$$

  in Theorem D.1. In Theorem 6.1, the first term vanishes.

Whether the bound converges depends on the scaling behavior of f. A helpful example is a scaled average $\sqrt{n}\big(\frac{1}{nk}\sum_{i,j}\phi_{ij}\mathbf{X}_i\big)$. Here, the second and third derivatives are respectively of order $\frac{1}{nk^2}$ and $\frac{1}{n^{3/2}k^3}$ (see Appendix D.6.1 for the exact calculation). The bound then scales as $\frac{1}{k} + \frac{1}{n^{1/2}}$ for $\delta > 0$, and as $\frac{1}{n^{1/2}}$ for $\delta = 0$.

D.4.2. Proof of Theorem D.1

We abbreviate $g := h \circ f$, and note that g is a smooth function from $\mathcal{D}^{nk}$ to $\mathbb{R}$. Recall that we have denoted

$$\mathbf{W}_i^\delta(\bullet) \;:=\; (\Phi_1\mathbf{X}_1, \ldots, \Phi_{i-1}\mathbf{X}_{i-1}, \bullet, \mathbf{Z}_{i+1}^\delta, \ldots, \mathbf{Z}_n^\delta) \,.$$

By a telescoping sum argument and the triangle inequality,

$$\big|\mathbb{E}h(f(\Phi\mathcal{X})) - \mathbb{E}h(f(\mathbf{Z}_1^\delta, \ldots, \mathbf{Z}_n^\delta))\big| \;=\; \big|\mathbb{E}\textstyle\sum_{i=1}^n \big[g(\mathbf{W}_i^\delta(\Phi_i\mathbf{X}_i)) - g(\mathbf{W}_i^\delta(\mathbf{Z}_i^\delta))\big]\big|$$

$$\leq \sum_{i=1}^{n} \left|\mathbb{E}\big[g(\mathbf{W}_i^\delta(\Phi_i\mathbf{X}_i)) - g(\mathbf{W}_i^\delta(\mathbf{Z}_i^\delta))\big]\right| . \tag{D.13}$$

Each summand can be written as a sum of two terms,

$$\big(g(\mathbf{W}_i^\delta(\Phi_i\mathbf{X}_i)) - g(\mathbf{W}_i^\delta(\mathbf{0}))\big) \;-\; \big(g(\mathbf{W}_i^\delta(\mathbf{Z}_i^\delta)) - g(\mathbf{W}_i^\delta(\mathbf{0}))\big) .$$

Since $\mathcal{D}^k$ is convex and contains $\mathbf{0} \in \mathbb{R}^{kd}$, we can expand the first term in a Taylor series in the i^{th} argument of g around $\mathbf{0}$ to third order. Then,

$$\begin{aligned}
&\Big|g(\mathbf{W}_i^\delta(\Phi_i\mathbf{X}_i)) - g(\mathbf{W}_i^\delta(\mathbf{0})) - \big(D_i g(\mathbf{W}_i^\delta(\mathbf{0}))\big)(\Phi_i\mathbf{X}_i) \\
&\qquad\qquad - \frac{1}{2}\big(D_i^2 g(\mathbf{W}_i^\delta(\mathbf{0}))\big)\big((\Phi_i\mathbf{X}_i)(\Phi_i\mathbf{X}_i)^\top\big)\Big| \\
&\leq \; \frac{1}{6}\sup\nolimits_{\mathbf{w}\in[\mathbf{0},\Phi_i\mathbf{X}_i]} \big|D_i^3 g\big(\mathbf{W}_i(\mathbf{w})\big)(\Phi_i\mathbf{X}_i)^{\otimes 3}\big|
\end{aligned}$$

holds almost surely. For the second term, we analogously obtain

$$\begin{aligned}
&\Big|g(\mathbf{W}_i^\delta(\mathbf{Z}_i^\delta)) - g(\mathbf{W}_i^\delta(\mathbf{0})) - \big(D_i g(\mathbf{W}_i^\delta(\mathbf{0}))\big)\mathbf{Z}_i^\delta - \frac{1}{2}\big(D_i^2 g(\mathbf{W}_i^\delta(\mathbf{0}))\big)\big((\mathbf{Z}_i^\delta)(\mathbf{Z}_i^\delta)^\top\big)\Big| \\
&\qquad\qquad \leq \; \frac{1}{6}\sup\nolimits_{\mathbf{w}\in[\mathbf{0},\mathbf{Z}_i^\delta]} \big|D_i^3 g\big(\mathbf{W}_i^\delta(\mathbf{w})\big)(\mathbf{Z}_i^\delta)^{\otimes 3}\big|
\end{aligned}$$

almost surely. Each summand in (D.13) is hence bounded above as

$$\left|\mathbb{E}\big[g(\mathbf{W}_i^\delta(\Phi_i\mathbf{X}_i)) - g(\mathbf{W}_i^\delta(\mathbf{Z}_i^\delta))\big]\right| \;\leq\; |\kappa_{1,i}| + \frac{1}{2}|\kappa_{2,i}| + \frac{1}{6}|\kappa_{3,i}| , \tag{D.14}$$

where

$$\begin{aligned}
\kappa_{1,i} &\coloneqq \mathbb{E}\big[\big(D_i g(\mathbf{W}_i^\delta(\mathbf{0}))\big)\big(\Phi_i\mathbf{X}_i - \mathbf{Z}_i^\delta\big)\big] \\
\kappa_{2,i} &\coloneqq \mathbb{E}\big[\big(D_i^2 g(\mathbf{W}_i^\delta(\mathbf{0}))\big)\big((\Phi_i\mathbf{X}_i)(\Phi_i\mathbf{X}_i)^\top - (\mathbf{Z}_i^\delta)(\mathbf{Z}_i^\delta)^\top\big)\big] \\
\kappa_{3,i} &\coloneqq \mathbb{E}\big[\sup_{\mathbf{w}\in[\mathbf{0},\Phi_i\mathbf{X}_i]} \big|D_i^3 g\big(\mathbf{W}_i^\delta(\mathbf{w})\big)(\Phi_i\mathbf{X}_i)^{\otimes 3}\big| + \sup_{\mathbf{w}\in[\mathbf{0},\mathbf{Z}_i^\delta]} \big|D_i^3 g\big(\mathbf{W}_i^\delta(\mathbf{w})\big)(\mathbf{Z}_i^\delta)^{\otimes 3}\big|\big] .
\end{aligned}$$

Substituting into (D.13) and applying the triangle inequality shows

$$\left|\mathbb{E}h(f(\Phi\mathcal{X})) - \mathbb{E}h(f(\mathbf{Z}_1^\delta,\ldots,\mathbf{Z}_n^\delta))\right| \;\leq\; \sum_{i=1}^{n}\big(|\kappa_{1,i}| + \frac{1}{2}|\kappa_{2,i}| + \frac{1}{6}|\kappa_{3,i}|\big) .$$

The next step is to obtain more specific upper bounds for the $\kappa_{r,i}$. To this end, first consider $\kappa_{1,i}$. Since $(\Phi_i\mathbf{X}_i, \mathbf{Z}_i^\delta)$ is independent of $(\Phi_j\mathbf{X}_j, \mathbf{Z}_j^\delta)_{j\neq i}$, we can factorize the expectation, and obtain

$$\kappa_{1,i} \;=\; \mathbb{E}\big[D_i g(\mathbf{W}_i(\mathbf{0}))\big]\big(\mathbb{E}[\Phi_i\mathbf{X}_i] - \mathbb{E}[\mathbf{Z}_i^\delta]\big) \;=\; 0 ,$$

where the second identity holds since $\mathbb{E}\mathbf{Z}_i^\delta = \mathbf{1}_{k\times 1} \otimes \mathbb{E}[\phi_{11}\mathbf{X}_1] = \mathbb{E}[\Phi_i\mathbf{X}_i]$. Factorizing the expectation in $\kappa_{2,i}$ shows

$$\begin{aligned}
\kappa_{2,i} \;&=\; \mathbb{E}\big[D_i^2 g(\mathbf{W}_i(\mathbf{0}))\big]\big(\mathbb{E}\big[(\Phi_i\mathbf{X}_i)(\Phi_i\mathbf{X}_i)^\top\big] - \mathbb{E}\big[(\mathbf{Z}_i^\delta)(\mathbf{Z}_i^\delta)^\top\big]\big) \\
&\leq\; \big\|\mathbb{E}\big[D_i^2 g(\mathbf{W}_i(\mathbf{0}))\big]\big\|\big\|\mathbb{E}\big[(\Phi_i\mathbf{X}_i)(\Phi_i\mathbf{X}_i)^\top\big] - \mathbb{E}\big[(\mathbf{Z}_i^\delta)(\mathbf{Z}_i^\delta)^\top\big]\big\| \\
&\overset{(a)}{=}\; \big\|\mathbb{E}\big[D_i^2 g(\mathbf{W}_i(\mathbf{0}))\big]\big\|\big\|\mathrm{Var}[\Phi_i\mathbf{X}_i] - \mathrm{Var}[\mathbf{Z}_i^\delta]\big\| .
\end{aligned}$$

where to obtain (a) we exploited once again the fact that $\mathbb{E}\mathbf{Z}_i^\delta = \mathbb{E}[\Phi_i \mathbf{X}_i]$. Consider the final norm. Since the covariance matrix of $\Phi_i \mathbf{X}_i$ is

$$\mathrm{Var}[\Phi_i \mathbf{X}_i] \;=\; \mathbf{I}_k \otimes \mathrm{Var}[\phi_{11}\mathbf{X}_1] + (\mathbf{1}_{k\times k} - \mathbf{I}_k) \otimes \mathrm{Cov}[\phi_{11}\mathbf{X}_1, \phi_{12}\mathbf{X}_1] \;,$$

the argument of the norm is

$$\mathrm{Var}[\Phi_i \mathbf{X}_i] - \mathrm{Var}[\mathbf{Z}_i^\delta] \;=\; \delta \mathbf{I}_k \otimes \big(\mathrm{Var}[\phi_{11}\mathbf{X}_1] - \mathrm{Cov}[\phi_{11}\mathbf{X}_1, \phi_{12}\mathbf{X}_1]\big) \;.$$

Lemma D.19 shows $\mathrm{Cov}[\phi_{11}\mathbf{X}_1, \phi_{12}\mathbf{X}_1] \;=\; \mathrm{Var}\mathbb{E}[\phi_{11}\mathbf{X}_1|\mathbf{X}_1]$. It follows that

$$\big\|\mathrm{Var}[\Phi_i \mathbf{X}_i] - \mathrm{Var}[\mathbf{Z}_i^\delta]\big\| \;=\; \delta \big\|\mathbf{I}_k \otimes \mathbb{E}\mathrm{Var}[\phi_{11}\mathbf{X}_1|\mathbf{X}_1]\big\| \;=\; 2\delta k^{1/2} c_1 \;,$$

and hence $\frac{1}{2}|\kappa_{2,i}| \;\leq\; \big\|\mathbb{E}\big[D_i^2 g(\mathbf{W}_i^\delta(\mathbf{0}))\big]\big\| \delta k^{1/2} c_1$. By applying Cauchy-Scwharz inequality and Hölder's inequality, the term $\kappa_{3,i}$ is upper-bounded by

$$\begin{aligned}
\kappa_{3,i} \;&\leq\; \big\|\|\Phi_i \mathbf{X}_i\|^3\big\|_{L_2} \big\|\sup_{\mathbf{w}\in[\mathbf{0},\Phi_i\mathbf{X}_i]} \|D_i^3 g(\mathbf{W}_i^\delta(\mathbf{w}))\|\big\|_{L_2} \\
&\quad+\; \big\|\|\mathbf{Z}_i^\delta\|^3\big\|_{L_2} \big\|\sup_{\mathbf{w}\in[\mathbf{0},\mathbf{Z}_i^\delta]} \|D_i^3 g(\mathbf{W}_i^\delta(\mathbf{w}))\|\big\|_{L_2} \;.
\end{aligned}$$

Since the function $x \mapsto x^3$ is convex on $\mathbb{R}_+$, we can apply Jensen's inequality to obtain

$$\begin{aligned}
\big\|\|\Phi_i \mathbf{X}_i\|^3\big\|_{L_2} \;&=\; \sqrt{\mathbb{E}[\|\Phi_i \mathbf{X}_i\|^6]} \\
&=\; \sqrt{\mathbb{E}\Big[\big(\textstyle\sum_{j=1}^k \|\phi_{ij}\mathbf{X}_i\|^2\big)^3\Big]} \;=\; k^{3/2}\sqrt{\mathbb{E}\Big[\big(\tfrac{1}{k}\textstyle\sum_{j=1}^k \|\phi_{ij}\mathbf{X}_i\|^2\big)^3\Big]} \\
&\leq\; k^{3/2}\sqrt{\mathbb{E}\Big[\tfrac{1}{k}\textstyle\sum_{j=1}^k \|\phi_{ij}\mathbf{X}_i\|^6\Big]} \overset{(a)}{=} k^{3/2}\sqrt{\mathbb{E}\|\phi_{11}\mathbf{X}_1\|^6} = 6k^{3/2} c_X \;,
\end{aligned}$$

where (a) is by noting that for all $i \leq n, j \leq k$, $\phi_{ij}\mathbf{X}_i$ is identically distributed as $\phi_{11}\mathbf{X}_1$. On the other hand, by noting that $\mathbf{Z}_i^\delta$ is identically distributed as $\mathbf{Z}_1^\delta$,

$$\big\|\|\mathbf{Z}_i^\delta\|^3\big\|_{L_2} \;=\; \sqrt{\mathbb{E}[\|\mathbf{Z}_1^\delta\|^6]} \;=\; k^{3/2}\sqrt{\mathbb{E}\Big[\Big(\frac{|Z_{111}^\delta|^2 + \ldots + |Z_{1kd}^\delta|^2}{k}\Big)^3\Big]} \;=\; 6k^{3/2} c_{Z^\delta} \;.$$

We can now abbreviate

$$M_i \coloneqq \max\big\{\big\|\sup_{\mathbf{w}\in[\mathbf{0},\Phi_i\mathbf{X}_i]} \|D_i^3 g\big(\mathbf{W}_i^\delta(\mathbf{w})\big)\|\big\|_{L_2}, \big\|\sup_{\mathbf{w}\in[\mathbf{0},\mathbf{Z}_i]} \|D_i^3 g\big(\mathbf{W}_i^\delta(\mathbf{w})\big)\|\big\|_{L_2}\big\} \;,$$

and obtain $\frac{1}{6}|\kappa_{3,i}| \;\leq\; k^{3/2}(c_X + c_{Z^\delta}) M_i$. In summary, the right-hand side of (D.13) is hence upper-bounded by

$$\begin{aligned}
\text{(D.13)} \;&\leq\; \delta k^{1/2} c_1 \Big(\textstyle\sum_{i=1}^n \big\|\mathbb{E}\big[D_i^2 g(\mathbf{W}_i^\delta(\mathbf{0}))\big]\big\|\Big) \;+\; k^{3/2}(c_X + c_Z)\Big(\textstyle\sum_{i=1}^n M_i\Big) \\
&\leq\; \delta n k^{1/2} c_1 \max_{i\leq n} \big\|\mathbb{E}\big[D_i^2 g(\mathbf{W}_i^\delta(\mathbf{0}))\big]\big\| \;+\; n k^{3/2}(c_X + c_Z) \max_{i\leq n} M_i \;.
\end{aligned}$$

Lemma D.27 below shows that the two maxima are in turn bounded by

$$\max_{i\leq n} \big\|\mathbb{E}\big[D_i^2 g(\mathbf{W}_i^\delta(\mathbf{0}))\big]\big\| \;\leq\; \gamma_2(h)\alpha_{1;2}(f)^2 + \gamma_1(h)\alpha_{2;1}(f) \;=\; \lambda_1(n,k), \quad \text{(D.15)}$$

$$\max_{i\leq n} M_i \;\leq\; \gamma_3(h)\alpha_{1;6}(f)^3 + 3\gamma_2(h)\alpha_{1;4}(f)\alpha_{2;4}(f) + \gamma_1(h)\alpha_{3;2}(f) = \lambda_2(n,k). \quad \text{(D.16)}$$

That yields the desired upper bound on (D.13),

$$\left|\mathbb{E}h(f(\Phi\mathcal{X})) - \mathbb{E}h(f(\mathbf{Z}_1^\delta, \dots, \mathbf{Z}_n^\delta))\right| \le \delta n k^{1/2}\lambda_1(n,k)c_1 + nk^{3/2}\lambda_2(n,k)(c_X + c_Z)\ ,$$

which finishes the proof.

Remark D.4. We remark that both Theorem 6.1 and Theorem D.1 can be generalised directly to independent but not identically distributed vectors $\mathbf{X}_1, \dots, \mathbf{X}_n$. , and that the suprema in the derivative terms can be removed by using a Taylor expansion with integral remainders instead. The resultant bound is the following: For some absolute constant $C > 0$, we have

$$\begin{aligned}&\left|\mathbb{E}h(f(\Phi\mathcal{X})) - \mathbb{E}h(f(\mathbf{Z}_1^\delta, \dots, \mathbf{Z}_n^\delta))\right| \\ &\le \sum_{i=1}^n \delta k^{1/2}\tilde{\chi}_1(n,k)\frac{\|\mathbb{E}\mathrm{Var}[\phi_{i1}\mathbf{X}_1|\mathbf{X}_1]\|}{2} + \sum_{i=1}^n Ck^{3/2}\tilde{\chi}_2(n,k)\frac{\sqrt{\mathbb{E}\|\phi_{i1}\mathbf{X}_i\|^6} + \sqrt{\mathbb{E}\|\mathbf{Z}_i\|^6}}{6}\ ,\end{aligned}$$

where

$$\begin{aligned}\tilde{\chi}_1(n,k) &:= \gamma_2(h)\tilde{\theta}_{1;2}(f)^2 + \gamma_1(h)\tilde{\theta}_{2;1}(f)\ , \\ \tilde{\chi}_2(n,k) &:= \gamma_3(h)\tilde{\theta}_{1;6}(f)^3 + 3\gamma_2(h)\tilde{\theta}_{1;4}(f)\tilde{\alpha}_{2;4}(f) + \gamma_1(h)\tilde{\theta}_{3;2}(f)\ , \\ \theta_{r;m}(f) &:= \sum_{s\le q}\max_{i\le n}\max\left\{\left\|\|D_i^r f_s(\mathbf{W}_i^\delta(\Theta\Phi_i\mathbf{X}_i))\|\right\|_{L_m}, \left\|\|D_i^r f_s(\mathbf{W}_i^\delta(\Theta\mathbf{Z}_i^\delta))\|\right\|_{L_m}\right\},\end{aligned}$$

where $\Theta \sim \mathrm{Uniform}[0,1]$ is independent of all other random variables and plays the role of the variable to be integrated against in the integral remainders.

D.4.3. The remaining bounds

It remains to establish the bounds in (D.15) and (D.16). To this end, we use a vector-valued version of the generalised chain rule, also known as the Faà di Bruno formula. Here is a form that is convenient for our purposes:

Lemma D.25. [Adapted from Theorem 2.1 of Constantine and Savits (1996)] *Consider functions $f \in \mathcal{F}_3(\mathcal{D}^{nk}, \mathbb{R}^q)$ and $h \in \mathcal{F}_3(\mathbb{R}^q, \mathbb{R})$, and write $g := h \circ f$. Then*

$$\begin{aligned}D_i^2 g(\mathbf{u}) &= \partial^2 h\big(f(\mathbf{u})\big)\big(D_i f(\mathbf{u})\big)^{\otimes 2} + \partial h\big(f(\mathbf{u})\big)D_i^2 f(\mathbf{u}), \\ D_i^3 g(\mathbf{u}) &= \partial^3 h\big(f(\mathbf{u})\big)\big(D_i f(\mathbf{u})\big)^{\otimes 3} + 3\partial^2 h\big(f(\mathbf{u})\big)\big(D_i f(\mathbf{u}) \otimes D_i^2 f(\mathbf{u})\big) \\ &\quad + \partial h\big(f(\mathbf{u})\big)D_i^3 f(\mathbf{u})\end{aligned}$$

for any $\mathbf{u} \in \mathcal{D}^{nk}$.

We also need the following result for bounding quantities that involve $\zeta_{i;m}$ in terms of noise stability terms $\alpha_{r;m}$ defined in Theorem D.1.

Lemma D.26. $\max_{i\le n}\zeta_{i;m}(\|D_i^r f(\mathbf{W}_i^\delta(\bullet))\|) \le \alpha_{r;m}(f)$.

Proof. Note that almost surely

$$\|D_i^r f(\mathbf{W}_i^\delta(\bullet))\| \;=\; \sqrt{\textstyle\sum_{s\le q}\|D_i^r f(\mathbf{W}_i^\delta(\bullet))\|^2} \;\le\; \textstyle\sum_{s\le q}\|D_i^r f_s(\mathbf{W}_i^\delta(\bullet))\|\,.$$

Therefore, by triangle inequality of $\zeta_{i;m}$ from Lemma D.20,

$$\begin{aligned}\alpha_{r;m}(f) &\coloneqq \sum_{s\le q}\max_{i\le n}\max\Big\{\Big\|\sup_{\mathbf{w}\in[\mathbf{0},\Phi_i\mathbf{X}_i]}\|D_i^r f_s(\mathbf{W}_i^\delta(\mathbf{w}))\|\Big\|_{L_m},\\ &\qquad\qquad \Big\|\sup_{\mathbf{w}\in[\mathbf{0},\mathbf{Z}_i^\delta]}\|D_i^r f_s(\mathbf{W}_i^\delta(\mathbf{w}))\|\Big\|_{L_m}\Big\}\\ &= \textstyle\sum_{s\le q}\max_{i\le n}\zeta_{i;m}(\|D_i^r f_s(\mathbf{W}_i^\delta(\bullet))\|)\\ &\ge \max_{i\le n}\zeta_{i;m}(\|D_i^r f(\mathbf{W}_i^\delta(\bullet))\|)\,,\end{aligned}$$

which gives the desired bound. □

We are now ready to complete the proof for Theorem 6.1 by proving (D.15) and (D.16).

Lemma D.27. *The bounds* (D.15) *and* (D.16) *hold.*

Proof. For a random function $\mathbf{T} : \mathcal{D}^k \to \mathbb{R}$, define $\zeta_{i;m}(\mathbf{T})$ as in Lemma D.20 with respect to $\Phi_1\mathbf{X}_i$ and $\mathbf{Z}_i^\delta$ from Theorem D.1,

$$\zeta_{i;m}(\mathbf{T}) \coloneqq \max\Big\{\big\|\sup\nolimits_{\mathbf{w}\in[\mathbf{0},\Phi_1\mathbf{X}_i]}\mathbf{T}(\mathbf{w})\big\|_{L_m}, \big\|\sup\nolimits_{\mathbf{w}\in[\mathbf{0},\mathbf{Z}_i^\delta]}\mathbf{T}(\mathbf{w})\big\|_{L_m}\Big\}.$$

We first consider (D.15). By Lemma D.25, almost surely,

$$D_i^2 g(\mathbf{W}_i^\delta(\mathbf{0})) \;= \partial^2 h\big(f(\mathbf{W}_i^\delta(\mathbf{0}))\big)\big(D_i f(\mathbf{W}_i^\delta(\mathbf{0}))\big)^{\otimes 2} + \partial h\big(f(\mathbf{W}_i^\delta(\mathbf{0}))\big) D_i^2 f(\mathbf{W}_i^\delta(\mathbf{0}))\,.$$

By Jensen's inequality to move $\|\bullet\|$ inside the expectation and Cauchy-Schwarz,

$$\begin{aligned}&\big\|\mathbb{E}\big[D_i^2 g(\mathbf{W}_i^\delta(\mathbf{0}))\big]\big\| \;\le\; \zeta_{i;1}\big(\big\|D_i^2 g(\mathbf{W}_i^\delta(\bullet))\big\|\big)\\ &\le \zeta_{i;1}\big(\big\|\partial^2 h\big(f(\mathbf{W}_i^\delta(\bullet))\big)\big\|\big\|D_i f(\mathbf{W}_i^\delta(\bullet))\big\|^2 + \big\|\partial h\big(f(\mathbf{W}_i^\delta(\bullet))\big)\big\|\big\|D_i^2 f(\mathbf{W}_i^\delta(\bullet))\big\|\big)\\ &\le \zeta_{i;1}\big(\gamma_2(h)\|D_i f(\mathbf{W}_i^\delta(\bullet))\|^2 + \gamma_1(h)\|D_i^2 f(\mathbf{W}_i^\delta(\bullet))\|\big)\\ &\overset{(a)}{\le} \gamma_2(h)\;\zeta_2(\|D_i f(\mathbf{W}_i^\delta(\bullet))\|)^2 + \gamma_1(h)\;\zeta_{i;1}(\|D_i^2 f(\mathbf{W}_i^\delta(\bullet))\|)\\ &\overset{(b)}{\le} \gamma_2(h)\,\alpha_{1;2}(f)^2 + \gamma_1(h)\,\alpha_{2;1}(f) \;=\; \lambda_1(n,k)\,,\end{aligned}$$

where (a) is by Hölder's inequality in Lemma D.20 and (b) is by Lemma D.26. Since $\lambda_1(n,k)$ is independent of i, we obtain (D.15) as desired:

$$\max\nolimits_{1\le i\le n}\big\|\mathbb{E}\big[D_i^2 g(\mathbf{W}_i^\delta(\mathbf{0}))\big]\big\| \;\le\; \lambda_1(n,k)\,.$$

We now want to establish that (D.16) holds. Using Lemma D.25 and the triangle

inequality, we obtain that $\|D_i^3 g(\mathbf{W}_i^\delta(\mathbf{w}))\| \leq \mathbf{T}_{1,i}(\mathbf{w}) + \mathbf{T}_{2,i}(\mathbf{w}) + \mathbf{T}_{3,i}(\mathbf{w})$, where

$$\begin{aligned}
\mathbf{T}_{1,i}(\mathbf{w}) &= \|\partial^3 h\big(f(\mathbf{W}_i^\delta(\mathbf{w}))\big)\|\|D_i f(\mathbf{W}_i^\delta(\mathbf{w}))\|^3 \leq \gamma_3(h)\|D_i f(\mathbf{W}_i^\delta(\mathbf{w}))\|^3 \,, \\
\mathbf{T}_{2,i}(\mathbf{w}) &\leq 3\gamma_2(h)\|D_i f(\mathbf{W}_i^\delta(\mathbf{w}))\|\|D_i^2 f(\mathbf{W}_i^\delta(\mathbf{w}))\| \,, \\
\mathbf{T}_{3,i}(\mathbf{w}) &\leq \gamma_1(h)\|D_i^3 f(\mathbf{W}_i^\delta(\mathbf{w}))\| \,.
\end{aligned}$$

Then, by triangle inequality of $\zeta_{i;2}$ from Lemma D.20 (i),

$$M_i = \zeta_{i;2}\big(\big\|D_i^3 g\big(\mathbf{W}_i^\delta(\bullet)\big)\big\|\big) \leq \zeta_{i;2}(\mathbf{T}_{1,i}) + \zeta_{i;2}(\mathbf{T}_{2,i}) + \zeta_{i;2}(\mathbf{T}_{3,i}) \,.$$

Hölder's inequality of ζ_m from Lemma D.20 allows each term to be further bounded as below:

$$\begin{aligned}
\zeta_{i;2}(\mathbf{T}_{1,i}) &\leq \gamma_3(h)\zeta_{i;2}\big(\|D_i f(\mathbf{W}_i^\delta(\bullet))\|^3\big) \leq \gamma_3(h)\zeta_{i;6}\big(\|D_i f(\mathbf{W}_i^\delta(\bullet))\|\big)^3 \\
&\leq \gamma_3(h)\alpha_{1;6}(f)^3 \,, \\
\zeta_{i;2}(\mathbf{T}_{2,i}) &\leq 3\gamma^2(h)\ \zeta_{i;2}\big(\|D_i f(\mathbf{W}_i^\delta(\bullet))\|\|D_i^2 f(\mathbf{W}_i^\delta(\bullet))\|\big) \\
&\leq 3\gamma^2(h)\ \zeta_{i;4}\big(\|D_i f(\mathbf{W}_i^\delta(\bullet))\|\big)\zeta_{i;4}\big(\|D_i^2 f(\mathbf{W}_i^\delta(\bullet))\|\big) \\
&\leq 3\gamma_2(h)\alpha_{1;4}(f)\alpha_{2;4}(f) \,, \\
\zeta_{i;2}(\mathbf{T}_{3,i}) &\leq \gamma_1(h)\ \zeta_{i;2}\big(\|D_i^3 f(\mathbf{W}_i^\delta(\bullet))\|\big) \leq \gamma_1(h)\alpha_{3;2}(f) \,.
\end{aligned}$$

We have again applied Lemma D.26 in each of the final inequalities above. Note that all bounds are again independent of i. Summing the bounds and taking a maximum recovers (D.16):

$$\max_{i\leq n} M_i \leq \gamma_3(h)\,\alpha_{1;6}(f)^3 + 3\gamma_2(h)\alpha_{1;4}(f)\alpha_{2;4}(f) + \gamma_1(h)\alpha_{3;2}(f) = \lambda_2(n,k) \,.$$

□

D.5 Proofs for Appendix D.1

D.5.1. Proofs for Appendix D.1.1

The proof for Theorem D.1 has been discussed in Appendix D.4. In this section we present the proof for Lemma D.2 and Lemma D.3, which shows how Theorem D.1 can be used to obtain bounds on convergence of variance and convergence in $d_{\mathcal{H}}$. They are generalisations of Corollary 6.2 and Corollary 6.4 in the main text.

The main idea in proving Lemma D.2 is to apply the bound on functions of the form $h \circ f$ from Theorem D.1 with h set to identity and f set to an individual coordinate of f and a product of two individual coordinates of f, both scaled up by $\sqrt{n}$.

Proof of Lemma D.2. Choose $h(y) \coloneqq y$ for $y \in \mathbb{R}$ and define

$$f_{rs}(\mathbf{x}_{11:nk}) \;\coloneqq\; f_r(\mathbf{x}_{11:nk}) f_s(\mathbf{x}_{11:nk}) \;, \qquad\qquad \mathbf{x}_{11:nk} \in \mathcal{D}^{nk} \;.$$

Let $[\,\bullet\,]_{r,s}$ denote the (r,s)-th coordinate of a matrix. The difference between $f(\Phi\mathcal{X})$ and $f(\mathcal{Z}^\delta)$ at each coordinate of their covariance matrices can be written in terms of quantities involving $h \circ f_{r_s}$ and $h \circ f_r$:

$$\begin{aligned}
&(\mathrm{Var}[f(\Phi\mathcal{X})])_{r,s} - (\mathrm{Var}[f(\mathcal{Z}^\delta)])_{r,s} \\
&= \mathrm{Cov}[f_r(\Phi\mathcal{X}), f_s(\Phi\mathcal{X})] - \mathrm{Cov}[f_r(\mathcal{Z}^\delta), f_s(\mathcal{Z}^\delta)] \\
&= \mathbb{E}[h(f_{rs}(\Phi\mathcal{X})) - h(f_{rs}(\mathcal{Z}^\delta))] \qquad\qquad (D.17) \\
&\quad - \big(\mathbb{E}[h(f_r(\Phi\mathcal{X}))]\mathbb{E}[h(f_s(\Phi\mathcal{X}))] - \mathbb{E}[h(f_r(\mathcal{Z}^\delta))]\mathbb{E}[h(f_s(\mathcal{Z}^\delta))]\big) \\
&\overset{(a)}{\leq} \big|\mathbb{E}[h(f_{rs}(\Phi\mathcal{X})) - h(f_{rs}(\mathcal{Z}^\delta))]\big| + \big|\mathbb{E}[h(f_r(\Phi\mathcal{X})) - h(f_r(\mathcal{Z}^\delta))]\big|\big|\mathbb{E}[h(f_s(\Phi\mathcal{X}))]\big| \\
&\quad + \big|\mathbb{E}[h(f_r(\mathcal{Z}^\delta))]\big|\big|\mathbb{E}[h(f_s(\Phi\mathcal{X})) - h(f_s(\mathcal{Z}^\delta))]\big| \\
&\overset{(b)}{\leq} T(f_{rs}) + T(f_r)\alpha_{0;1}(f_s) + T(f_s)\alpha_{0;1}(f_r) \;, \qquad\qquad (D.18)
\end{aligned}$$

In (a), we have added and subtracted $\mathbb{E}[h(f_r(\mathcal{Z}))]\mathbb{E}[h(f_s(\Phi\mathcal{X}))]$ from the second difference before applying Cauchy-Schwarz inequality. In (b), we have used the noise stability term $\alpha_{r;m}$ defined in Theorem D.1 and defined the quantity $T(f^*) \coloneqq \mathbb{E}[h(f^*(\Phi\mathcal{X})) - h(f^*(\mathcal{Z}))]$.

We now proceed to bound $T(f)$ using Theorem D.1. First note that $\gamma_1(h) = |\partial h(0)| = 1$ and $\gamma_2(h) = \gamma_3(h) = 0$. To bound $T(f^*)$ for a given $f^* : \mathbb{R} \to \mathbb{R}$, making the dependence on f^* explicit, the mixed smoothness terms in Theorem D.1 is given by

$$\lambda_1(n,k;f^*) \;=\; \alpha_{2;1}(f^*) \;, \qquad\qquad \lambda_2(n,k;f^*) \;=\; \alpha_{3;4}(f^*) \;,$$

and therefore Theorem D.1 implies

$$T(f^*) \;\leq\; \delta n k^{1/2} \alpha_{2;1}(f^*) c_1 + n k^{3/2} \alpha_{3;2}(f^*)(c_X + c_{Z^\delta}) \;. \qquad (D.19)$$

Applying (D.19) to f_r and f_s allows the last two terms in (D.18) to be bounded as:

$$\begin{aligned}
&T(f_r)\alpha_{0;1}(f_s) + T(f_s)\alpha_{0;1}(f_r) \\
&\qquad \leq \delta n k^{1/2}\big(\alpha_{2;1}(f_r)\alpha_{0;1}(f_s) + \alpha_{2;1}(f_s)\alpha_{0;1}(f_r)\big)c_1 \\
&\qquad\quad + n k^{3/2}\big(\alpha_{3;2}(f_r)\alpha_{0;1}(f_s) + \alpha_{3;2}(f_s)\alpha_{0;1}(f_r)\big)(c_X + c_{Z^\delta}) \;.
\end{aligned} \qquad (D.20)$$

To apply (D.19) to $T(f_{rs})$, we need to compute bounds on the partial derivatives of f_{rs}:

$$\begin{aligned}
\|D_i f_{rs}(\mathbf{x}_{11:nk})\| &\leq |f_r(\mathbf{x}_{11:nk})|\,\|\partial f_s(\mathbf{x}_{11:nk})\| + \|\partial f_r(\mathbf{x}_{11:nk})\|\,|f_s(\mathbf{x}_{11:nk})| \;, \\
\|D_i^2 f_{rs}(\mathbf{x}_{11:nk})\| &\leq |f_r(\mathbf{x}_{11:nk})|\,\|\partial^2 f_s(\mathbf{x}_{11:nk})\| + 2\|\partial f_r(\mathbf{x}_{11:nk})\|\,\|\partial f_s(\mathbf{x}_{11:nk})\| \\
&\quad + \|\partial^2 f_r(\mathbf{x}_{11:nk})\|\,|f_s(\mathbf{x}_{11:nk})| \;,
\end{aligned}$$

$$\begin{aligned}\|D_i^3 f_{rs}(\mathbf{x}_{11:nk})\| \;&\le |f_r(\mathbf{x}_{11:nk})|\,\|\partial^3 f_s(\mathbf{x}_{11:nk})\| + 3\|\partial f_r(\mathbf{x}_{11:nk})\|\,\|\partial^2 f_s(\mathbf{x}_{11:nk})\| \\ &\quad + 3\|\partial^2 f_r(\mathbf{x}_{11:nk})\|\,\|\partial f_s(\mathbf{x}_{11:nk})\| + \|\partial^3 f_r(\mathbf{x}_{11:nk})\|\,|f_s(\mathbf{x}_{11:nk})|\;.\end{aligned}$$

Since f_{rs} and f_r both output variables in 1 dimension, recall from Lemma D.26 that noise stability terms can be rewritten in terms of $\zeta_{i;m}$ in Lemma D.20:

$$\alpha_{R;m}(f_{rs}) = \max_{i\le n}\zeta_{i;m}(\|D_i^R f_{rs}(\mathbf{W}_i(\bullet))\|)\;,\;\; \alpha_{R;m}(f_r) = \max_{i\le n}\zeta_{i;m}(\|D_i^R f_r(\mathbf{W}_i(\bullet))\|)\;.$$

By triangle inequality, positive homogeneity and Hölder's inequality of ζ_m from Lemma D.20, we get

$$\begin{aligned}\alpha_{2;1}(f_{rs}) \;&=\; \max_{i\le n}\zeta_{i;2}(\|D_i^2 f_{rs}(\mathbf{W}_i(\bullet))\|) \\ &\le \max_{i\le n}\Big(\zeta_{i;4}(|f_r(\mathbf{W}_i(\bullet))|)\,\zeta_{i;4}(\|\partial^2 f_s(\mathbf{W}_i(\bullet))\|) \\ &\qquad + 2\zeta_{i;4}(\|\partial f_r(\mathbf{W}_i(\bullet))\|)\,\zeta_{i;4}(\|\partial f_s(\mathbf{W}_i(\bullet))\|) \\ &\qquad + \zeta_{i;4}(\|\partial^2 f_r(\mathbf{W}_i(\bullet))\|)\,\zeta_{i;4}(|f_s(\mathbf{W}_i(\bullet))|)\Big) \\ &\le \alpha_{0;4}(f_r)\alpha_{2;4}(f_s) + 2\alpha_{1;4}(f_r)\alpha_{1;4}(f_s) + \alpha_{2;4}(f_r)\alpha_{0;4}(f_s)\;, \qquad \text{(D.21)} \\ \alpha_{3;2}(f_{rs}) \;&=\; \max_{i\le n}\zeta_{i;2}(\|D_i^3 f_{rs}(\mathbf{W}_i(\bullet))\|) \\ &\le \max_{i\le n}\Big(\zeta_{i;4}(|f_r(\mathbf{W}_i(\bullet))|)\,\zeta_{i;4}(\|\partial^3 f_s(\mathbf{W}_i(\bullet))\|) \\ &\qquad + 3\zeta_{i;4}(\|\partial f_r(\mathbf{W}_i(\bullet))\|)\,\zeta_{i;4}(\|\partial^2 f_s(\mathbf{W}_i(\bullet))\|) \\ &\qquad + 3\zeta_{i;4}(\|\partial^2 f_r(\mathbf{W}_i(\bullet))\|)\,\zeta_{i;4}(\|\partial f_s(\mathbf{W}_i(\bullet))\|) \\ &\qquad + \zeta_{i;4}(\|\partial^3 f_r(\mathbf{W}_i(\bullet))\|)\,\zeta_{i;4}(|f_s(\mathbf{W}_i(\bullet))|)\Big) \\ &\le \alpha_{0;4}(f_r)\alpha_{3;4}(f_s) + 3\alpha_{1;4}(f_r)\alpha_{2;4}(f_s) + 3\alpha_{2;4}(f_r)\alpha_{1;4}(f_s) + \alpha_{3;4}(f_r)\alpha_{0;4}(f_s)\;. \qquad \text{(D.22)}\end{aligned}$$

Therefore by (D.19), we get

$$T(f_{rs}) \;\le\; \delta nk^{1/2}\times\text{(D.21)}\times c_1 + nk^{3/2}\times\text{(D.22)}\times(c_X + c_{Z^\delta})\;.$$

Substitute this and the bound obtained in (D.20) for $T(f_r)$ and $T(f_s)$ into (D.18), we get

$$\begin{aligned}&(\text{Var}[f(\Phi\mathcal{X})])_{r,s} - (\text{Var}[f(\mathcal{Z})])_{r,s} \\ &\le \delta nk^{1/2}c_1\times\big(\alpha_{2;1}(f_r)\alpha_{0;1}(f_s) + \alpha_{2;1}(f_s)\alpha_{0;1}(f_r) + \alpha_{0;4}(f_r)\alpha_{2;4}(f_s) \\ &\qquad + 2\alpha_{1;4}(f_r)\alpha_{1;4}(f_s) + \alpha_{2;4}(f_r)\alpha_{0;4}(f_s)\big) \\ &\quad + nk^{3/2}(c_X + c_{Z^\delta})\times\big(\alpha_{3;2}(f_r)\alpha_{0;1}(f_s) + \alpha_{3;2}(f_s)\alpha_{0;1}(f_r) + \alpha_{0;4}(f_r)\alpha_{3;4}(f_s) \\ &\qquad + 3\alpha_{1;4}(f_r)\alpha_{2;4}(f_s) + 3\alpha_{2;4}(f_r)\alpha_{1;4}(f_s) + \alpha_{3;4}(f_r)\alpha_{0;4}(f_s)\big)\;.\end{aligned}$$

Note that summation of each term above over $1\le r,s\le q$ can be computed as

$$\big(\textstyle\sum_{r=1}^q \alpha_{R_1;m_1}(f_r)\big)\big(\sum_{s=1}^q \alpha_{R_2;m_2}(f_s)\big) \;\overset{(a)}{=}\; \alpha_{R_1;m_1}(f)\alpha_{R_2;m_2}(f)\;.$$

Therefore,

$$\|\text{Var}[f(\Phi\mathcal{X})] - \text{Var}[f(\mathcal{Z})]\| \;\le\; \textstyle\sum_{r,s=1}^q \big|[\text{Var}[f(\Phi\mathcal{X})]]_{r,s} - [\text{Var}[f(\mathcal{Z})]]_{r,s}\big|$$

$$
\begin{aligned}
&\le \delta n k^{1/2} c_1 (2\alpha_{2;1}(f)\alpha_{0;1}(f) + 2\alpha_{0;4}(f)\alpha_{2;4}(f) + 2\alpha_{1;4}(f)\alpha_{1;4}(f)) \\
&\quad + nk^{3/2}(c_X + c_{Z^\delta})(2\alpha_{3;2}(f)\alpha_{0;1}(f) + 2\alpha_{0;4}(f)\alpha_{3;4}(f) + 6\alpha_{1;4}(f)\alpha_{2;4}(f)) \\
&\overset{(b)}{\le} 4\delta n k^{1/2}(\alpha_{0;4}\alpha_{2;4} + \alpha_{1;4}^2)c_1 + 6nk^{3/2}(\alpha_{0;4}\alpha_{3;4} + \alpha_{1;4}\alpha_{2;4})(c_X + c_{Z^\delta}) \, .
\end{aligned}
$$

In (a), we have used Lemma D.26. In (b), we have omitted f-dependence and used that $\alpha_{2;1}\alpha_{0;1} \le \alpha_{0;4}\alpha_{2;4}$ and $\alpha_{3;2}\alpha_{0;1} \le \alpha_{3;4}\alpha_{0;4}$. Multiplying across by n gives the desired result. □

To prove Lemma D.3, we only need to apply the bound on $h \circ f$ from Theorem D.1 with f replaced by $\sqrt{n}f$.

Proof of Lemma D.3. Recall that for any $h \in \mathcal{H}$, $\gamma^1(h), \gamma^2(h), \gamma^3(h) \le 1$. Moreover, for $\zeta_{i;m}$ defined in Lemma D.20,

$$
\begin{aligned}
\alpha_{r;m}(\sqrt{n}f) &= \max_{i \le n} \zeta_{i;m}(\|\sqrt{n}\, D_i^r f(\mathbf{W}_i(\bullet))\|) \\
&= \sqrt{n}\max_{i \le n} \zeta_{i;m}(\| D_i^r f(\mathbf{W}_i(\bullet))\|) = \sqrt{n}\alpha_{r;m}(f) \, .
\end{aligned}
$$

Therefore, Theorem D.1 implies that for every $h \in \mathcal{H}$,

$$
\begin{aligned}
&\big|\mathbb{E}h(\sqrt{n}f(\Phi\mathcal{X})) - \mathbb{E}h(\sqrt{n}f(\mathcal{Z}^\delta))\big| \\
&\qquad \le \delta n k^{1/2} c_1 \big(n\alpha_{1;2}(f)^2 + n^{1/2}\alpha_{2;1}(f)\big) \\
&\qquad\quad + nk^{3/2}\big(n^{3/2}\alpha_{1;6}(f)^3 + 3n\alpha_{1;4}(f)\alpha_{2;4}(f) + n^{1/2}\alpha_{3;2}(f)\big)(c_X + c_{Z^\delta}) \, .
\end{aligned}
$$

Taking a supremum over all $h \in \mathcal{H}$ and omitting f-dependence imply that

$$
\begin{aligned}
&d_{\mathcal{H}}(\sqrt{n}f(\Phi\mathcal{X}), \sqrt{n}f(\mathcal{Z}^\delta)) = \sup_{h \in \mathcal{H}}\big|\mathbb{E}h(\sqrt{n}f(\Phi\mathcal{X})) - \mathbb{E}h(\sqrt{n}f(\mathcal{Z}^\delta))\big| \\
&\le \delta n^{3/2} k^{1/2} c_1 \big(n^{1/2}\alpha_{1;2}^2 + \alpha_{2;1}\big) + (nk)^{3/2}(n\alpha_{1;6}^3 + 3n^{1/2}\alpha_{1;4}\alpha_{2;4} + \alpha_{3;2})(c_X + c_{Z^\delta}) \, ,
\end{aligned}
$$

which is the desired bound. □

D.5.2. Proofs for Appendix D.1.2

We give the proofs for Lemma D.4, which concerns convergence when dimension of the statistic q is allowed to grow, and for Corollary D.5, which formulates our main result with the assumption of invariance. Both proofs are direct applications of Theorem D.1. The proof of Corollary D.6 is not stated as it is just obtained by setting $\delta = 1$ in Theorem D.1.

Proof of Lemma D.4. By assumption, the noise stability terms satisfy

$$
\alpha_1 = o(n^{-5/6}k^{-1/2}d^{-1/2}), \ \alpha_3 = o(n^{-3/2}k^{-3/2}d^{-3/2}), \ \alpha_0\alpha_3, \alpha_1\alpha_2 = o(n^{-2}k^{-3/2}d^{-3/2}).
$$

Since each coordinate of $\phi_{11}\mathbf{X}_1$ and $\mathbf{Z}_1$ is $O(1)$, the moment terms satisfy

$$c_X \;=\; \frac{1}{6}\big(\mathbb{E}[\|\phi_{11}\mathbf{X}_1\|^4]\big)^{3/4} \;=\; \frac{1}{6}\big(\mathbb{E}\big[\big(\textstyle\sum_{s=1}^{d}(\mathbf{e}_s^\top\phi_{11}\mathbf{X}_1)^2\big)^2\big]\big)^{3/4} \;=\; O(d^{3/2})\;,$$
$$c_Z \;=\; \frac{1}{6}\big(\mathbb{E}\big[\big(\frac{1}{k}\textstyle\sum_{j\le k,s\le d}|Z_{1jd}|^2\big)^2\big]\big)^{3/4} \;=\; O(d^{3/2})\;.$$

The condition on α_r's imply that the bound in Corollary 6.2, with δ set to 0, becomes

$$n\big\|\mathrm{Var}[f(\Phi\mathcal{X})] - \mathrm{Var}[f(\mathbf{Z}_1,\dots,\mathbf{Z}_n)]\big\| \;\le 6n^2k^{3/2}(c_X+c_Z)(\alpha_0\alpha_3+\alpha_1\alpha_2) \;=\; o(1)\;.$$

Since $\alpha_r(f_s)\le\alpha_r(f)$ by definition of α_r, the above bounds hold for $\alpha_r(f_s)$. Applying Corollary 6.4 to f_s gives

$$\begin{aligned}&d_{\mathcal{H}}(\sqrt{n}f_s(\Phi\mathcal{X}),\sqrt{n}f_s(\mathbf{Z}_1,\dots,\mathbf{Z}_n))\\ &\le\; n^{3/2}k^{3/2}(n\alpha_1(f_s)^3+3n^{1/2}\alpha_1(f_s)\alpha_2(f_s)+\alpha_3(f_s))(c_X+c_Z) \;=\; o(1)\;.\end{aligned}$$

By Lemma 6.3, convergence in $d_{\mathcal{H}}$ implies weak convergence, which gives the desired result. □

Proof of Corollary D.5. By law of total variance,

$$\Sigma_{11} \;\coloneqq\; \mathrm{Var}[\phi_{11}\mathbf{X}_1] \;=\; \tilde{\Sigma}_{11} + \mathrm{Var}\mathbb{E}[\phi_{11}\mathbf{X}_1|\phi_{11}]\;,$$

and by distributional invariance assumption, almost surely,

$$\mathbb{E}[\phi_{11}\mathbf{X}_1|\phi_{11}] \;=\; \mathbb{E}[\phi_{12}\mathbf{X}_1|\phi_{12}] \;=\; \mathbb{E}[\mathbf{X}_1]\;.$$

This implies $\mathrm{Var}\mathbb{E}[\phi_{11}\mathbf{X}_1|\phi_{11}]$ vanishes and therefore $\Sigma_{11}=\tilde{\Sigma}_{11}$. The equality in Σ_{12} is directly from Lemma D.19. □

D.5.3. Proofs for Appendix D.1.3

We present the proofs for the two results of Lemma D.7 for plug-in estimates. The following lemma is analogous to Lemma D.26 but for $\kappa_{r;m}$, and will be useful in the proof.

Lemma D.28. $\big\|\sup_{\mathbf{w}\in[\mathbf{0},\bar{\mathbf{X}}]}\|\partial^r g(\mu+\mathbf{w})\|\big\|_{L_m} \le \kappa_{r;m}(g)\;.$

Proof. By the definition of $\kappa_{r;m}$ and the triangle inequality,

$$\kappa_{r;m}(g) \coloneqq \sum_{s\le q}\big\|\sup_{\mathbf{w}\in[\mathbf{0},\bar{\mathbf{X}}]}\big\|\partial^r g_s\big(\mu+\mathbf{w}\big)\big\|\big\|_{L_m} \ge \big\|\sup_{\mathbf{w}\in[\mathbf{0},\bar{\mathbf{X}}]}\big\|\partial^r g\big(\mu+\mathbf{w}\big)\big\|\big\|_{L_m}\;,$$

which is the desired bound. □

For the proof of Lemma D.7(i), we first compare g to its first-order Taylor expansion. The Taylor expansion only involves an empirical average, whose weak convergence and

equality in variance are given by Lemma D.2 and Lemma D.3 in a similar manner as the proof for Proposition 6.7. We recall that $\mathcal{D}$ is assumed to be a convex subset in $\mathbb{R}^d$ containing $\mathbf{0}$, which is important for the Taylor expansion argument.

Proof of Lemma D.7(i). We first prove the bound in $d_{\mathcal{H}}$. Using the triangle inequality to separate the bound into two parts, we get

$$\begin{aligned} d_{\mathcal{H}}(\sqrt{n}f(\Phi\mathcal{X}),\sqrt{n}f^T(\mathcal{Z}^\delta)) &= \sup_{h\in\mathcal{H}} |\mathbb{E}[h(\sqrt{n}f(\Phi\mathcal{X})) - \mathbb{E}[h(\sqrt{n}f^T(\mathcal{Z}^\delta))]| \\ &\leq d_{\mathcal{H}}(\sqrt{n}f(\Phi\mathcal{X}), \sqrt{n}f^T(\Phi\mathcal{X})) + d_{\mathcal{H}}(\sqrt{n}f^T(\Phi\mathcal{X}), \sqrt{n}f^T(\mathcal{Z}^\delta)) \,. \end{aligned} \tag{D.23}$$

Consider bounding the first term of (D.23). Since $f(\Phi\mathcal{X}) = g\big(\bar{\mathbf{X}} + \mu\big)$ and $f^T(\Phi\mathcal{X}) = g(\mu) + \partial g(\mu)\bar{\mathbf{X}}$, a Taylor expansion argument on $g\big(\bar{\mathbf{X}} + \mu\big)$ gives

$$\big\|f(\Phi\mathcal{X}) - f^T(\Phi\mathcal{X})\big\| \leq \sup_{\mathbf{w}\in[\mathbf{0},\bar{\mathbf{X}}]} \big\|\partial^2 g(\mu + \mathbf{w})\big\| \, \big\|\bar{\mathbf{X}}\big\|^2 \,.$$

Recall that $\gamma_1(h) = \sup_{\mathbf{w}\in\mathbb{R}^q}\{\|\partial h(\mathbf{w})\|\}$. By mean value theorem, the above bound and Hölder's inequality, we get

$$\begin{aligned} |\mathbb{E}h(\sqrt{n}f(\Phi\mathcal{X})) - \mathbb{E}\,h(\sqrt{n}f^T(\Phi\mathcal{X}))| &\leq \sqrt{n}\,\gamma_1(h)\,\mathbb{E}\|f(\Phi\mathcal{X}) - f^T(\Phi\mathcal{X})\| \\ &\leq \sqrt{n}\,\gamma_1(h)\,\mathbb{E}\big[\sup_{\mathbf{w}\in[\mathbf{0},\bar{\mathbf{X}}]} \|\partial^2 g(\mu + \mathbf{w})\| \, \big\|\bar{\mathbf{X}}\big\|^2\big] \\ &\leq \sqrt{n}\,\gamma_1(h)\,\big\|\sup_{\mathbf{w}\in[\mathbf{0},\bar{\mathbf{X}}]} \|\partial^2 g(\mu + \mathbf{w})\|\big\|_{L_3} \, \big\|\,\|\bar{\mathbf{X}}\|\,\big\|_{L_3}^2 \\ &\leq \sqrt{n}\,\gamma_1(h)\,\kappa_{2;3}(g)\,\big\|\,\|\bar{\mathbf{X}}\|\,\big\|_{L_3}^2 \,. \end{aligned}$$

In the last inequality we have used Lemma D.28. To control the moment term, we use Rosenthal's inequality for vectors from Corollary D.22. Since $\phi_{ij}\mathbf{X}_i$ have bounded 6th moments, for each $2 \leq m \leq 6$, there exists a constant K_m depending only on m such that

$$\begin{aligned} \big\|\,\|\bar{\mathbf{X}}\|\,\big\|_{L_m} &= \Big\|\Big\|\frac{1}{nk}\textstyle\sum_{i=1}^n\sum_{j=1}^k \phi_{ij}\mathbf{X}_i - \mu\Big\|\Big\|_{L_m} \\ &\leq \frac{K_m}{n}\Bigg(\sum_{s=1}^d \max\Bigg\{\Bigg(\sum_{i=1}^n \Big\|\frac{1}{k}\sum_{j=1}^k(\phi_{ij}\mathbf{X}_i - \mu)_s\Big\|_{L_m}^m\Bigg)^{2/m}, \\ &\qquad\qquad \sum_{i=1}^n \Big\|\frac{1}{k}\sum_{j=1}^k(\phi_{ij}\mathbf{X}_i - \mu)_s\Big\|_{L_2}^2\Bigg\}\Bigg)^{1/2} \\ &= \frac{K_m}{\sqrt{n}}\Big(\sum_{s=1}^d \max\big\{n^{\frac{2}{m}-1}\big\|\tfrac{1}{k}\textstyle\sum_{j=1}^k(\phi_{1j}\mathbf{X}_1 - \mu 0_s\big\|_{L_m}^2, \, \big\|\tfrac{1}{k}\sum_{j=1}^k(\phi_{1j}\mathbf{X}_1 - \mu)_s\big\|_{L_2}^2\big\}\Big)^{1/2} \\ &= O(n^{-1/2}\bar{c}_m) \,. \end{aligned} \tag{D.24}$$

Substituting this into the bound above, we get

$$\big|\mathbb{E}h(\sqrt{n}f(\Phi\mathcal{X})) - \mathbb{E}h\big(\sqrt{n}f^T(\Phi\mathcal{X})\big)\big| = O\big(n^{-1/2}\gamma^1(h)\,\kappa_{2;3}(g)\,\bar{c}_3^2\big) \,.$$

Since for all $h \in \mathcal{H}$, $\gamma^1(h) \le 1$, taking supremum of the above over $h \in \mathcal{H}$ gives the bound for the first term of (D.23):

$$d_H\big(\sqrt{n}f(\Phi\mathcal{X}), \sqrt{n}f^T(\Phi\mathcal{X})\big) \;=\; O\big(n^{-1/2}\,\kappa_{2;3}(g)\,\bar{c}_3^2\big)\;. \tag{D.25}$$

The second term of (D.23) can be bounded in the usual way by applying Lemma D.3 to $f^T(\mathbf{x}_{11:nk}) = g(\mu) + \partial g(\mu)\big(\frac{1}{nk}\sum_{i,j}\mathbf{x}_{ij} - \mu\big)$. Let f_s^T denote the sth coordinate of f^T. The partial derivatives are given by:

$$\Big\|\frac{\partial f_s^T(\mathbf{x}_{11:nk})}{\partial \mathbf{x}_{ij}}\Big\| = \frac{1}{nk}\|\partial g_s(\mu)\|, \qquad \Big\|\frac{\partial^2 f_s^T(\mathbf{x}_{11:nk})}{\partial \mathbf{x}_{ij_1}\partial \mathbf{x}_{ij_2}}\Big\| = \Big\|\frac{\partial^3 f_s^T(\mathbf{x}_{11:nk})}{\partial \mathbf{x}_{ij_1}\partial \mathbf{x}_{ij_2}\partial \mathbf{x}_{ij_3}}\Big\| = 0.$$

This implies that for $s \le q$,

$$\|D_i f_s^T(\mathbf{x}_{11:nk})\| \;=\; \frac{1}{nk^{1/2}}\|\partial g_s(\mu)\|\;,\quad \|D_i^2 f_s^T(\mathbf{x}_{11:nk})\| \;=\; \|D_i^3 f_s^T(\mathbf{x}_{11:nk})\| \;=\; 0\;.$$

Thus we have $\alpha_{1;m}\big(f^T\big) = \sum_{s=1}^q n^{-1}k^{-1/2}\|\partial g_s(\mu)\| \le n^{-1}k^{-1/2}\kappa_{1;1}(g)$ by Lemma D.28, and $\alpha_{2;m}\big(f^T\big) = \alpha_{3;m}\big(f^T\big) = 0$. The bound in Lemma D.3 then becomes

$$\begin{aligned}&\delta n^{3/2}k^{1/2}c_1\big(n^{1/2}(\alpha_{1;2})^2 + \alpha_{2;1}\big) + (nk)^{3/2}(n(\alpha_{1;6})^3 + 3n^{1/2}\alpha_{1;4}\alpha_{2;4} + \alpha_{3;2})(c_X + c_{Z^\delta})\\ &= O\big(\delta k^{-1/2}\kappa_{1;1}(g)^2 c_1 + n^{-1/2}\kappa_{1;1}(g)^3(c_X + c_{Z^\delta})\big)\;,\end{aligned}$$

which implies

$$d_{\mathcal{H}}(\sqrt{n}f^T(\Phi\mathcal{X}), \sqrt{n}f^T(\mathcal{Z}^\delta)) \;=\; O\big(\delta k^{-1/2}\kappa_{1;1}(g)^2 c_1 + n^{-1/2}\kappa_{1;1}(g)^3(c_X + c_{Z^\delta})\big)\;.$$

Substituting this into (D.23) together with the bound in (D.25) gives the required bound

$$d_{\mathcal{H}}\big(\sqrt{n}f(\Phi\mathcal{X}), \sqrt{n}f^T(\mathcal{Z}^\delta)\big) \;=\; O\big(n^{-1/2}\kappa_{2;3}\,\bar{c}_3^2 + \delta k^{-1/2}\kappa_{1;1}^2 c_1 + n^{-1/2}\kappa_{1;1}^3(c_X + c_{Z^\delta})\big)\;,$$

where we have omitted g-dependence.

Recall that $\Sigma_{11} = \mathrm{Var}[\phi_{11}\mathbf{X}_1]$ and $\Sigma_{12} = \mathrm{Cov}[\phi_{11}\mathbf{X}_1, \phi_{12}\mathbf{X}_1]$. For the bound on variance, we first note that by the variance condition on $\mathbf{Z}_i^\delta$ from (D.1), we get

$$\begin{aligned}\mathrm{Var}[\bar{\mathbf{X}}] - \mathrm{Var}[\bar{\mathbf{Z}}^\delta] \;&=\; \frac{1}{n}\Big(\frac{1}{k}\Sigma_{11} + \frac{k-1}{k}\Sigma_{12}\Big) - \frac{1}{n}\Big(\frac{1}{k}((1-\delta)\Sigma_{11} + \delta\Sigma_{12}) + \frac{k-1}{k}\Sigma_{12}\Big)\\ &= \frac{\delta}{nk}(\Sigma_{11} - \Sigma_{12})\;.\end{aligned}$$

This implies

$$\begin{aligned}n\|\mathrm{Var}[f^T(\Phi\mathcal{X})] - \mathrm{Var}[f^T(\mathcal{Z})]\| \;&= n\|\mathrm{Var}[g(\mu) + \partial g(\mu)\bar{\mathbf{X}}] - \mathrm{Var}[g(\mu) + \partial g(\mu)\bar{\mathbf{Z}}^\delta]\|\\ &= n\big\|\partial g(\mu)\mathrm{Var}[\bar{\mathbf{X}}]\partial g(\mu)^\top - \partial g(\mu)\mathrm{Var}[\bar{\mathbf{Z}}^\delta]\partial g(\mu)^\top\big\|\\ &= n\big\|\frac{\delta}{nk}\partial g(\mu)(\Sigma_{11} - \Sigma_{12})\partial g(\mu)^\top\big\|\\ &\le \frac{4\delta}{k}\|\partial g(\mu)\|_2^2\, c_1^2\;,\end{aligned} \tag{D.26}$$

where in the inequality we have recalled that $2c_1 \coloneqq \|\mathbb{E}\mathrm{Var}[\phi_{11}\mathbf{X}_1|\mathbf{X}_1]\| = \|\Sigma_{11} - \Sigma_{12}\|$

by Lemma D.19. Next, we bound the quantity

$$n\|\mathrm{Var}[f(\Phi\mathcal{X})] - \mathrm{Var}[f^T(\Phi\mathcal{X})]\| \ ,$$

for which we use a second-order Taylor expansion on each coordinate of the covariance matrix. For every $s \le q$, let $f_s(\mathbf{x}_{11:nk})$ and $g_s(\mathbf{x}_{11:nk})$ be the s^{th} coordinate of $f(\mathbf{x}_{11:nk})$ and $g(\mathbf{x}_{11:nk})$ respectively, i.e. f_s, g_s are both functions $\mathcal{D} \to \mathbb{R}$. Then there exists $\tilde{\mathbf{X}}^{(s)} \in \big[\mathbf{0}, \bar{\mathbf{X}}\big]$ such that

$$f_s(\Phi\mathcal{X}) \ = \ g_s(\mu) + (\partial g_s(\mu))^\top \bar{\mathbf{X}} + \mathrm{Tr}\big((\partial^2 g_s(\mu + \tilde{\mathbf{X}}^{(s)}))^\top \bar{\mathbf{X}}\bar{\mathbf{X}}^\top\big) \ . \tag{D.27}$$

Denote for convenience

$$\mathbf{R}_s^1 \ = \ (\partial g_s(\mu))^\top \bar{\mathbf{X}} \ , \qquad\qquad \mathbf{R}_s^2 \ = \ \mathrm{Tr}\big((\partial^2 g_s(\mu + \tilde{\mathbf{X}}^{(s)}))^\top \bar{\mathbf{X}}\bar{\mathbf{X}}^\top\big) \ ,$$

The Taylor expansion above allows us to control the difference in variance at (r,s)-th coordinate:

$$\begin{aligned}
n\big(\mathrm{Var}[f(\Phi\mathcal{X})] - \mathrm{Var}\big[f^T(\Phi\mathcal{X})\big]\big)_{r,s} \ &= \ n\big((\mathrm{Var}[f(\Phi\mathcal{X})])_{r,s} - \big(\mathrm{Var}\big[g(\mu) + \partial g(\mu)\bar{\mathbf{X}}\big]\big)_{r,s}\big) \\
&= n\big(\mathrm{Cov}[f_r(\Phi\mathcal{X}), f_s(\Phi\mathcal{X})] - \mathrm{Cov}\big[g_r(\mu) + \mathbf{R}_r^1, \ g_s(\mu) + \mathbf{R}_s^1\big]\big) \\
&\overset{(a)}{=} n\big(\mathrm{Cov}\big[\mathbf{R}_r^1 + \mathbf{R}_r^2, \ \mathbf{R}_s^1 + \mathbf{R}_s^2\big] - \mathrm{Cov}\big[\mathbf{R}_r^1, \ \mathbf{R}_s^1\big]\big) \\
&= n\big(\mathrm{Cov}[\mathbf{R}_r^1, \mathbf{R}_s^2] + \mathrm{Cov}[\mathbf{R}_r^2, \mathbf{R}_s^1] + \mathrm{Cov}[\mathbf{R}_r^2, \mathbf{R}_s^2]\big) \ .
\end{aligned} \tag{D.28}$$

In (a), we have used (D.27) and the fact that $g_r(\mu)$ and $g_s(\mu)$ are deterministic. To control the first covariance term, by noting that $\mathbb{E}[\bar{\mathbf{X}}] = \mathbf{0}$, Cauchy-Schwarz and Hölder's inequality, we get

$$\begin{aligned}
\mathrm{Cov}[\mathbf{R}_r^1, \mathbf{R}_s^2] \ &= \mathbb{E}[\mathbf{R}_r^1 \mathbf{R}_s^2] \ = \ \mathbb{E}\big[(\partial g_r(\mu))^\top \bar{\mathbf{X}} \ \mathrm{Tr}\big((\partial^2 g_s(\mu + \tilde{\mathbf{X}}^{(s)}))^\top \bar{\mathbf{X}}\bar{\mathbf{X}}^\top\big)\big] \\
&\le \|\partial g_r(\mu)\| \mathbb{E}\big[\|\partial^2 g_s(\mu + \tilde{\mathbf{X}}^{(s)})\| \|\bar{\mathbf{X}}\|^3\big] \\
&\le \|\partial g_r(\mu)\| \ \big\|\|\partial^2 g_s(\mu + \tilde{\mathbf{X}}^{(s)})\|\big\|_{L_4} \ \big\|\|\bar{\mathbf{X}}\|\big\|_{L_4}^3 \\
&\overset{(b)}{=} O\big(n^{-3/2} \kappa_{1;1}(g_r) \kappa_{2;4}(g_s) \, \bar{c}_4^3\big) \ .
\end{aligned}$$

In (b), we have used the definition of κ_m^r and the bound on moments of $\bar{\mathbf{X}}$ computed in (D.24). An analogous argument gives

$$\mathrm{Cov}[\mathbf{R}_r^1, \mathbf{R}_s^2] \ = \ O\big(n^{-3/2} \kappa_{1;1}(g_s) \kappa_{2;4}(g_r) \, \bar{c}_4^3\big) \ ,$$

and also

$$\begin{aligned}
&\mathrm{Cov}[\mathbf{R}_r^2, \mathbf{R}_s^2] \\
&\le \ \big|\mathbb{E}\big[\mathrm{Tr}\big((\partial^2 g_r(\mu + \tilde{\mathbf{X}}^{(r)}))^\top \bar{\mathbf{X}}\bar{\mathbf{X}}^\top\big) \ \mathrm{Tr}\big((\partial^2 g_s(\mu + \tilde{\mathbf{X}}^{(s)}))^\top \bar{\mathbf{X}}\bar{\mathbf{X}}^\top\big)\big]\big| \\
&\quad + \big|\mathbb{E}\big[\mathrm{Tr}\big((\partial^2 g_r(\mu + \tilde{\mathbf{X}}^{(r)}))^\top \bar{\mathbf{X}}\bar{\mathbf{X}}^\top\big)\big]\big| \ \big|\mathbb{E}\big[\mathrm{Tr}\big((\partial^2 g_s(\mu + \tilde{\mathbf{X}}^{(s)}))^\top \bar{\mathbf{X}}\bar{\mathbf{X}}^\top\big)\big]\big| \\
&\le \ 2\big\|\|\partial^2 g_r(\mu + \tilde{\mathbf{X}}^{(r)})\|\big\|_{L_6} \ \big\|\|\partial^2 g_s(\mu + \tilde{\mathbf{X}}^{(s)})\|\big\|_{L_6} \ \big\|\|\bar{\mathbf{X}}\|\big\|_{L_6}^4
\end{aligned}$$

$$= O\big(n^{-2}\kappa_{2;6}(g_r)\kappa_{2;6}(g_s)\,\bar{c}_6^4\big)\ .$$

Substituting the bounds on each covariance term back into (D.28), we get that

$$\begin{aligned} &n\big((\mathrm{Var}[f(\Phi\mathcal{X})])_{r,s} - \big(\mathrm{Var}\big[f^T(\Phi\mathcal{X})\big]\big)_{r,s}\big) \\ &= O\big(n^{-1/2}(\kappa_{1;1}(g_r)\kappa_{2;4}(g_s) + \kappa_{1;1}(g_s)\kappa_{2;4}(g_r))\bar{c}_4^3 + n^{-1}\kappa_{2;6}(g_r)\kappa_{2;6}(g_s)\,\bar{c}_6^4\big)\ . \end{aligned}$$

Note that by the definition of $\kappa_{R;m}$ in Lemma D.7,

$$\textstyle\sum_{r,s=1}^{q}\kappa_{R_1;m_1}(g_r)\kappa_{R_2;m_2}(g_s) \;=\; \kappa_{R_1;m_1}(g)\,\kappa_{R_2;m_2}(g)\ ,$$

so summing the bound above over $r, s \le q$ gives the bound,

$$n\big\|\mathrm{Var}[f(\Phi\mathcal{X})] - \mathrm{Var}\big[f^T(\Phi\mathcal{X})\big]\big\| \;=\; O\big(n^{-1/2}\kappa_{1;1}(g)\kappa_{2;4}(g)\bar{c}_4^3 + n^{-1}\kappa_{2;6}(g)^2\bar{c}_6^4\big)\ .$$

Combine this with the bound from (D.26) and omitting g-dependence gives

$$n\big\|\mathrm{Var}[f(\Phi\mathcal{X})] - \mathrm{Var}\big[f^T(\mathcal{Z}^\delta)\big]\big\| = O\big(\delta k^{-1}\|\partial g(\mu)\|_2^2\,c_1^2 + n^{-1/2}\kappa_{1;1}\kappa_{2;4}\bar{c}_4^3 + n^{-1}\kappa_{2;6}^2\bar{c}_6^4\big)\ .$$

□

For Lemma D.7(ii), we only need to rewrite the noise stability terms $\alpha_{r;m}(f)$ in Lemma D.2 and D.3 in terms of $\nu_{r;m}(g)$.

Proof of Lemma D.7(ii). We just need to compute the bounds in Lemma D.2 (concerning variance) and Lemma D.3 (concerning $d_{\mathcal{H}}$) in terms of $\nu_{r;m}(g)$, which boils down to rewriting $\alpha_{r;m}(f)$ in terms of $\nu_{r;m}(g)$. As usual, we start with computing partial derivatives of $f_s(\mathbf{x}_{11:nk}) = g\big(\frac{1}{nk}\sum_{i\le n,j\le k}\mathbf{x}_{ij}\big)$:

$$\begin{aligned} \frac{\partial}{\partial\mathbf{x}_{ij}}f_s(\mathbf{x}_{11:nk}) &= \frac{1}{nk}\partial g_s\big(\frac{1}{nk}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{k}\mathbf{x}_{ij}\big)\ , \\ \frac{\partial^2}{\partial\mathbf{x}_{ij_1}\partial x_{ij_2}}\tilde{f}_s(\mathbf{x}_{11:nk}) &= \frac{1}{n^2k^2}\partial^2 g_s\big(\frac{1}{nk}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{k}\mathbf{x}_{ij}\big)\ , \\ \frac{\partial^3}{\partial\mathbf{x}_{ij_1}\partial\mathbf{x}_{ij_2}\partial\mathbf{x}_{ij_3}}\tilde{f}_s(\mathbf{x}_{11:nk}) &= \frac{1}{n^3k^3}\partial^3 g_s\big(\frac{1}{nk}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{k}\mathbf{x}_{ij}\big)\ . \end{aligned}$$

Norm of the first partial derivative is given by

$$\|D_i f_s(\mathbf{x}_{11:nk})\| \;=\; \sqrt{\textstyle\sum_{j=1}^{k}\Big\|\frac{\partial}{\partial\mathbf{x}_{ij}}f_s(\mathbf{x}_{11:nk})\Big\|^2} \;=\; \frac{1}{nk^{1/2}}\Big\|\partial g_s\Big(\frac{1}{nk}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{k}\mathbf{x}_{ij}\Big)\Big\|\ ,$$

and therefore, by the definitions of $\alpha_{r;m}$ from Theorem D.1 and $\nu_{r;m}$ from (D.3),

$$\begin{aligned} \alpha_{1;m}(f) &:= \textstyle\sum_{s\le q}\max_{i\le n}\zeta_{i;m}\big(|D_i f_s(\mathbf{W}_i(\bullet))|\big) \\ &= \frac{1}{nk^{1/2}}\textstyle\sum_{s\le q}\max_{i\le n}\zeta_{i;m}\big(|D_i g_s(\overline{\mathbf{W}}_i(\bullet)|\big) \;=\; \frac{1}{nk^{1/2}}\nu_{1;m}(g)\ . \end{aligned}$$

Similarly we get $\alpha_{2;m}(f) = \frac{1}{n^2k}\nu_{2;m}(g)$, $\alpha_{3;m}(f) = \frac{1}{n^3k^{3/2}}\nu_{3;m}(g)$ and $\alpha_{0;m}(f) =$

$\nu_{0;m}(g)$.The bound in Lemma D.3 can then be computed as

$$\begin{aligned}&\delta n^{3/2}k^{1/2}c_1\big(n^{1/2}\alpha_{1;2}^2+\alpha_{2;1}\big)+(nk)^{3/2}(n\alpha_{1;6}^3+3n^{1/2}\alpha_{1;4}\alpha_{2;4}+\alpha_{3;2})(c_X+c_{Z^\delta})\,.\\&=\delta\big(k^{-1/2}\nu_{1;2}^2+n^{-1/2}k^{-1/2}\nu_{2;1}\big)c_1+\big(n^{-1/2}\nu_{1;6}^3+3n^{-1}\nu_{1;4}\nu_{2;4}+n^{-3/2}\nu_{3;2}\big)(c_X+c_{Z^\delta})\,,\end{aligned}$$

while the bound in Lemma D.2 can be computed as

$$\begin{aligned}&4\delta n^2k^{1/2}(\alpha_{0;4}\alpha_{2;4}+\alpha_{1;4}^2)c_1+6n^2k^{3/2}(\alpha_{0;4}\alpha_{3;4}+\alpha_{1;4}\alpha_{2;4})(c_X+c_{Z^\delta})\\&=O\big(\delta k^{-1/2}(\nu_{0;4}\nu_{2;4}+\nu_{1;4}^2)c_1+n^{-1}(\nu_{0;4}\nu_{3;4}+\nu_{1;4}\nu_{2;4})(c_X+c_{Z^\delta})\big)\,.\end{aligned}$$

These give the desired bounds on the differences $d_{\mathcal{H}}(\sqrt{n}f(\Phi\mathcal{X}),\sqrt{n}f(\mathcal{Z}^\delta))$ and $n\|\mathrm{Var}[f(\Phi\mathcal{X})]-\mathrm{Var}[f(\mathcal{Z}^\delta)]\|$. □

D.5.4. Proofs for Appendix D.1.4

We present the proof for Theorem D.8, which concerns non-smooth statistics in high-dimensions. We also prove Corollary D.9 and D.10, which are respectively variants of Lemma D.7(ii) for non-smooth statistics and Lemma D.7(i) for high-dimensions. Throughout this section, we use x_{ijl} to denote the l^{th} coordinate of a d-dimensional vector $\mathbf{x}_{ij}$.

The general idea for proving Theorem D.8 is to apply an intermediate result of the proof of Theorem D.1 to $f^{(t)}$, some smooth approximation of f. By taking Hölder's inequality differently, we obtain vector-∞ norm for the moments as desired. The final bound is then obtained by combining a bound analogous to that of Theorem D.1 and an approximation error term using $\varepsilon(t)$ and moment terms of $f^{(t)}$.

Proof of Theorem D.8. Recall from an intermediate equation (D.14) in the proof of Theorem D.1 (for which Theorem 4.1 is a special case), with g replaced by $h\circ f^{(t)}$ and δ set to 0, that

$$\big|\mathbb{E}h\big(f^{(t)}(\Phi\mathcal{X})\big)-\mathbb{E}h\big(f^{(t)}(\mathcal{Z})\big)\big|\;\le\;\textstyle\sum_{i\le n}\big(|\kappa_{1,i}|+\frac{1}{2}|\kappa_{2,i}|+\frac{1}{6}|\kappa_{3,i}|\big)\,,\qquad\text{(D.29)}$$

where we have shown $\kappa_{1,i}=0$ and $\kappa_{2,i}=0$ for $\delta=0$ in the proof of Theorem D.1, and written

$$\begin{aligned}\kappa_{3,i}\;:=\;&\mathbb{E}\big[\,\sup\nolimits_{\mathbf{w}\in[\mathbf{0},\Phi_i\mathbf{X}_i]}\big|D_i^3(h\circ f^{(t)})\big(\mathbf{W}_i(\mathbf{w})\big)(\Phi_i\mathbf{X}_i)^{\otimes3}\big|\\&+\sup\nolimits_{\mathbf{w}\in[\mathbf{0},\mathbf{Z}_i]}\big|D_i^3(h\circ f^{(t)})\big(\mathbf{W}_i(\mathbf{w})\big)(\mathbf{Z}_i)^{\otimes3}\big|\big]\,.\end{aligned}$$

We now seek to bound $\kappa_{3,i}$. We apply the same argument as the original proof, except that in step (a) below we provide a bound with a vector-1 norm and a vector-∞ norm

instead of the Cauchy-Schwarz inequality:

$$
\begin{aligned}
&\big|D_i^3(h\circ f^{(t)})\big(\mathbf{W}_i(\mathbf{w})\big)(\Phi_i\mathbf{X}_i)^{\otimes 3}\big| \\
&= \sum_{j_1,j_2,j_3=1}^{k}\sum_{l_1,l_2,l_3=1}^{d}\Big(\frac{\partial^3}{\partial x_{ij_1l_1}\partial x_{ij_2l_2}\partial x_{ij_3l_3}}(h\circ f^{(t)})\big(\mathbf{W}_i(\mathbf{w})\big)\Big)(\phi_{ij_1}\mathbf{X}_i)_{l_1}(\phi_{ij_2}\mathbf{X}_i)_{l_2}(\phi_{ij_3}\mathbf{X}_i)_{l_3} \\
&\overset{(a)}{\leq} \sum_{j_1,j_2,j_3=1}^{k}\Big(\sum_{l_1,l_2,l_3=1}^{d}\Big|\frac{\partial^3}{\partial x_{ij_1l_1}\partial x_{ij_2l_2}\partial x_{ij_3l_3}}(h\circ f^{(t)})\big(\mathbf{W}_i(\mathbf{w})\big)\Big|\Big)\prod_{r=1}^{3}\max_{l\leq d}|(\phi_{ij_r}\mathbf{X}_i)_l| \\
&\leq \Big(\sum_{j_1,j_2,j_3=1}^{k}\Big(\sum_{l_1,l_2,l_3=1}^{d}\Big|\frac{\partial^3(h\circ f^{(t)})(\mathbf{W}_i(\mathbf{w}))}{\partial x_{ij_1l_1}\partial x_{ij_2l_2}\partial x_{ij_3l_3}}\Big|\Big)^2\Big)^{1/2}\Big(\sum_{j=1}^{k}\max_{l\leq d}|(\phi_{ij}\mathbf{X}_i)_l|^2\Big)^{3/2} \\
&=: U_i(\mathbf{w})\Big(\textstyle\sum_{j=1}^{k}\max_{l\leq d}|(\phi_{ij}\mathbf{X}_i)_l|^2\Big)^{3/2}.
\end{aligned}
$$

This together with the Cauchy-Schwarz inequality implies

$$
\begin{aligned}
&\mathbb{E}\big[\sup_{\mathbf{w}\in[\mathbf{0},\Phi_i\mathbf{X}_i]}\big|D_i^3(h\circ f^{(t)})\big(\mathbf{W}_i(\mathbf{w})\big)(\Phi_i\mathbf{X}_i)^{\otimes 3}\big|\big] \\
&\qquad\leq \mathbb{E}\Big[U_i(\mathbf{w})\Big(\textstyle\sum_{j=1}^{k}\max_{l\leq d}|(\phi_{ij}\mathbf{X}_i)_l|^2\Big)^{3/2}\Big] \\
&\qquad\leq \zeta_{i;2}(U_i(\bullet))\,\Big\|\Big(\textstyle\sum_{j=1}^{k}\max_{l\leq d}|(\phi_{ij}\mathbf{X}_i)_l|^2\Big)^{3/2}\Big\|_{L_2}.
\end{aligned}
$$

Moreover, by the Jensen's inequality on the convex function $x\to x^3$ defined on $\mathbb{R}^+$ and noting that $\phi_{ij}\mathbf{X}_i$ is identically distributed as $\phi_{i1}\mathbf{X}_i$, we have

$$
\begin{aligned}
\Big\|\Big(\textstyle\sum_{j=1}^{k}\max_{l\leq d}|(\phi_{ij}\mathbf{X}_i)_l|^2\Big)^{3/2}\Big\|_{L_2} &= k^{3/2}\sqrt{\mathbb{E}\Big[\Big(\frac{1}{k}\sum_{j=1}^{k}\max_{l\leq d}|(\phi_{ij}\mathbf{X}_i)_l|^2\Big)^3\Big]} \\
&\leq k^{3/2}\sqrt{\mathbb{E}\Big[\Big(\frac{1}{k}\sum_{j=1}^{k}\max_{l\leq d}|(\phi_{ij}\mathbf{X}_i)_l|^6\Big)\Big]} \\
&= k^{3/2}\sqrt{\mathbb{E}\big[\max_{l\leq d}|(\phi_{ij}\mathbf{X}_i)_l|^6\big]} \;=\; 6k^{3/2}\tilde{c}_X\,,
\end{aligned}
$$

which implies that the term in $\kappa_{3,i}$ involving $\Phi_i\mathbf{X}_i$ can be bounded as

$$
\mathbb{E}\big[\sup_{\mathbf{w}\in[\mathbf{0},\Phi_i\mathbf{X}_i]}\big|D_i^3(h\circ f^{(t)})\big(\mathbf{W}_i(\mathbf{w})\big)(\Phi_i\mathbf{X}_i)^{\otimes 3}\big|\big]\leq 6k^{3/2}\zeta_{i;2}(U_i(\bullet))\,\tilde{c}_X.
$$

On the other hand, the same argument applies to the term in $\kappa_{3,i}$ involving $\mathbf{Z}_i$ to give

$$
\begin{aligned}
&\mathbb{E}\big[\sup_{\mathbf{w}\in[\mathbf{0},\mathbf{Z}_i]}\big|D_i^3(h\circ f^{(t)})\big(\mathbf{W}_i(\mathbf{w})\big)(\Phi_i\mathbf{X}_i)^{\otimes 3}\big|\big] \\
&\qquad\leq \zeta_{i;2}(U_i(\bullet))\,\Big\|\Big(\textstyle\sum_{j=1}^{k}\max_{l\leq d}|(\mathbf{Z}_{ij})_l|^2\Big)^{3/2}\Big\|_{L_2}.
\end{aligned}
$$

The moment term can similarly be bounded by the Jensen's inequality as

$$
\begin{aligned}
\Big\|\Big(\textstyle\sum_{j=1}^{k}\max_{l\leq d}|(\mathbf{Z}_{ij})_l|^2\Big)^{3/2}\Big\|_{L_2} &= 6k^{3/2}\sqrt{\mathbb{E}\Big[\Big(\frac{1}{k}\sum_{j=1}^{k}\max_{l\leq d}|(\mathbf{Z}_{ij})_l|^2\Big)^3\Big]} \\
&\leq 6k^{3/2}\sqrt{\mathbb{E}\Big[\frac{1}{k}\sum_{j=1}^{k}\max_{l\leq d}|(\mathbf{Z}_{ij})_l|^6\Big]} \;=\; 6k^{3/2}\tilde{c}_Z\,.
\end{aligned}
$$

Therefore, we get the following bound on the $\kappa_{3,i}$ term in each summand of (D.29),

$$\frac{1}{6}|\kappa_{3;i}| \;\le\; k^{3/2}\zeta_{i;2}(U_i(\,\bullet\,))\,(\tilde{c}_X + \tilde{c}_Z)\;.$$

Now to bound $\zeta_{i;2}(U_i(\,\bullet\,))$, which involves a third derivative term, recall the chain rule from Lemma D.25,

$$\begin{aligned} D_i^3(h\circ f^{(t)})(\mathbf{u}) \;&= \partial^3 h\big(f^{(t)}(\mathbf{u})\big)\big(D_i f^{(t)}(\mathbf{u})\big)^{\otimes 3} + 3\partial^2 h\big(f^{(t)}(\mathbf{u})\big)\big(D_i f^{(t)}(\mathbf{u})\otimes D_i^2 f(\mathbf{u})\big) \\ &\quad + \partial h\big(f^{(t)}(\mathbf{u})\big) D_i^3 f^{(t)}(\mathbf{u})\;. \end{aligned}$$

Note that for $r = 1,2,3$, $\big\|\partial^r h\big(f^{(t)}(\mathbf{u})\big)\big\| \le \gamma_r(h)$. By Cauchy-Schwarz and triangle inequality, this implies that almost surely

$$\begin{aligned} U_i(\mathbf{w}) \;&\le \Big(\sum_{j_1,j_2,j_3=1}^{k}\Big(\sum_{l_1,l_2,l_3=1}^{d}\Big(\gamma_3(h)\Big\|\frac{\partial f^{(t)}(\mathbf{W}_i(\mathbf{w}))}{\partial x_{ij_1l_1}}\Big\|\Big\|\frac{\partial f^{(t)}(\mathbf{W}_i(\mathbf{w}))}{\partial x_{ij_2l_2}}\Big\|\Big\|\frac{\partial f^{(t)}(\mathbf{W}_i(\mathbf{w}))}{\partial x_{ij_3l_3}}\Big\| \\ &\qquad + 3\gamma_2(h)\Big\|\frac{\partial f^{(t)}(\mathbf{W}_i(\mathbf{w}))}{\partial x_{ij_1l_1}}\Big\|\Big\|\frac{\partial^2 f^{(t)}(\mathbf{W}_i(\mathbf{w}))}{\partial x_{ij_2l_2}\partial x_{ij_3l_3}}\Big\| \\ &\qquad + \gamma_1(h)\Big\|\frac{\partial^3 f^{(t)}(\mathbf{W}_i(\mathbf{w}))}{\partial x_{ij_3l_3}\partial x_{ij_2l_2}\partial x_{ij_3l_3}}\Big\|\Big)\Big)^2\Big)^{1/2} \\ &\le \gamma_3(h)\Big(\textstyle\sum_{j_1,j_2,j_3=1}^{k}\Big(\sum_{l_1=1}^{d}\Big\|\frac{\partial f^{(t)}(\mathbf{W}_i(\mathbf{w}))}{\partial x_{ij_1l_1}}\Big\|\Big)^2\Big(\sum_{l_2=1}^{d}\Big\|\frac{\partial f^{(t)}(\mathbf{W}_i(\mathbf{w}))}{\partial x_{ij_2l_2}}\Big\|\Big)^2 \\ &\qquad \Big(\sum_{l_3=1}^{d}\Big\|\frac{\partial f^{(t)}(\mathbf{W}_i(\mathbf{w}))}{\partial x_{ij_3l_3}}\Big\|\Big)^2\Big)^{1/2} \\ &\quad + 3\gamma_2(h)\Big(\sum_{j_1,j_2,j_3=1}^{k}\Big(\sum_{l_1=1}^{d}\Big\|\frac{\partial f^{(t)}(\mathbf{W}_i(\mathbf{w}))}{\partial x_{ij_1l_1}}\Big\|\Big)^2\Big(\sum_{l_2,l_3=1}^{d}\Big\|\frac{\partial^2 f^{(t)}(\mathbf{W}_i(\mathbf{w}))}{\partial x_{ij_2l_2}\partial x_{ij_3l_3}}\Big\|\Big)^2\Big)^{1/2} \\ &\quad + \gamma_1(h)\Big(\textstyle\sum_{j_1,j_2,j_3=1}^{k}\Big(\sum_{l_1,l_2,l_3=1}^{d}\Big\|\frac{\partial^3 f^{(t)}(\mathbf{W}_i(\mathbf{w}))}{\partial x_{ij_3l_3}\partial x_{ij_2l_2}\partial x_{ij_3l_3}}\Big\|\Big)^2\Big)^{1/2} \\ &= \gamma_3(h)\,T_{1;i}(\mathbf{w})^3 + 3\gamma_2(h)\,T_{1;i}(\mathbf{w})T_{2;i}(\mathbf{w}) + \gamma_1(h)\,T_{3;i}(\mathbf{w})\;, \end{aligned}$$

where for $r = 1,2,3$,

$$\begin{aligned} T_{r;i}(\mathbf{w}) \;&:= \Big(\textstyle\sum_{j_1,\ldots,j_r=1}^{k}\Big(\sum_{l_1,\ldots,l_r=1}^{d}\Big\|\frac{\partial^r f^{(t)}(\mathbf{W}_i(\mathbf{w}))}{\partial x_{ij_1l_1}\ldots\partial x_{ij_rl_r}}\Big\|\Big)^2\Big)^{1/2} \\ &\le \textstyle\sum_{s=1}^{q}\Big(\sum_{j_1,\ldots,j_r=1}^{k}\Big(\sum_{l_1,\ldots,l_r=1}^{d}\Big|\frac{\partial^r f_s^{(t)}(\mathbf{W}_i(\mathbf{w}))}{\partial x_{ij_1l_1}\ldots\partial x_{ij_rl_r}}\Big|\Big)^2\Big)^{1/2} \\ &= \textstyle\sum_{s=1}^{q}\Big(\sum_{j_1,\ldots,j_r=1}^{k}\Big\|\frac{\partial^r f_s^{(t)}(\mathbf{W}_i(\mathbf{w}))}{\partial \mathbf{x}_{ij_1}\ldots\partial \mathbf{x}_{ij_r}}\Big\|\Big)^2\Big)^{1/2}\;. \end{aligned}$$

Therefore, by properties of $\zeta_{i;m}$ and the definition of $\tilde{\alpha}_r^{(t)}$, we get

$$\frac{1}{6}|\kappa_{3,i}| \;\le k^{3/2}\big(\gamma_3(h)(\tilde{\alpha}_1^{(t)})^3 + 3\gamma_2(h)\tilde{\alpha}_1^{(t)}\tilde{\alpha}_2^{(t)} + \gamma_1(h)\tilde{\alpha}_3^{(t)}\big)(\tilde{c}_X + \tilde{c}_Z)\;,$$

which then yields the bound

$$\begin{aligned} &\big|\mathbb{E}h\big(f^{(t)}(\Phi\mathcal{X})\big) - \mathbb{E}h\big(f^{(t)}(\mathcal{Z})\big)\big| \\ &\le\; nk^{3/2}\big(\gamma_3(h)(\tilde{\alpha}_1^{(t)})^3 + 3\gamma_2(h)\tilde{\alpha}_1^{(t)}\tilde{\alpha}_2^{(t)} + \gamma_1(h)\tilde{\alpha}_3^{(t)}\big)(\tilde{c}_X + \tilde{c}_Z)\;. \end{aligned}$$

To bound the approximation error introduced by replacing f with $f^{(t)}$, we apply mean value theorem and Cauchy-Schwarz inequality to obtain

$$\begin{aligned}
&\big|\mathbb{E}h\big(f(\Phi\mathcal{X})\big) - \mathbb{E}h\big(f^{(t)}(\Phi\mathcal{X})\big)\big| \\
&\leq \Big|\mathbb{E}\Big[\sup_{\mathbf{w}\in[f(\Phi\mathcal{X}),f^{(t)}(\Phi\mathcal{X})]} \|\partial h(\mathbf{w})\|\|f(\Phi\mathcal{X}) - f^{(t)}(\Phi\mathcal{X})\|\Big]\Big| \\
&\leq \gamma_1(h)\,\varepsilon(t).
\end{aligned}$$

An analogous argument shows that $\big|\mathbb{E}h\big(f(\mathcal{Z})\big) - \mathbb{E}h\big(f^{(t)}(\mathcal{Z})\big)\big| \leq \gamma_1(h)\,\varepsilon(t)$, and the desired bound is obtained by triangle inequality. □

We now prove Corollary D.9, which is a variant of Lemma D.7(ii) and concerns convergence for a non-smooth plug-in estimate in high dimensions.

Proof of Corollary D.9. Write $f^{(t)}(\mathbf{x}_{11:nk}) \coloneqq g^{(t)}\big(\frac{1}{nk}\sum_{i\leq n,j\leq k}\mathbf{x}_{ij}\big)$. An argument analogous to the proof of Lemma D.7(ii) shows that $\tilde{\alpha}_r^{(t)}$ defined in Theorem D.8 satisfies

$$\tilde{\alpha}_0^{(t)} = \tilde{\nu}_0^{(t)}\,, \quad \tilde{\alpha}_1^{(t)} = \frac{1}{nk^{1/2}}\tilde{\nu}_1^{(t)}\,, \quad \tilde{\alpha}_2^{(t)} = \frac{1}{n^2k}\tilde{\nu}_2^{(t)}\,, \quad \tilde{\alpha}_3^{(t)} = \frac{1}{n^3k^{3/2}}\tilde{\nu}_3^{(t)}\,.$$

Notice that the above substitutions of $\tilde{\alpha}_r^{(t)}$ by $\tilde{\nu}_r^{(t)}$ are of the exact same form of those from the proof of Lemma D.7(ii), and the following bound from the proof of Theorem D.8 is of the exact same form as Theorem 4.1 except that α_r is replaced by $\tilde{\alpha}_r^{(t)}$ and $c_X + c_Z$ is replaced by $\tilde{c}_X + \tilde{c}_Z$):

$$\begin{aligned}
&\big|\mathbb{E}h\big(f^{(t)}(\Phi\mathcal{X})\big) - \mathbb{E}h\big(f^{(t)}(\mathcal{Z})\big)\big| \\
&\leq nk^{3/2}\big(\gamma_3(h)(\tilde{\alpha}_1^{(t)})^3 + 3\gamma_2(h)\tilde{\alpha}_1^{(t)}\tilde{\alpha}_2^{(t)} + \gamma_1(h)\tilde{\alpha}_3^{(t)}\big)(\tilde{c}_X + \tilde{c}_Z)\,.
\end{aligned}$$

Therefore, a repetition of the proof of Lemma D.7(ii) with $\delta = 0$ yields the analogous bounds

$$\begin{aligned}
d_{\mathcal{H}}(\sqrt{n}f^{(t)}(\Phi\mathcal{X}), \sqrt{n}f^{(t)}(\mathcal{Z})) &= O\big((n^{-1/2}(\tilde{\nu}_1^{(t)})^3 + 3n^{-1}\tilde{\nu}_1^{(t)}\tilde{\nu}_2^{(t)} + n^{-3/2}\tilde{\nu}_3^{(t)})(\tilde{c}_X + \tilde{c}_Z)\big)\,, \\
n\|\mathrm{Var}[f^{(t)}(\Phi\mathcal{X})] - \mathrm{Var}[f^{(t)}(\mathcal{Z})]\| &= O\big(n^{-1}(\tilde{\nu}_0^{(t)}\tilde{\nu}_3^{(t)} + \tilde{\nu}_1^{(t)}\tilde{\nu}_2^{(t)})(\tilde{c}_X + \tilde{c}_Z)\big)\,.
\end{aligned}$$

Now by an argument analogous to the proof of Theorem D.8, we can bound the difference between f_t and f in $d_{\mathcal{H}}$ as

$$\begin{aligned}
&d_{\mathcal{H}}(\sqrt{n}f(\Phi\mathcal{X}), \sqrt{n}f^{(t)}(\Phi\mathcal{X})) = \sup_{h\in\mathcal{H}}\big|\mathbb{E}h\big(\sqrt{n}f(\Phi\mathcal{X})\big) - \mathbb{E}h\big(\sqrt{n}f^{(t)}(\Phi\mathcal{X})\big)\big| \\
&\leq \sup_{h\in\mathcal{H}}\Big|\mathbb{E}\Big[\sup_{\mathbf{w}\in[\sqrt{n}f(\Phi\mathcal{X}),\sqrt{n}f^{(t)}(\Phi\mathcal{X})]}\|\partial h(\mathbf{w})\|\sqrt{n}\|f(\Phi\mathcal{X}) - f_t(\Phi\mathcal{X})\|\Big]\Big| \\
&\leq \sqrt{n}\sup_{h\in\mathcal{H}}\gamma_1(h)\varepsilon(t) \leq \sqrt{n}\varepsilon(t)\,.
\end{aligned}$$

Moreover, by the triangle inequality of $\|\bullet\|$, the Jensen's inequality to move $\|\bullet\|$ inside the expectation and the Cauchy-Schwarz inequality,

$$n\|\mathrm{Var}[f(\Phi\mathcal{X})] - \mathrm{Var}[f^{(t)}(\Phi\mathcal{X})]\|$$

$$
\begin{aligned}
&= \tfrac{n}{2}\big\|\mathbb{E}\big[(f(\Phi\mathcal{X}) + f^{(t)}(\Phi\mathcal{X}))(f(\Phi\mathcal{X}) - f^{(t)}(\Phi\mathcal{X}))^\top\big] \\
&\qquad + \mathbb{E}\big[(f(\Phi\mathcal{X}) - f^{(t)}(\Phi\mathcal{X}))(f(\Phi\mathcal{X}) + f^{(t)}(\Phi\mathcal{X}))^\top\big] \\
&\qquad - \mathbb{E}[f(\Phi\mathcal{X}) + f^{(t)}(\Phi\mathcal{X})]\mathbb{E}[f(\Phi\mathcal{X}) - f^{(t)}(\Phi\mathcal{X})]^\top \\
&\qquad - \mathbb{E}[f(\Phi\mathcal{X}) - f^{(t)}(\Phi\mathcal{X})]\mathbb{E}[f(\Phi\mathcal{X}) + f^{(t)}(\Phi\mathcal{X})]^\top\big\| \\
&\le n\big\|\, \|f(\Phi\mathcal{X}) + f^{(t)}(\Phi\mathcal{X})\|\, \big\|_{L_2}\big\|\, \|f(\Phi\mathcal{X}) - f^{(t)}(\Phi\mathcal{X})\|\, \big\|_{L_2} \\
&\quad + n\big\|\, \|f(\Phi\mathcal{X}) + f^{(t)}(\Phi\mathcal{X})\|\, \big\|_{L_1}\big\|\, \|f(\Phi\mathcal{X}) - f^{(t)}(\Phi\mathcal{X})\|\, \big\|_{L_1} \\
&\le 2n\big\|\, \|2f(\Phi\mathcal{X}) + f^{(t)}(\Phi\mathcal{X}) - f(\Phi\mathcal{X})\|\, \big\|_{L_2}\big\|\, \|f(\Phi\mathcal{X}) - f^{(t)}(\Phi\mathcal{X})\|\, \big\|_{L_2} \\
&\le 2n\big(2\|\, \|f(\Phi\mathcal{X})\|\, \|_{L_2} + \varepsilon(t)\big)\varepsilon(t) \;=\; 4n\|\, \|f(\Phi\mathcal{X})\|\, \|_{L_2}\varepsilon(t) + 2n\varepsilon(t)^2\,.
\end{aligned}
$$

The same argument applies to $f(\mathcal{Z})$, and the desired bound is obtained by applying triangle inequalities. □

We now prove Corollary D.10, which is a variant of Lemma D.7(i) and concerns convergence for a plug-in estimate in high dimensions to the first-order Taylor expansion defined in (6.11):

$$
f^T(\mathbf{x}_{11},\dots,\mathbf{x}_{nk}) \coloneqq g(\mathbb{E}[\phi_{11}\mathbf{X}_1]) + \partial g(\mathbb{E}[\phi_{11}\mathbf{X}_1])\big(\tfrac{1}{nk}\textstyle\sum_{i=1}^n\sum_{j=1}^k \mathbf{x}_{ij} - \mathbb{E}[\phi_{11}\mathbf{X}_1]\big)\,.
$$

Proof of Corollary D.10. Similar to the proof of Lemma D.7(i), a Taylor expansion argument followed by Hölder's inequality and noting the definition of $\tilde{\kappa}_r$ and that $\gamma_1(h) = 1$ gives

$$
\begin{aligned}
|\mathbb{E}h(\sqrt{n}f(\Phi\mathcal{X})) - \mathbb{E}h(\sqrt{n}f^T(\Phi\mathcal{X}))| &\le \sqrt{n}\,\gamma_1(h)\,\mathbb{E}\|f(\Phi\mathcal{X}) - f^T(\Phi\mathcal{X})\| \\
&\le \sqrt{n}\,\gamma_1(h)\,\mathbb{E}\big[\sup_{\mathbf{w}\in[\mathbf{0},\bar{\mathbf{X}}]}\|\partial^2 g(\mu+\mathbf{w})\|_1\,(\max_{l\le d}|(\bar{\mathbf{X}})_l|)^2\big] \\
&\le \sqrt{n}\,\gamma_1(h)\,\big\|\sup_{\mathbf{w}\in[\mathbf{0},\bar{\mathbf{X}}]}\|\partial^2 g(\mu+\mathbf{w})\|_1\big\|_{L_3}\,\big\|\,\max_{l\le d}|(\bar{\mathbf{X}})_l|\,\big\|_{L_3}^2 \\
&\le \sqrt{n}\,\tilde{\kappa}_2\,\big\|\,\max_{l\le d}|(\bar{\mathbf{X}})_l|\,\big\|_{L_3}^2\,. \qquad \text{(D.30)}
\end{aligned}
$$

Note that by properties of a maximum and the triangle inequality,

$$
\begin{aligned}
\big\|\,\max_{l\le d}|(\bar{\mathbf{X}})_l|\,\big\|_{L_3} &= \big\|\,\max\{\max_{l\le d}(\bar{\mathbf{X}})_l\,,\,\max_{l\le d}(-\bar{\mathbf{X}})_l\}\,\big\|_{L_3} \\
&\le \big\|\,\max\{|\max_{l\le d}(\bar{\mathbf{X}})_l|\,,\,|\max_{l\le d}(-\bar{\mathbf{X}})_l|\}\,\big\|_{L_3} \\
&\le \big\|\,|\max_{l\le d}(\bar{\mathbf{X}})_l| + |\max_{l\le d}(-\bar{\mathbf{X}})_l|\,\big\|_{L_3} \\
&\le \big\|\max_{l\le d}(\bar{\mathbf{X}})_l\big\|_{L_3} + \big\|\max_{l\le d}(-\bar{\mathbf{X}})_l\big\|_{L_3}\,.
\end{aligned}
$$

$\bar{\mathbf{X}}$ is an average of n i.i.d. zero-mean terms $\frac{1}{k}\sum_j \phi_{ij}\mathbf{X}_i - \mu$, and $-\bar{\mathbf{X}}$ is an average of n i.i.d. zero-mean terms $-\frac{1}{k}\sum_j \phi_{ij}\mathbf{X}_i + \mu$. Therefore, Lemma D.38 applies and we get that there exists a universal constant C_m such that

$$
\Big\|\max_{l\le d}(\bar{\mathbf{X}})_l\Big\|_{L_m}\,,\,\Big\|\max_{l\le d}(-\bar{\mathbf{X}})_l\Big\|_{L_m} \le \inf_{\nu\in\mathbb{R}}\big[2n^{-(1-\nu)} + \log(d)n^{-\nu}\,M_2^2 + n^{-1/2}CM_m\big],
$$

where for $m \le 6$,

$$M_m \;:=\; \big\| \max_{l\le d} |(\phi_{ij}\mathbf{X}_i - \mathbb{E}[\phi_{ij}\mathbf{X}_i])_l| \big\|_{L_m} \le 2 \big\| \max_{l\le d} |(\phi_{ij}\mathbf{X}_i)_l| \big\|_{L_m} \le 2(6\tilde{c}_X)^{1/3} \,.$$

The bound on M_m together with the assumption $\log d = o(n^\alpha)$ for some $\alpha \ge 0$ implies, for $m \le 6$,

$$\begin{aligned}
\big\| \max_{l\le d} |(\bar{\mathbf{X}})_l| \big\|_{L_m} &\le \inf_{\nu\in\mathbb{R}} \big[4n^{-(1-\nu)} + 48\log(d)n^{-\nu}\,(\tilde{c}_X)^{2/3} + 24n^{-1/2}C(\tilde{c}_X)^{1/3}\big] \\
&\le o\big(\inf_{\nu\in\mathbb{R}} \big[4n^{-(1-\nu)} + 48n^{\alpha-\nu}\,(\tilde{c}_X)^{2/3} + 24n^{-1/2}C(\tilde{c}_X)^{1/3}\big]\big) \\
&\le o\big(\max\{1,(\tilde{c}_X)^{2/3}\}\big(4n^{-(1-\alpha)/2} + 48n^{-(1-\alpha)/2} + 24n^{-1/2}C\big)\big) \\
&= o(n^{-(1-\alpha)/2}\max\{1,(\tilde{c}_X)^{2/3}\}) \,, \qquad \text{(D.31)}
\end{aligned}$$

so by substituting into (D.30) and noting that the bounds are independent of $\mathcal{H}$,

$$\begin{aligned}
d_{\mathcal{H}}(\sqrt{n}f^T(\mathcal{Z}), \sqrt{n}f^T(\Phi\mathcal{X})) &= \sup_{h\in\mathcal{H}} |\mathbb{E}h(\sqrt{n}f^T(\mathcal{Z})) - \mathbb{E}h(\sqrt{n}f^T(\Phi\mathcal{X}))| \\
&= o(n^{-1/2+\alpha}\,\tilde{\kappa}_2\,\max\{1,(\tilde{c}_X)^{4/3}\}) \,.
\end{aligned}$$

Now to bound $d_{\mathcal{H}}(\sqrt{n}f(\Phi\mathcal{X}), \sqrt{n}f^T(\Phi\mathcal{X}))$, we seek to apply Theorem D.8. First define $f^{(t)} := f^T$ which is thrice-differentiable with approximation quality $\epsilon(t) = 0$. Note that for $\tilde{\alpha}_r^{(t)}$ defined in terms of $f^{(t)}$,

$$\tilde{\alpha}_1^{(t)} \;=\; n^{-1}k^{-1/2}\|\partial g(\mu)\|_1 \;\le\; n^{-1}k^{-1/2}\kappa_1 \,, \qquad \tilde{\alpha}_2^{(t)} \;=\; \tilde{\alpha}_3^{(t)} \;=\; 0 \,.$$

Notice that the above substitutions of $\tilde{\alpha}_r^{(t)}$ by $\tilde{\kappa}_r$ are of the exact same form of those from the proof of Lemma D.7(i). Then the following bound from the proof of Theorem D.8 applies, which is of the exact same form as Theorem 4.1 except that α_r is replaced by $\tilde{\alpha}_r^{(t)}$ and $c_X + c_Z$ is replaced by $\tilde{c}_X + \tilde{c}_Z$):

$$\begin{aligned}
&\big|\mathbb{E}h\big(f^{(t)}(\Phi\mathcal{X})\big) - \mathbb{E}h\big(f^{(t)}(\mathcal{Z})\big)\big| \\
&\qquad\qquad \le nk^{3/2}\big(\gamma_3(h)(\tilde{\alpha}_1^{(t)})^3 + 3\gamma_2(h)\tilde{\alpha}_1^{(t)}\tilde{\alpha}_2^{(t)} + \gamma_1(h)\tilde{\alpha}_3^{(t)}\big)(\tilde{c}_X + \tilde{c}_Z) \,.
\end{aligned}$$

Therefore, a repetition of the proof of Lemma D.7(i) with $\delta = 0$ yields the analogous bound

$$d_{\mathcal{H}}(\sqrt{n}f^T(\Phi\mathcal{X}), \sqrt{n}f^T(\mathcal{Z})) \;=\; d_{\mathcal{H}}(\sqrt{n}f^{(t)}(\Phi\mathcal{X}), \sqrt{n}f^{(t)}(\mathcal{Z})) \;=\; O\big(n^{-1/2}\tilde{\kappa}_1^3(\tilde{c}_X + \tilde{c}_Z)\big) \,.$$

By the triangle inequality we get the desired bound

$$d_{\mathcal{H}}(\sqrt{n}f(\Phi\mathcal{X}), \sqrt{n}f^T(\mathcal{Z})) \;=\; o\big(n^{-1/2+\alpha}\tilde{\kappa}_2\max\{1,(\tilde{c}_X)^{4/3}\}\big) + O\big(n^{-1/2}\tilde{\kappa}_1^3(\tilde{c}_X + \tilde{c}_Z)\big) \,.$$

For the variance bound, by the proof of Lemma D.7(i) with $\delta = 0$, we get

$$\mathrm{Var}[f^T(\Phi\mathcal{X})] \;=\; \mathrm{Var}[f^T(\mathcal{Z})] \,.$$

Moreover by an analogous argument to the proof of Lemma D.7(i), almost surely there

exists some $\tilde{\mathbf{X}}^{(s)} \in [\mathbf{0}, \bar{\mathbf{X}}]$ that depends on $f_s(\Phi\mathcal{X})_s$ and $f_s^T(\Phi\mathcal{X})$, such that for

$$\mathbf{R}_s^1 \;\coloneqq\; (\partial g_s(\mu))^\top \bar{\mathbf{X}} \;, \qquad\qquad \mathbf{R}_s^2 \;\coloneqq\; \mathrm{Tr}\big((\partial^2 g_s(\mu + \tilde{\mathbf{X}}^{(s)}))^\top \bar{\mathbf{X}}\bar{\mathbf{X}}^\top\big) \;,$$

we have

$$n\big(\mathrm{Var}[f(\Phi\mathcal{X})] - \mathrm{Var}\big[f^T(\Phi\mathcal{X})\big]\big)_{r,s} \;= n\big(\mathrm{Cov}[\mathbf{R}_r^1, \mathbf{R}_s^2] + \mathrm{Cov}[\mathbf{R}_r^2, \mathbf{R}_s^1] + \mathrm{Cov}[\mathbf{R}_r^2, \mathbf{R}_s^2]\big) \;.$$

The only part that differs from the proof of Lemma D.7(i) is how we bound the covariance terms, which is similar to how the bound on (D.30) is obtained:

$$\begin{aligned}
\mathrm{Cov}[\mathbf{R}_r^1, \mathbf{R}_s^2] \;&= \mathbb{E}[\mathbf{R}_r^1 \mathbf{R}_s^2] \;=\; \mathbb{E}\big[(\partial g_r(\mu))^\top \bar{\mathbf{X}}\, \mathrm{Tr}\big((\partial^2 g_s(\mu + \tilde{\mathbf{X}}^{(s)}))^\top \bar{\mathbf{X}}\bar{\mathbf{X}}^\top\big)\big] \\
&\le \|\partial g_r(\mu)\|_1 \mathbb{E}\big[\|\partial^2 g_s(\mu + \tilde{\mathbf{X}}^{(s)})\|_1 \|\max_{l \le d} |(\bar{\mathbf{X}})_l|\|^3\big] \\
&\le \|\partial g_r(\mu)\|_1 \, \big\|\|\partial^2 g_s(\mu + \tilde{\mathbf{X}}^{(s)})\|_1\big\|_{L_4} \, \big\|\max_{l \le d} |(\bar{\mathbf{X}})_l|\big\|_{L_4}^3 \\
&= o\big(n^{-3/2+3\alpha/2} \tilde{\kappa}_1(g_r) \tilde{\kappa}_2(g_s) \max\{1, (\tilde{c}_X)^2\}\big) \;.
\end{aligned}$$

In the last bound, we have specified $\tilde{\kappa}_1(g_r)$ to be the quantity $\tilde{\kappa}_1$ defined in terms of g_r, and used (D.31) to bound $\|\max_{l \le d} |(\bar{\mathbf{X}})_l|\|_{L_4} = o(n^{-(1-\alpha)/2} \max\{1, (\tilde{c}_X)^{2/3}\})$. Analogously

$$\mathrm{Cov}[\mathbf{R}_r^2, \mathbf{R}_s^1] \;=\; o\big(n^{-3/2+3\alpha/2} \tilde{\kappa}_1(g_r) \tilde{\kappa}_2(g_s) \max\{1, (\tilde{c}_X)^2\}\big) \;,$$

and also

$$\begin{aligned}
\mathrm{Cov}[\mathbf{R}_r^2, \mathbf{R}_s^2] \;&\le \big|\mathbb{E}\big[\mathrm{Tr}\big((\partial^2 g_r(\mu + \tilde{\mathbf{X}}^{(r)}))^\top \bar{\mathbf{X}}\bar{\mathbf{X}}^\top\big)\, \mathrm{Tr}\big((\partial^2 g_s(\mu + \tilde{\mathbf{X}}^{(s)}))^\top \bar{\mathbf{X}}\bar{\mathbf{X}}^\top\big)\big]\big| \\
&\quad + \big|\mathbb{E}\big[\mathrm{Tr}\big((\partial^2 g_r(\mu + \tilde{\mathbf{X}}^{(r)}))^\top \bar{\mathbf{X}}\bar{\mathbf{X}}^\top\big)\big]\big| \, \big|\mathbb{E}\big[\mathrm{Tr}\big((\partial^2 g_s(\mu + \tilde{\mathbf{X}}^{(s)}))^\top \bar{\mathbf{X}}\bar{\mathbf{X}}^\top\big)\big]\big| \\
&\le 2\big\|\|\partial^2 g_r(\mu + \tilde{\mathbf{X}}^{(r)})\|_1\big\|_{L_6} \, \big\|\|\partial^2 g_s(\mu + \tilde{\mathbf{X}}^{(s)})\|_1\big\|_{L_6} \, \big\|\max_{l \le d} |(\bar{\mathbf{X}})_l|\big\|_{L_6}^4 \\
&= o\big(n^{-2+2\alpha} \tilde{\kappa}_2(g_r) \tilde{\kappa}_2(g_s) \max\{1, (\tilde{c}_X)^{8/3}\}\big) \;.
\end{aligned}$$

Therefore

$$\begin{aligned}
&n\big(\mathrm{Var}[f(\Phi\mathcal{X})] - \mathrm{Var}\big[f^T(\mathcal{Z})\big]\big)_{r,s} \\
&\qquad = o\big(\big(n^{-1/2+3\alpha/2} \tilde{\kappa}_1(g_r) \tilde{\kappa}_2(g_s) + n^{-1+2\alpha} \tilde{\kappa}_2(g_r) \tilde{\kappa}_2(g_s)\big) \max\{1, (\tilde{c}_X)^{8/3}\}\big) \;.
\end{aligned}$$

Since $\sum_{r,s=1}^q \tilde{\kappa}_{R_1}(g_r) \tilde{\kappa}_{R_2}(g_s) = \tilde{\kappa}_{R_1} \tilde{\kappa}_{R_2}$ where $\tilde{\kappa}_R$ is defined in terms of g, we get

$$n\big\|\mathrm{Var}[f(\Phi\mathcal{X})] - \mathrm{Var}\big[f^T(\mathcal{Z})\big]\big\| \;=\; o\big(\big(n^{-1/2+3\alpha/2} \tilde{\kappa}_1 \tilde{\kappa}_2 + n^{-1+2\alpha} \tilde{\kappa}_2^2\big) \max\{1, (\tilde{c}_X)^{8/3}\}\big) \;,$$

as required. This completes the proof. □

D.5.5. Proofs for Appendix D.1.5

In this section, we first prove Lemma D.12, a toy example showing how repeated augmentation adds additional complexity, and then prove D.11, the main result concerning repeated augmentation.

Proof of Lemma D.12. By the invariance $\phi_1 \mathbf{X}_1 \stackrel{d}{=} \mathbf{X}_1$ and the fact that $\mathbf{X}_1$, $\mathbf{X}_2$, ϕ_1 and ϕ_2 are independent, we get that

$$\mathrm{Var} f_1(\mathbf{X}_1, \mathbf{X}_2) \;=\; \mathrm{Var} f_1(\phi_1 \mathbf{X}_1, \phi_2 \mathbf{X}_2) \;, \quad \mathrm{Var} f_2(\mathbf{X}_1, \mathbf{X}_2) \;=\; \mathrm{Var} f_2(\phi_1 \mathbf{X}_1, \phi_2 \mathbf{X}_2) \;.$$

For repeated augmentation, notice that for any $\mathbf{v} \in \mathbb{R}^d$,

$$\begin{aligned}
\mathbf{v}^\top \mathrm{Var} f_1(\phi_1 \mathbf{X}_1, \phi_1 \mathbf{X}_2) \mathbf{v} \;&= \mathbf{v}^\top \mathrm{Var}[\phi_1 \mathbf{X}_1 + \phi_1 \mathbf{X}_2] \mathbf{v} \\
&= \mathbf{v}^\top \mathrm{Var}[\phi_1 \mathbf{X}_1] \mathbf{v} + \mathbf{v}^\top \mathrm{Var}[\phi_1 \mathbf{X}_2] \mathbf{v} + 2 \mathbf{v}^\top \mathrm{Cov}[\phi_1 \mathbf{X}_1, \phi_1 \mathbf{X}_2] \mathbf{v} \\
&= 2 \mathbf{v}^\top \mathrm{Var}[\mathbf{X}_1] \mathbf{v} + 2 \mathbf{v}^\top \mathrm{Cov}[\phi_1 \mathbf{X}_1, \phi_1 \mathbf{X}_2] \mathbf{v} \\
&= \mathbf{v}^\top \mathrm{Var}[\phi_1 \mathbf{X}_1 + \phi_2 \mathbf{X}_2] \mathbf{v} + 2 \mathbf{v}^\top \mathrm{Cov}[\phi_1 \mathbf{X}_1, \phi_1 \mathbf{X}_2] \mathbf{v} \\
&= \mathbf{v}^\top \mathrm{Var} f_1(\phi_1 \mathbf{X}_1, \phi_2 \mathbf{X}_2) \mathbf{v} + 2 \mathbf{v}^\top \mathrm{Cov}[\phi_1 \mathbf{X}_1, \phi_1 \mathbf{X}_2] \mathbf{v} \;,
\end{aligned}$$

and similarly

$$\mathbf{v}^\top \mathrm{Var} f_2(\phi_1 \mathbf{X}_1, \phi_1 \mathbf{X}_2) \mathbf{v} \;= \mathbf{v}^\top \mathrm{Var} f_2(\phi_1 \mathbf{X}_1, \phi_2 \mathbf{X}_2) \mathbf{v} - 2 \mathbf{v}^\top \mathrm{Cov}[\phi_1 \mathbf{X}_1, \phi_1 \mathbf{X}_2] \mathbf{v} \;.$$

Now note that for all $\mathbf{v} \in \mathbb{R}^d$,

$$\begin{aligned}
\mathbf{v}^\top \mathrm{Cov}[\phi_1 \mathbf{X}_1, \phi_1 \mathbf{X}_2] \mathbf{v} \;&= \mathbb{E}\big[(\mathbf{X}_1^\top \phi_1^\top \mathbf{v})^\top (\mathbf{X}_2^\top \phi_1^\top \mathbf{v})\big] - \mathbb{E}\big[\mathbf{X}_1^\top \phi_1^\top \mathbf{v}\big]^\top \mathbb{E}\big[\mathbf{X}_2^\top \phi_1^\top \mathbf{v}\big] \\
&= \mathbb{E}\big[(\mu^\top \phi_1^\top \mathbf{v})^\top (\mu^\top \phi_1^\top \mathbf{v})\big] - \mathbb{E}\big[\mu^\top \phi_1^\top \mathbf{v}\big]^\top \mathbb{E}\big[\mu^\top \phi_1^\top \mathbf{v}\big] \\
&= \mathrm{Var}[\mathbf{v}^\top \phi_1 \mu] \geq 0 \;,
\end{aligned}$$

and therefore for all $\mathbf{v} \in \mathbb{R}^d$,

$$\begin{aligned}
\mathbf{v}^\top \mathrm{Var} f_1(\phi \mathbf{X}_1, \phi_1 \mathbf{X}_2) \mathbf{v} \;&\geq\; \mathbf{v}^\top \mathrm{Var} f_1(\phi_1 \mathbf{X}_1, \phi_2 \mathbf{X}_2) \mathbf{v} \;, \\
\mathbf{v}^\top \mathrm{Var} f_2(\phi \mathbf{X}_1, \phi_1 \mathbf{X}_2) \mathbf{v} \;&\leq\; \mathbf{v}^\top \mathrm{Var} f_2(\phi_1 \mathbf{X}_1, \phi_2 \mathbf{X}_2) \mathbf{v} \;,
\end{aligned}$$

which completes the proof.

□

The broad stroke idea in proving Theorem D.11 for repeated augmentation is similar to that of our main result, Theorem 4.1, and we refer readers to Appendix D.4 for a proof overview. The only difference is that in proving Theorem 4.1, we can group data into independent blocks due to i.i.d. augmentations being used for different data points. In the proof of Theorem D.11, the strategy must be modified: The additional dependence introduced by reusing transformations means moments can no longer be factored off from derivatives, so stronger assumptions on the derivatives are required to control terms. This is achieved by using the symmetry assumption on f from (D.4).

Proof of Theorem D.11. Similar to the proof for Theorem D.1 (a generalised version of

Theorem 4.1), we abbreviate $g = h \circ f$ and denote

$$\mathbf{V}_i(\bullet) \;\coloneqq\; (\tilde{\Phi}_1\mathbf{X}_1, \ldots, \tilde{\Phi}_{i-1}\mathbf{X}_{i-1}, \bullet, \mathbf{Y}_{i+1}, \ldots, \mathbf{Y}_n)\,.$$

The same telescoping sum and Taylor expansion argument follows, yielding

$$\begin{aligned}\big|\mathbb{E}h(f(\tilde{\Phi}\mathcal{X})) - \mathbb{E}h(f(\mathbf{Y}_1, \ldots, \mathbf{Y}_n))\big| \;&=\; \big|\mathbb{E}\textstyle\sum_{i=1}^n \big[g(\mathbf{V}_i(\tilde{\Phi}_i\mathbf{X}_i)) - g(\mathbf{V}_i(\mathbf{Y}_i))\big]\big| \\ &\le \textstyle\sum_{i=1}^n \big|\mathbb{E}\big[g(\mathbf{V}_i(\tilde{\Phi}_i\mathbf{X}_i)) - g(\mathbf{V}_i(\mathbf{Y}_i))\big]\big|,\end{aligned} \tag{D.32}$$

and each summand is bounded above as

$$\big|\mathbb{E}\big[g(\mathbf{V}_i(\tilde{\Phi}_i\mathbf{X}_i)) - g(\mathbf{V}_i(\mathbf{Y}_i))\big]\big| \;\le\; |\tau_{1,i}| + \frac{1}{2}|\tau_{2,i}| + \frac{1}{6}|\tau_{3,i}|\,,$$

where

$$\begin{aligned}\tau_{1,i} &\coloneqq \mathbb{E}\big[\big(D_i g(\mathbf{V}_i(\mathbf{0}))\big)\big(\tilde{\Phi}_i\mathbf{X}_i - \mathbf{Y}_i\big)\big] \\ \tau_{2,i} &\coloneqq \mathbb{E}\big[\big(D_i^2 g(\mathbf{V}_i(\mathbf{0}))\big)\big((\tilde{\Phi}_i\mathbf{X}_i)(\tilde{\Phi}_i\mathbf{X}_i)^\top - \mathbf{Y}_i\mathbf{Y}_i^\top\big)\big] \\ \tau_{3,i} &\coloneqq \mathbb{E}\big[\|\tilde{\Phi}_i\mathbf{X}_i\|^3 \sup_{\mathbf{w}\in[\mathbf{0},\tilde{\Phi}_i\mathbf{X}_i]} \big\|D_i^3 g\big(\mathbf{V}_i(\mathbf{w})\big)\big\| + \|\mathbf{Y}_i\|^3 \sup_{\mathbf{w}\in[\mathbf{0},\mathbf{Y}_i]} \big\|D_i^3 g\big(\mathbf{V}_i(\mathbf{w})\big)\big\|\big]\,.\end{aligned}$$

With a slight abuse of notation, we view $D_i g(\mathbf{V}_i(\mathbf{0}))$ as a function $\mathbb{R}^{dk} \to \mathbb{R}$ and $D_i^2 g(\mathbf{V}_i(\mathbf{0}))$ as a function $\mathbb{R}^{dk\times dk} \to \mathbb{R}$. Substituting into (D.32), and applying the triangle inequality, shows

$$\big|\mathbb{E}h(f(\tilde{\Phi}\mathcal{X})) - \mathbb{E}h(f(\mathbf{Y}_1, \ldots, \mathbf{Y}_n))\big| \;\le\; \textstyle\sum_{i=1}^n \big(|\tau_{1,i}| + \frac{1}{2}|\tau_{2,i}| + \frac{1}{6}|\tau_{3,i}|\big)\,.$$

The next step is to bound the terms $\tau_{1,i}$, $\tau_{2,i}$ and $\tau_{3,i}$. $\tau_{3,i}$ is analogous to $\kappa_{3,i}$ in the proof of Theorem 4.1. Define

$$M_i \coloneqq \max\big\{\big\|\sup_{\mathbf{w}\in[\mathbf{0},\tilde{\Phi}_i\mathbf{X}_i]}\|D_i^3 g\big(\mathbf{V}_i(\mathbf{w})\big)\|\big\|_{L_2},\; \big\|\sup_{\mathbf{w}\in[\mathbf{0},\mathbf{Y}_i]}\|D_i^3 g\big(\mathbf{V}_i(\mathbf{w})\big)\|\big\|_{L_2}\big\}\,,$$

we can handle $\tau_{3,i}$ in the exact same way as in Theorem 4.1 to obtain

$$\frac{1}{6}|\tau_{3,i}| \;\le\; k^{3/2}(c_X + c_Y)M_i.$$

However, bounding $\tau_{1,i}$ and $\tau_{2,i}$ works differently, since $(\tilde{\Phi}_i\mathbf{X}_i, \mathbf{Y}_i)$ is no longer independent of $(\tilde{\Phi}_j\mathbf{X}_j, \mathbf{Y}_j)_{j\neq i}$ and therefore not independent of $\mathbf{V}_i(\mathbf{0})$. To this end, we invoke the permutation invariance assumption (D.4) on f, which implies the function $g(\mathbf{V}_i(\bullet)) = h(f(\mathbf{V}_i(\bullet)))$ that takes input in $\mathbb{R}^{kd}$ satisfies (D.50) in Lemma D.31. Then Lemma D.31 shows that, for each $i \le n$ and for $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ik} \in \mathbb{R}^d$,

$$\frac{\partial}{\partial\mathbf{x}_{i1}} g(\mathbf{V}_i(\mathbf{0})) \;=\; \ldots \;=\; \frac{\partial}{\partial\mathbf{x}_{ik}} g(\mathbf{V}_i(\mathbf{0})), \tag{D.33}$$

$$\frac{\partial^2}{\partial\mathbf{x}_{i1}^2} g(\mathbf{V}_i(\mathbf{0})) \;=\; \ldots \;=\; \frac{\partial^2}{\partial\mathbf{x}_{ik}^2} g(\mathbf{V}_i(\mathbf{0})), \tag{D.34}$$

$$\frac{\partial^2}{\partial\mathbf{x}_{ir}\partial\mathbf{x}_{is}} g(\mathbf{V}_i(\mathbf{0})) \text{ is the same for all } r \neq s, 1 \le r, s \le k. \tag{D.35}$$

Consider bounding $\tau_{1,i}$. Rewrite $\tau_{1,i}$ as a sum of k terms and denote $\mathbf{Y}_{ij} \in \mathbb{R}^d$ as $Y_{ij1:ijd}$, the subvector of $\mathbf{Y}_i$ analogous to $\phi_j \mathbf{X}_i$ in $\tilde{\Phi}_i \mathbf{X}_i$. Since that (D.33) allows $\frac{\partial}{\partial \mathbf{x}_{i1}} g(\mathbf{V}_i(\mathbf{0}))$ to be taken outside the summation in (a) below, we get

$$
\begin{aligned}
|\tau_{1,i}| \;&=\; \mathbb{E}\big[\textstyle\sum_{j=1}^k \frac{\partial}{\partial \mathbf{x}_{ij}} g(\mathbf{V}_i(\mathbf{0}))\big(\phi_j \mathbf{X}_i - \mathbf{Y}_{ij}\big)\big] \\
&\overset{(a)}{=}\; \mathbb{E}\big[\frac{\partial}{\partial \mathbf{x}_{i1}} g(\mathbf{V}_i(\mathbf{0})) \textstyle\sum_{j=1}^k \big(\phi_j \mathbf{X}_i - \mathbf{Y}_{ij}\big)\big] \\
&\overset{(b)}{=}\; \mathbb{E}\Big[\mathbb{E}\big[\frac{\partial}{\partial \mathbf{x}_{i1}} g(\mathbf{V}_i(\mathbf{0}))\big|\tilde{\Phi}, \Psi\big]\mathbb{E}\big[\textstyle\sum_{j=1}^k \big(\phi_j \mathbf{X}_i - \mathbf{Y}_{ij}\big)\big|\tilde{\Phi}, \Psi\big]\Big] \\
&\leq\; \mathbb{E}\big[\big\|\mathbb{E}\big[\frac{\partial}{\partial \mathbf{x}_{i1}} g(\mathbf{V}_i(\mathbf{0}))\big|\tilde{\Phi}, \Psi\big]\big\| \, \big\|\mathbb{E}\big[\textstyle\sum_{j=1}^k \big(\phi_j \mathbf{X}_i - \mathbf{Y}_{ij}\big)\big|\tilde{\Phi}, \Psi\big]\big\|\big] \\
&\leq\; \Big\| \Big\|\mathbb{E}\big[\frac{\partial}{\partial \mathbf{x}_{i1}} g(\mathbf{V}_i(\mathbf{0}))\big|\tilde{\Phi}, \Psi\big]\Big\| \Big\|_{L_2} \Big\| \Big\|\mathbb{E}\big[\textstyle\sum_{j=1}^k \big(\phi_j \mathbf{X}_i - \mathbf{Y}_{ij}\big)\big|\tilde{\Phi}, \Psi\big]\Big\| \Big\|_{L_2} \\
&=:\; (t1i)\,(t2i)\;.
\end{aligned}
$$

where to get (b), we apply conditional independence conditioning on Φ and Ψ, the augmentations for $\mathcal{X}$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ respectively, and to obtain the final bound we exploited Cauchy-Schwarz inequality. We will first upper bound $(t2i)$ by the trace of the variance of the augmented (X_i). Moving the summation outside the expectation,

$$
\begin{aligned}
(t2i) = \Big\| \Big\|\mathbb{E}\big[\textstyle\sum_{j=1}^k \big(\phi_j \mathbf{X}_i - \mathbf{Y}_{ij}\big)\big|\tilde{\Phi}, \Psi\big]\Big\| \Big\|_{L_2} \\
&= \sqrt{\mathbb{E}\Big[\mathbb{E}\Big[\textstyle\sum_{j_1=1}^k \big(\phi_{j_1} \mathbf{X}_i - \mathbf{Y}_{ij_1}\big)\Big|\tilde{\Phi}, \Psi\Big]^\top \mathbb{E}\Big[\textstyle\sum_{j_2=1}^k \big(\phi_{j_2} \mathbf{X}_i - \mathbf{Y}_{ij_2}\big)\Big|\tilde{\Phi}, \Psi\Big]\Big]} \\
&= \sqrt{\textstyle\sum_{j_1, j_2=1}^k \mathbb{E}\Big[\mathbb{E}\big[\phi_{j_1} \mathbf{X}_i - \mathbf{Y}_{ij_1}\big|\phi_{j_1}, \psi_{j_1}\big]^\top \mathbb{E}\big[\phi_{j_2} \mathbf{X}_i - \mathbf{Y}_{ij_2}\big|\phi_{j_2}, \psi_{j_2}\big]\Big]}.
\end{aligned}
\tag{D.36}
$$

In each summand, the expectation is taken over a product of two quantities, which are respectively functions of $\{\phi_{j_1}, \psi_{j_1}\}$ and $\{\phi_{j_2}, \psi_{j_2}\}$. For $j_1 \neq j_2$, the two quantities are independent, and are also zero-mean since

$$
\begin{aligned}
\mathbb{E}\Big[\mathbb{E}\big[\phi_j \mathbf{X}_i - (\mathbf{Y}_i)_j\big|\phi_j, \psi_j\big]\Big] \;&=\; \mathbb{E}\big[\mathbb{E}[\phi_j \mathbf{X}_i|\phi_j]\big] - \mathbb{E}\big[\mathbb{E}[\mathbf{Y}_{ij}|\psi_j]\big] \\
&=\; \mathbb{E}\big[\mathbb{E}[\phi_j \mathbf{X}_i|\phi_j]\big] - \mathbb{E}\big[\mathbb{E}[\psi_j \mathbf{X}_1|\psi_j]\big] = \mathbf{0}.
\end{aligned}
$$

Therefore, summands with $j_1 \neq j_2$ vanish, and (D.36) becomes

$$
\begin{aligned}
(t2i) \;&=\; \Big\| \Big\|\mathbb{E}\big[\textstyle\sum_{j=1}^k \big(\phi_j \mathbf{X}_i - \mathbf{Y}_{ij}\big)\big|\tilde{\Phi}, \Psi\big]\Big\| \Big\|_{L_2} \\
&=\; \sqrt{\textstyle\sum_{j=1}^k \mathbb{E}\Big[\mathbb{E}\big[\phi_j \mathbf{X}_i - \mathbf{Y}_{ij}\big|\phi_j, \psi_j\big]^\top \mathbb{E}\big[\phi_j \mathbf{X}_i - \mathbf{Y}_{ij}\big|\phi_j, \psi_j\big]\Big]} \\
&\overset{(c)}{=}\; \sqrt{k}\sqrt{\mathbb{E}\Big[\mathbb{E}\big[\phi_1 \mathbf{X}_i - \mathbf{Y}_{i1}\big|\phi_1, \psi_1\big]^\top \mathbb{E}\big[\phi_1 \mathbf{X}_i - \mathbf{Y}_{i1}\big|\phi_1, \psi_1\big]\Big]} \\
&=\; \sqrt{k}\sqrt{\mathrm{TrVar}\big[\mathbb{E}[\phi_1 \mathbf{X}_i|\phi_1] - \mathbb{E}[\mathbf{Y}_{i1}|\psi_1]\big]} \\
&\overset{(d)}{=}\; \sqrt{k}\sqrt{\mathrm{TrVar}\big[\mathbb{E}[\phi_1 \mathbf{X}_1|\phi_1] - \mathbb{E}[\psi_1 \mathbf{X}_1|\psi_1]\big]}
\end{aligned}
$$

$$\overset{(e)}{=} \sqrt{2k}\sqrt{\mathrm{TrVar}\big[\mathbb{E}[\phi_1 \mathbf{X}_1|\phi_1]\big]} \;=\; \sqrt{k}m_1.$$

where we have used that $(\phi_1, \psi_1), \ldots, (\phi_k, \psi_k)$ are i.i.d. in (c) and that $\mathbb{E}[\phi_1 \mathbf{X}_1|\phi_1]$ and $\mathbb{E}[\psi_1 \mathbf{X}_1|\psi_1]$ are i.i.d. in (d) and (e). Define

$$C_i := \Big\| \Big\| \mathbb{E}\big[\frac{\partial}{\partial x_{i11:i1d}} g(\mathbf{V}_i(\mathbf{0}))\big|\tilde{\Phi}, \Psi\big] \Big\| \Big\|_{L_2},$$

we note that $(t1i) \leq C_i$. Therefore we obtain

$$|\tau_{1,i}| \;\leq\; \sqrt{k}m_1 C_i \;.$$

$\tau_{2,i}$ can be bounded similarly by rewriting as a sum of k^2 terms and making use of conditional independence. We defer the detailed computation to Lemma D.29. Define

$$E_i := \Big\| \big\| \mathbb{E}\big[\frac{\partial^2}{\partial \mathbf{x}_{i1}^2} g(\mathbf{V}_i(\mathbf{0}))\big|\tilde{\Phi}, \Psi\big] \big\| \Big\|_{L_2}, \quad F_i := \Big\| \big\| \mathbb{E}\big[\frac{\partial^2}{\partial \mathbf{x}_{i1} \partial \mathbf{x}_{i2}} g(\mathbf{V}_i(\mathbf{0}))\big|\tilde{\Phi}, \Psi\big] \big\| \Big\|_{L_2}. \tag{D.37}$$

Lemma D.29 below shows that

$$\frac{1}{2}|\tau_{2,i}| \;\leq\; k^{1/2} m_2 E_i + k^{3/2} m_3 F_i. \tag{D.38}$$

In summary, the right hand side of (D.32) is hence bounded by

$$\begin{aligned}(\text{D.32}) &\leq \textstyle\sum_{i=1}^n |\tau_{1,i}| + \frac{1}{2}|\tau_{2,i}| + \frac{1}{6}|\tau_{1,i}| \\ &\leq nk^{-1/2} m_1 \max_{i \leq n} C_i + k^{1/2} m_2 \max_{i \leq n} E_i + k^{3/2} m_3 \max_{i \leq n} F_i + nk^{3/2}(c_2 + c_3) \max_{i \leq n} M_i.\end{aligned}$$

Lemma D.30 below shows that the the maximums $\max_{i \leq n} E_i, \max_{i \leq n} C_i, \max_{i \leq n} D_i$ $\max_{i \leq n} M_i$ are in turn bounded by

$$\max_{i \leq n} C_i \;\leq\; k^{-1/2} \gamma_1(h) \alpha_1, \tag{D.39}$$

$$\max_{i \leq n} E_i \;\leq\; k^{-1/2}(\gamma_2(h)\alpha_1^2 + \gamma_1(h)\alpha_2), \tag{D.40}$$

$$\max_{i \leq n} F_i \;\leq\; k^{-3/2}(\gamma_2(h)\alpha_1^2 + \gamma_1(h)\alpha_2), \tag{D.41}$$

$$\max_{i \leq n} M_i \;\leq\; \lambda_h(n, k). \tag{D.42}$$

That yields the desired upper bound on (D.32),

$$\begin{aligned}&\big|\mathbb{E}h(f(\tilde{\Phi}\mathcal{X})) - \mathbb{E}h(f(\mathbf{Y}_1, \ldots, \mathbf{Y}_n))\big| \\ &\qquad \leq n\gamma_1(h)\alpha_1 m_1 + n\omega_2(n,k)(\gamma_2(h)\alpha_1^2 + \gamma_1(h)\alpha_2) + nk^{3/2}\lambda_h(n,k)(c_X + c_Y)\;.\end{aligned}$$

which finishes the proof. □

We complete the computation of bounds in Lemma D.29 and Lemma D.30.

Lemma D.29. *The bound on* $|\tau_{2,i}|$ *in* (D.38) *holds.*

Proof. Rewrite $\tau_{2,i}$ as a sum of k^2 terms,

$$|\tau_{2,i}| = \mathbb{E}\Big[\sum_{j_1,j_2=1}^{k} \frac{\partial^2}{\partial \mathbf{x}_{ij_1}\partial \mathbf{x}_{ij_2}} g(\mathbf{V}_i(\mathbf{0}))\big((\phi_{j_1}\mathbf{X}_i)(\phi_{j_2}\mathbf{X}_i)^\top - (\mathbf{Y}_{ij_1})(\mathbf{Y}_{ij_2})^\top\big)\Big]. \quad \text{(D.43)}$$

Consider the terms with $j_1 = j_2$. (D.34) says that the derivatives are the same for $j_1 = 1, \ldots, k$ and allows $\frac{\partial^2}{\partial x^2_{ij1:ijd}} g(\mathbf{V}_i(\mathbf{0}))$ to be taken out of the following sum,

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{j=1}^{k} \frac{\partial^2}{\partial \mathbf{x}_{ij}^2} g(\mathbf{V}_i(\mathbf{0}))\big((\phi_j\mathbf{X}_i)(\phi_j\mathbf{X}_i)^\top - (\mathbf{Y}_{ij})(\mathbf{Y}_{ij})^\top\big)\Big] \\
&= \mathbb{E}\Big[\frac{\partial^2}{\partial \mathbf{x}_{i1}^2} g(\mathbf{V}_i(\mathbf{0}))\sum_{j=1}^{k}\big((\phi_j\mathbf{X}_i)(\phi_j\mathbf{X}_i)^\top - (\mathbf{Y}_{ij})(\mathbf{Y}_{ij})^\top\big)\Big] \\
&\overset{(a)}{=} \mathbb{E}\Big[\mathbb{E}\big[\frac{\partial^2}{\partial \mathbf{x}_{i1}^2} g(\mathbf{V}_i(\mathbf{0}))\big|\tilde{\Phi}, \Psi\big]\mathbb{E}\big[\sum_{j=1}^{k}\big((\phi_j\mathbf{X}_i)(\phi_j\mathbf{X}_i)^\top - (\mathbf{Y}_{ij})(\mathbf{Y}_{ij})^\top\big)\big|\tilde{\Phi}, \Psi\big]\Big] \\
&\leq \Big\| \big\|\mathbb{E}\big[\frac{\partial^2}{\partial \mathbf{x}_{i1}^2} g(\mathbf{V}_i(\mathbf{0}))\big|\tilde{\Phi}, \Psi\big]\big\|\Big\|_{L_2} \big\|\sum_{j=1}^{k}\|\mathbf{T}_{jj}\|\big\|_{L_2} \\
&= E_i \big\|\sum_{j=1}^{k}\|\mathbf{T}_{jj}\|\big\|_{L_2}, \qquad \text{(D.44)}
\end{aligned}$$

where we have used conditional independence conditioning on $\tilde{\Phi}$ and Ψ in (a), defined E_i as in (D.37) and denoted

$$\begin{aligned}
\mathbf{T}_{j_1j_2} &:= \mathbb{E}\big[(\phi_{j_1}\mathbf{X}_i)(\phi_{j_2}\mathbf{X}_i)^\top - (\mathbf{Y}_{ij_1})(\mathbf{Y}_{ij_2})^\top\big|\tilde{\Phi}, \Psi\big] \\
&= \mathbb{E}\big[(\phi_{j_1}\mathbf{X}_i)(\phi_{j_2}\mathbf{X}_i)^\top|\phi_{j_1}, \phi_{j_2}\big] - \mathbb{E}\big[(\mathbf{Y}_{ij_1})(\mathbf{Y}_{ij_2})^\top\big|\psi_{j_1}, \psi_{j_2}\big] \\
&= \mathbb{E}\big[(\phi_{j_1}\mathbf{X}_1)(\phi_{j_2}\mathbf{X}_1)^\top|\phi_{j_1}, \phi_{j_2}\big] - \mathbb{E}\big[(\psi_{j_1}\mathbf{X}_1)(\psi_{j_2}\mathbf{X}_1)^\top\big|\psi_{j_1}, \psi_{j_2}\big]\ .
\end{aligned}$$

Consider the terms in (D.43) with $j_1 \neq j_2$. (D.35) says that the derivatives are the same for $1 \leq j_1, j_2 \leq k$ with $j_1 \neq j_2$, so by a similar argument,

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{j_1\neq j_2} \frac{\partial^2}{\partial \mathbf{x}_{ij_1}\partial \mathbf{x}_{ij_2}} g(\mathbf{V}_i(\mathbf{0}))\big((\phi_{j_1}\mathbf{X}_i)(\phi_{j_2}\mathbf{X}_i)^\top - (\mathbf{Y}_{ij_1})(\mathbf{Y}_{ij_2})^\top\big)\Big] \\
&= \mathbb{E}\Big[\frac{\partial^2}{\partial \mathbf{x}_{i1}\partial \mathbf{x}_{i2}} g(\mathbf{V}_i(\mathbf{0}))\sum_{j_1\neq j_2}\big((\phi_{j_1}\mathbf{X}_i)(\phi_{j_2}\mathbf{X}_i)^\top - (\mathbf{Y}_{ij_1})(\mathbf{Y}_{ij_2})^\top\big)\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\big[\frac{\partial^2}{\partial \mathbf{x}_{i1}\partial \mathbf{x}_{i2}} g(\mathbf{V}_i(\mathbf{0}))\big|\tilde{\Phi}, \Psi\big]\mathbb{E}\big[\sum_{j_1\neq j_2}\big((\phi_{j_1}\mathbf{X}_i)(\phi_{j_2}\mathbf{X}_i)^\top - (\mathbf{Y}_{ij_1})(\mathbf{Y}_{ij_2})^\top\big)\big|\tilde{\Phi}, \Psi\big]\Big] \\
&\leq \Big\| \big\|\mathbb{E}\big[\frac{\partial^2}{\partial x_{i11:i1d}\partial x_{i21:i2d}} g(\mathbf{V}_i(\mathbf{0}))\big|\tilde{\Phi}, \Psi\big]\big\|\Big\|_{L_2} \big\|\sum_{j_1\neq j_2}\|\mathbf{T}_{j_1j_2}\|\big\|_{L_2} \\
&= F_i \big\|\sum_{j_1\neq j_2}\|\mathbf{T}_{j_1j_2}\|\big\|_{L_2}\ , \qquad \text{(D.45)}
\end{aligned}$$

where we have used F_i defined in (D.37). To obtain a bound for (D.44)and (D.45), we need to bound $\big\|\sum_{j=1}^{k}\|\mathbf{T}_{jj}\|\big\|_{L_2}$ and $\big\|\sum_{j_1\neq j_2}\|\mathbf{T}_{j_1j_2}\|\big\|_{L_2}$. To this end, we denote

$$\begin{aligned}
\mathbf{A}_\phi &:= \mathrm{vec}\big(\mathbb{E}\big[(\phi_1\mathbf{X}_1)(\phi_1\mathbf{X}_1)^\top|\phi_1\big]\big), & \mathbf{A}_\psi &:= \mathrm{vec}\big(\mathbb{E}\big[(\psi_1\mathbf{X}_1)(\psi_1\mathbf{X}_1)^\top\big|\psi_1\big]\big), \\
\mathbf{B}_\phi &:= \mathrm{vec}\big(\mathbb{E}\big[(\phi_1\mathbf{X}_1)(\phi_2\mathbf{X}_1)^\top|\phi_1, \phi_2\big]\big), & \mathbf{B}_\psi &:= \mathrm{vec}\big(\mathbb{E}\big[(\psi_1\mathbf{X}_1)(\psi_2\mathbf{X}_1)^\top\big|\psi_1, \psi_2\big]\big),
\end{aligned}$$

where $\mathrm{vec}(\{M_{rs}\}_{i,j\leq d}) = (M_{11}, M_{12}, \ldots, M_{dd}) \in \mathbb{R}^{d^2}$ converts a matrix to its vector representation. Then WLOG we can write $\mathbf{T}_{11} = \mathbf{A}_\phi - \mathbf{A}_\psi$, $\mathbf{T}_{12} = \mathbf{B}_\phi - \mathbf{B}_\psi$. Before we

proceed, we compute several useful quantities in terms of $\mathbf{T}$'s. Recall that

$$m_2 \coloneqq \sqrt{\sum_{r,s\le d} \frac{\mathrm{Var}\mathbb{E}\big[(\phi_1\mathbf{X}_1)_r(\phi_1\mathbf{X}_1)_s\big|\phi_1\big]}{2}},$$
$$m_3 \coloneqq \sqrt{\sum_{r,s\le d} 12\mathrm{Var}\mathbb{E}\big[(\phi_1\mathbf{X}_1)_r(\phi_2\mathbf{X}_1)_s\big|\phi_1,\phi_2\big]}\,.$$

Since $\mathbf{A}_\phi$ and $\mathbf{A}_\psi$ are i.i.d.,

$$\begin{aligned}\mathbb{E}\big[\big\|\mathbf{T}_{jj}\big\|^2\big] &= \mathbb{E}\big[\big\|\mathbf{T}_{11}\big\|^2\big] = \mathbb{E}\big[\mathrm{Tr}(\mathbf{T}_{11}\mathbf{T}_{11}^\top)\big] = \mathrm{Tr}\mathbb{E}\big[\mathbf{T}_{11}\mathbf{T}_{11}^\top\big]\\ &= \mathrm{Tr}(\mathbb{E}[\mathbf{A}_\phi\mathbf{A}_\phi^\top] - \mathbb{E}[\mathbf{A}_\phi\mathbf{A}_\psi^\top] - \mathbb{E}[\mathbf{A}_\psi\mathbf{A}_\phi^\top] + \mathbb{E}[\mathbf{A}_\psi\mathbf{A}_\psi^\top])\\ &= 2\mathrm{Tr}\big(\mathbb{E}[\mathbf{A}_\phi\mathbf{A}_\phi^\top] - \mathbb{E}[\mathbf{A}_\phi]\mathbb{E}[\mathbf{A}_\phi]^\top\big)\\ &= 2\sum_{r,s=1}^d \big(\mathbb{E}[(\mathbf{A}_\phi)_{rs}^2] - \mathbb{E}[(\mathbf{A}_\phi)_{rs}]^2\big)\\ &= 2\sum_{r,s=1}^d \mathrm{Var}\mathbb{E}\big[(\phi_1\mathbf{X}_1)_r(\phi_1\mathbf{X}_1)_s\big|\phi_1\big] = 4(m_2)^2. \end{aligned} \tag{D.46}$$

Similarly by noting that $\mathbf{B}_\phi$ and $\mathbf{B}_\psi$ are i.i.d., for $j_1 \neq j_2$,

$$\begin{aligned}\mathbb{E}\big[\big\|\mathbf{T}_{j_1j_2}\big\|^2\big] &= \mathbb{E}\big[\big\|\mathbf{T}_{12}\big\|^2\big] = \mathrm{Tr}\mathbb{E}\big[\mathbf{T}_{12}\mathbf{T}_{12}^\top\big]\\ &= \mathrm{Tr}(\mathbb{E}[\mathbf{B}_\phi\mathbf{B}_\phi^\top] - \mathbb{E}[\mathbf{B}_\phi\mathbf{B}_\psi^\top] - \mathbb{E}[\mathbf{B}_\psi\mathbf{B}_\phi^\top] + \mathbb{E}[\mathbf{B}_\psi\mathbf{B}_\psi^\top])\\ &= 2\sum_{r,s=1}^d \big(\mathbb{E}[(\mathbf{B}_\phi)_{rs}^2] - \mathbb{E}[(\mathbf{B}_\phi)_{rs}]^2\big)\\ &= 2\sum_{r,s=1}^d \mathrm{Var}\mathbb{E}\big[(\phi_1\mathbf{X}_1)_r(\phi_2\mathbf{X}_1)_s\big|\phi_1\big] = \frac{1}{6}(m_3)^2. \end{aligned} \tag{D.47}$$

On the other hand, by Cauchy-Schwarz with respect to the Frobenius inner product, for $j_1 \neq j_2$ and $l_1 \neq l_2$,

$$\begin{aligned}\big|\mathbb{E}[\mathrm{Tr}(\mathbf{T}_{j_1j_2}\mathbf{T}_{l_1l_2}^\top)]\big| &\le \Big|\mathbb{E}\Big[\sqrt{\mathrm{Tr}(\mathbf{T}_{j_1j_2}\mathbf{T}_{j_1j_2}^\top)}\sqrt{\mathrm{Tr}(\mathbf{T}_{l_1l_2}\mathbf{T}_{l_1l_2}^\top)}\Big]\Big|\\ &\le \sqrt{\mathbb{E}\mathrm{Tr}(\mathbf{T}_{j_1j_2}\mathbf{T}_{j_1j_2}^\top)}\sqrt{\mathbb{E}\mathrm{Tr}(\mathbf{T}_{l_1l_2}\mathbf{T}_{l_1l_2}^\top)}\\ &= \sqrt{\mathbb{E}\big[\big\|\mathbf{T}_{j_1j_2}\big\|^2\big]}\sqrt{\mathbb{E}\big[\big\|\mathbf{T}_{l_1l_2}\big\|^2\big]} \le \frac{1}{6}(m_3)^2\,, \end{aligned} \tag{D.48}$$

which can be computed using the above relations for each $j_1, j_2, l_1, l_2 \le k$. Moreover we note that, since $\mathbb{E}[\mathbf{A}_\phi] = \mathbb{E}[\mathbf{A}_\psi]$ and $\mathbb{E}[\mathbf{B}_\phi] = \mathbb{E}[\mathbf{B}_\psi]$ this directly implies that $\mathbb{E}[\mathbf{T}_{11}] = \mathbb{E}[\mathbf{T}_{12}] = 0$. We are now ready to bound $\big\|\sum_{j=1}^k \|\mathbf{T}_{jj}\|\big\|_{L_2}$ and $\big\|\sum_{j_1,j_2=1}^k \|\mathbf{T}_{j_1j_2}\|\big\|_{L_2}$:

$$\begin{aligned}\Big\|\,\|\textstyle\sum_{j=1}^k \mathbf{T}_{jj}\|\,\Big\|_{L_2} &\coloneqq \sqrt{\mathbb{E}\Big[\mathrm{Tr}\big(\big(\sum_{j_1=1}^k \mathbf{T}_{j_1j_1}\big)\big(\sum_{j_2=1}^k \mathbf{T}_{j_2j_2}\big)^\top\big)\Big]}\\ &= \sqrt{\sum_{j_1,j_2=1}^k \mathrm{Tr}\mathbb{E}[\mathbf{T}_{j_1j_1}\mathbf{T}_{j_2j_2}^\top]} \overset{(a)}{=} \sqrt{\sum_{j=1}^k \mathrm{Tr}\mathbb{E}[\mathbf{T}_{j_1j_1}\mathbf{T}_{j_1j_1}^\top]} \overset{(b)}{=} 2\sqrt{k}m_2, \end{aligned}$$

where (a) uses the independence of $\mathbf{T}_{j_1,j_1}$ and $\mathbf{T}_{j_2,j_2}$, and (b) uses (D.46). On the other hand,

$$\Big\|\,\|\textstyle\sum_{j_1\neq j_2} \mathbf{T}_{j_1j_2}\|\,\Big\|_{L_2} \coloneqq \sqrt{\sum_{j_1\neq j_2, l_1\neq l_2} \mathrm{Tr}\mathbb{E}[\mathbf{T}_{j_1j_2}\mathbf{T}_{l_1l_2}^\top]}. \tag{D.49}$$

Consider each summand in (D.49). If j_1, j_2, l_1, l_2 are all distinct, the summand vanishes

since $\mathbf{T}_{j_1 j_2}$ and $\mathbf{T}_{l_1 l_2}$ are independent and zero-mean. Otherwise, we can use (D.48) and (D.47) to upper bound each summand by $\frac{1}{6}(m_3)^2$. The number of non-zero terms is $k^4 - k(k-1)(k-2)(k-3) = 6k^3 - 11k^2 + 6k \le 6k^3 + 6k \le 12k^3$, so (D.49) can be upper bounded by $2k^{3/2} m_3$. In summary,

$$\frac{1}{2}|\tau_{2,i}| \;\le\; \frac{1}{2}E_i \big\| \textstyle\sum_{j=1}^k \|\mathbf{T}_{jj}\| \big\|_{L_2} + \frac{1}{2}F_i \big\| \sum_{j_1 \neq j_2} \|\mathbf{T}_{j_1 j_2}\| \big\|_{L_2} \;\le\; k^{1/2} m_2 E_i + k^{3/2} m_3 F_i \;,$$

which finishes the proof. □

Lemma D.30. *The bounds* (D.39)*,* (D.40)*,* (D.41) *and* (D.42) *hold.*

Proof. The argument is mostly the same as Lemma D.27, except that we use the permutation invariance assumption (D.4) and Lemma D.31 to handle C_i, E_i and F_i. To obtain (D.39), note that the vector norm $\|\bullet\|$ is a convex function, so by Jensen's inequality,

$$\begin{aligned} C_i &= \Big\| \Big\| \mathbb{E}\big[\frac{\partial}{\partial \mathbf{x}_{i1}} g(\mathbf{V}_i(\mathbf{0})) \big| \tilde{\Phi}, \Psi \big] \Big\| \Big\|_{L_2} \le \Big\| \mathbb{E}\Big[\Big\| \frac{\partial}{\partial \mathbf{x}_{i1}} g(\mathbf{V}_i(\mathbf{0})) \Big\| \Big| \tilde{\Phi}, \Psi \Big] \Big\|_{L_2} \\ &= \Big\| \Big\| \frac{\partial}{\partial \mathbf{x}_{i1}} g(\mathbf{V}_i(\mathbf{0})) \Big\| \Big\|_{L_2} \overset{(a)}{=} k^{-1/2} \Big\| \Big\| D_i g(\mathbf{V}_i(\mathbf{0})) \Big\| \Big\|_{L_2}. \end{aligned}$$

In the last equality (a), we have invoked the permutation invariance assumption on f and Lemma D.31, which implies that

$$\sqrt{\mathbb{E}\Big\| \frac{\partial}{\partial \mathbf{x}_{i1}} g(\mathbf{V}_i(\mathbf{0})) \Big\|^2} = \sqrt{\mathbb{E}\Big(\frac{1}{k} \textstyle\sum_{j=1}^k \Big\| \frac{\partial}{\partial \mathbf{x}_{ij}} g(\mathbf{V}_i(\mathbf{0})) \Big\|^2 \Big)} = k^{-1/2} \sqrt{\mathbb{E}\| D_i g(\mathbf{V}_i(\mathbf{0})) \|}.$$

This allows us to apply a similar argument to that in Lemma D.27. By chain rule, almost surely, $D_i g(\mathbf{V}_i(\mathbf{0})) = \partial h\big(f(\mathbf{V}_i(\mathbf{0}))\big)\big(D_i f(\mathbf{V}_i(\mathbf{0}))\big)$. For a random function $\mathbf{T} : \mathbb{R}^{dk} \to \mathbb{R}_0^+$ and $m \in \mathbb{N}$, define

$$\zeta'_{i;m}(\mathbf{T}) := \max\Big\{ \big\| \sup_{\mathbf{w} \in [\mathbf{0}, \tilde{\Phi}_1 \mathbf{X}_i]} \mathbf{T}(\mathbf{w}) \big\|_{L_m}, \big\| \sup_{\mathbf{w} \in [\mathbf{0}, \mathbf{Z}_i]} \mathbf{T}(\mathbf{w}) \big\|_{L_m} \Big\},$$

which is analogous to the definition of $\zeta_{i;m}$ in Lemma D.20 and satisfies all the properties in Lemma D.20. Then

$$\begin{aligned} &\| \, \|D_i g(\mathbf{V}_i(\mathbf{0}))\| \, \|_{L_2} \overset{(a)}{\le} \zeta'_{i;2}\big(\big\| D_i g(\mathbf{V}_i(\bullet)) \big\| \big) \\ &\le \zeta'_{i;2}\big(\big\| \partial h\big(f(\mathbf{V}_i(\bullet))\big) \big\| \big\| D_i f(\mathbf{V}_i(\bullet)) \big\| \big) \;\le\; \zeta'_{i;2}\big(\gamma_1(h) \big\| D_i f(\mathbf{V}_i(\bullet)) \big\| \big) \\ &\overset{(b)}{\le} \gamma_1(h) \zeta'_{i;2}(\| D_i f(\mathbf{V}_i(\bullet)) \|) \;\le\; \gamma_1(h) \alpha_1 \;. \end{aligned}$$

where we have used Lemma D.20 for (a) and (b). Therefore we obtain the bound (D.39) as

$$\max_{i \le n} C_i \le \max_{i \le n} k^{-1/2} \, \| \, \|D_i g(\mathbf{V}_i(\mathbf{0}))\| \, \|_{L_2} \le k^{-1/2} \gamma_1(h) \alpha_1.$$

To obtain (D.40) for the second partial derivatives, we use Jensen's inequality and Lemma D.31 again to get

$$E_i \;=\; \Big\| \Big\| \mathbb{E}\big[\frac{\partial^2}{\partial \mathbf{x}_{i1}^2} g(\mathbf{V}_i(\mathbf{0})) \big| \tilde{\Phi}, \Psi \big] \Big\| \Big\|_{L_2}$$

$$\leq \Big\| \Big\| \frac{\partial^2}{\partial \mathbf{x}_{i1}^2} g(\mathbf{V}_i(\mathbf{0})) \Big\| \Big\|_{L_2} = \sqrt{\frac{1}{k} \sum_{j=1}^{k} \Big\| \frac{\partial^2}{\partial \mathbf{x}_{ij}^2} g(\mathbf{V}_i(\mathbf{0})) \Big\|^2}$$

$$\leq \frac{1}{\sqrt{k}} \sqrt{\sum_{j_1, j_2 = 1}^{k} \Big\| \frac{\partial^2}{\partial \mathbf{x}_{ij_1} \partial \mathbf{x}_{ij_2}} g(\mathbf{V}_i(\mathbf{0})) \Big\|^2} = \frac{1}{\sqrt{k}} \| \, \| D_i^2 g(\mathbf{V}_i(\mathbf{0})) \| \, \|_{L_2}.$$

By the same argument for the mixed the derivatives, in (D.41),

$$F_i = \Big\| \, \| \mathbb{E}\big[\frac{\partial^2}{\partial \mathbf{x}_{i1} \partial \mathbf{x}_{i2}} g(\mathbf{V}_i(\mathbf{0})) \big| \tilde{\Phi}, \Psi \big] \| \Big\|_{L_2} \leq \Big\| \, \| \frac{\partial^2}{\partial \mathbf{x}_{i1} \partial \mathbf{x}_{i2}} g(\mathbf{V}_i(\mathbf{0})) \| \Big\|_{L_2}$$
$$\leq \frac{1}{\sqrt{k(k-1)}} \Big\| \, \| D_i^2 g(\mathbf{V}_i(\mathbf{0})) \| \, \|_{L_2} .$$

$\| \, \| D_i^2 g(\mathbf{V}_i(\mathbf{0})) \| \, \|_{L_2}$ is bounded similarly in Lemma D.27 except that we are bounding an L_2 norm instead of an L_1 norm.

$$\begin{aligned}
&\big\| \|D_i^2 g(\mathbf{V}_i(\mathbf{0}))\| \big\|_{L_2} \big\| \\
&\leq \zeta_2'\big(\big\| D_i^2 g(\mathbf{V}_i(\bullet)) \big\| \big) \\
&\overset{(a)}{\leq} \zeta_2'\big(\big\| \partial^2 h\big(f(\mathbf{V}_i(\bullet))\big) \big\| \big\| D_i f(\mathbf{V}_i(\bullet)) \big\|^2 + \big\| \partial h\big(f(\mathbf{V}_i(\bullet))\big) \big\| \big\| D_i^2 f(\mathbf{V}_i(\bullet)) \big\| \big) \\
&\leq \zeta_2'\Big(\gamma_2(h) \|D_i f(\mathbf{V}_i(\bullet))\|^2 + \gamma_1(h) \|D_i^2 f(\mathbf{V}_i(\bullet))\| \Big) \\
&\overset{(b)}{\leq} \gamma_2(h) \, \zeta_4'(\|D_i f(\mathbf{W}_i(\bullet))\|)^2 + \gamma_1(h) \, \zeta_2'(\|D_i^2 f(\mathbf{W}_i(\bullet))\|) \\
&\leq \gamma_2(h) \alpha_1^2 + \gamma_1(h) \alpha_2 \, ,
\end{aligned}$$

where we used Lemma D.25 to obtain (a) and Lemma D.20 to get (b). Therefore, the bounds (D.40) and (D.41) are obtained as

$$\max_{i \leq n} E_i \leq k^{-1/2} (\gamma_2(h) \alpha_1^2 + \gamma_1(h) \alpha_2) \, , \quad \max_{i \leq n} F_i \leq k^{-3/2} (\gamma_2(h) \alpha_1^2 + \gamma_1(h) \alpha_2) \, .$$

Finally for (D.42), recall that

$$M_i := \max \big\{ \big\| \sup_{\mathbf{w} \in [\mathbf{0}, \tilde{\Phi}_i \mathbf{X}_i]} \| D_i^3 g\big(\mathbf{V}_i(\mathbf{w})\big) \| \big\|_{L_2}, \, \big\| \sup_{\mathbf{w} \in [\mathbf{0}, \mathbf{Y}_i]} \| D_i^3 g\big(\mathbf{V}_i(\mathbf{w})\big) \| \big\|_{L_2} \big\} \, ,$$

and notice that it is the same quantity as M_i from Lemma D.27 except that $\mathbf{W}_i$ is replaced by $\mathbf{V}_i$, $\Phi_i \mathbf{X}_i$ is replaced by $\tilde{\Phi}_i \mathbf{X}_i$ and $\mathbf{Z}_i$ is replaced by $\mathbf{Y}_i$. The same argument applies to give

$$\max_{i \leq n} M_i \leq \lambda_h(n, k) \, ,$$

which completes the proof. □

Finally we present the following lemma that describes properties of derivatives of a function satisfying permutation invariance condition:

Lemma D.31. *Suppose $f \in \mathcal{F}(\mathbb{R}^{kd}, \mathbb{R}^q)$ is a function that satisfies the permutation invariance assumption*

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_k) = f(\mathbf{x}_{\pi(1)}, \ldots, \mathbf{x}_{\pi(k)}) \tag{D.50}$$

for any permutation π of k elements. Then at $\mathbf{0} \in \mathbb{R}^{kd}$, the derivatives of f satisfy, for $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^d$,

(i) $\frac{\partial}{\partial \mathbf{x}_1} f(\mathbf{0}) = \ldots = \frac{\partial}{\partial \mathbf{x}_d} f(\mathbf{0})$,
(ii) $\frac{\partial^2}{\partial \mathbf{x}_1^2} f(\mathbf{0}) = \ldots = \frac{\partial^2}{\partial \mathbf{x}_k^2} f(\mathbf{0})$,
(iii) $\frac{\partial^2}{\partial \mathbf{x}_r \partial \mathbf{x}_s} f(\mathbf{0})$ *is the same for* $r \neq s$, $1 \leq r, s \leq k$.

Proof. For $j \leq k, l \leq d$, denote $\mathbf{e}_{jl}$ as the $\big((j-1)d + l\big)^{\text{th}}$ basis vector in $\mathbb{R}^{kd}$ and x_{jl} as the lth coordinate of $\mathbf{x}_d$. Without loss of generality we can set $q = 1$, because it suffices to prove the results coordinate-wise over the q coordinates.. Consider $\frac{\partial}{\partial \mathbf{x}_j} f(\mathbf{0})$, which exists by assumption and can be written as

$$\frac{\partial}{\partial \mathbf{x}_j} f(\mathbf{0}) = \Big(\frac{\partial}{\partial x_{j1}} f(\mathbf{0}), \ldots, \frac{\partial}{\partial x_{jd}} f(\mathbf{0}) \Big)^\top.$$

For each $l \leq d$, the one-dimensional derivative is defined as

$$\frac{\partial}{\partial x_{jl}} f(\mathbf{0}) \;\coloneqq\; \lim_{\epsilon \to 0} \frac{f(\epsilon \mathbf{e}_{jl}) - f(\mathbf{0})}{\epsilon} \;\overset{(a)}{=}\; \lim_{\epsilon \to 0} \frac{f(\epsilon \mathbf{e}_{1l}) - f(\mathbf{0})}{\epsilon} \;=\; \frac{\partial}{\partial x_{1l}} f(\mathbf{0}).$$

In (a) above, we have used the permutation invariance assumption (D.4) across $j \leq k$. This implies $\frac{\partial}{\partial \mathbf{x}_1} f(\mathbf{0}) = \ldots = \frac{\partial}{\partial \mathbf{x}_k} f(\mathbf{0})$ as required. The second derivative $\frac{\partial^2}{\partial \mathbf{x}_j^2} f(\mathbf{0})$ is a $\mathbb{R}^{d \times d}$ matrix with the $(l_1, l_2)^{\text{th}}$ coordinate given by $\frac{\partial^2}{\partial x_{jl_1} \partial x_{jl_2}} f(\mathbf{0})$, which is in turn defined by

$$\begin{aligned}
\frac{\partial^2}{\partial x_{jl_1} \partial x_{jl_2}} f(\mathbf{0}) \;&\coloneqq\; \lim_{\delta \to 0} \frac{\frac{\partial}{\partial x_{jl_1}} f(\delta \mathbf{e}_{jl_2}) - \frac{\partial}{\partial x_{jl_1}} f(\mathbf{0})}{\delta} \\
&\overset{(b)}{=}\; \lim_{\delta \to 0} \lim_{\epsilon \to 0} \frac{\big(f(\delta \mathbf{e}_{jl_2} + \epsilon \mathbf{e}_{jl_1}) - f(\delta \mathbf{e}_{jl_2})\big) - \big(f(\epsilon \mathbf{e}_{jl_1}) - f(\mathbf{0})\big)}{\epsilon \delta} \\
&\overset{(c)}{=}\; \lim_{\delta \to 0} \lim_{\epsilon \to 0} \frac{\big(f(\delta \mathbf{e}_{1l_2} + \epsilon \mathbf{e}_{1l_1}) - f(\delta \mathbf{e}_{1l_2})\big) - \big(f(\epsilon \mathbf{e}_{1l_1}) - f(\mathbf{0})\big)}{\epsilon \delta} \\
&=\; \frac{\partial^2}{\partial x_{1l_2} \partial x_{1l_1}} f(\mathbf{0}).
\end{aligned}$$

We have used the definition for the first derivatives in (b) and assumption (D.4) in (c). This implies, as before, $\frac{\partial^2}{\partial \mathbf{x}_1^2} f(\mathbf{0}) = \ldots = \frac{\partial^2}{\partial \mathbf{x}_k^2} f(\mathbf{0})$. For the mixed derivatives, notice that assumption (D.4) implies, for $r \neq s$, $1 \leq r, s, \leq k$ and $1 \leq l_1, l_2 \leq d$,

$$f(\delta \mathbf{e}_{rl_2} + \epsilon \mathbf{e}_{sl_1}) \;=\; f(\delta \mathbf{e}_{1l_2} + \epsilon \mathbf{e}_{2l_1}),$$

by considering a permutation that brings (r, s) to $(1, 2)$. Therefore, by an analogous argument,

$$\begin{aligned}
\frac{\partial^2}{\partial x_{rl_1} \partial x_{sl_2}} f(\mathbf{0}) \;&\coloneqq\; \lim_{\delta \to 0} \frac{\frac{\partial}{\partial x_{rl_1}} f(\delta \mathbf{e}_{sl_2}) - \frac{\partial}{\partial x_{rl_1}} f(\mathbf{0})}{\delta} \\
&=\; \lim_{\delta \to 0} \lim_{\epsilon \to 0} \frac{\big(f(\delta \mathbf{e}_{sl_2} + \epsilon \mathbf{e}_{rl_1}) - f(\delta \mathbf{e}_{sl_2})\big) - \big(f(\epsilon \mathbf{e}_{rl_1}) - f(\mathbf{0})\big)}{\epsilon \delta} \\
&=\; \lim_{\delta \to 0} \lim_{\epsilon \to 0} \frac{\big(f(\delta \mathbf{e}_{1l_2} + \epsilon \mathbf{e}_{2l_1}) - f(\delta \mathbf{e}_{1l_2})\big) - \big(f(\epsilon \mathbf{e}_{2l_1}) - f(\mathbf{0})\big)}{\epsilon \delta}
\end{aligned}$$

$$= \frac{\partial^2}{\partial x_{1l_2} \partial x_{1l_1}} f(\mathbf{0}).$$

This implies $\frac{\partial^2}{\partial \mathbf{x}_r \partial \mathbf{x}_s} f(\mathbf{0})$ is the same for $r \neq s$, $1 \leq r, s \leq k$. □

D.6 Derivation of examples

Different versions of Gaussian surrogates are used throughout the computation in this section. For clarity, we denote $\mathbf{x}_{11:nk} \coloneqq \{\mathbf{x}_{11}, \ldots, \mathbf{x}_{nk}\}$ and define

$$\begin{aligned} \mathbf{W}(\bullet) &\coloneqq (\Phi_1 \mathbf{X}_1, \ldots, \Phi_{i-1} \mathbf{X}_{i-1}, \bullet, \mathbf{Z}_{i+1}, \ldots, \mathbf{Z}_n), \\ \tilde{\mathbf{W}}(\bullet) &\coloneqq (\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_1, \bullet, \tilde{\mathbf{Z}}_{i+1}, \ldots, \tilde{\mathbf{Z}}_n), \end{aligned}$$

where:

- $\Phi_1 \mathbf{X}_1, \ldots, \Phi_n \mathbf{X}_n \in \mathcal{D}^k$ are the augmented data vectors and $\mathbf{Z}_1, \ldots, \mathbf{Z}_n \in \mathcal{D}^k$ are the i.i.d. surrogate vectors, both defined in Theorem 6.1 (corresponding to $\mathbf{Z}_i^\delta$ defined with $\delta = 0$ in Theorem D.1);
- $\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_n \in \mathcal{D}^k$ are the unaugmented data vectors (k-replicate of original data) whereas the surrogate vectors are denoted $\tilde{\mathbf{Z}}_1, \ldots, \tilde{\mathbf{Z}}_n \in \mathcal{D}^k$, both defined in (6.7).

As before, we write $\Phi\mathcal{X} = \{\Phi_1 \mathbf{X}_1, \ldots, \Phi_n \mathbf{X}_n\}$, $\mathcal{Z} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$, $\tilde{\mathcal{X}} = \{\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_n\}$ and $\tilde{\mathcal{Z}} = \{\tilde{\mathbf{Z}}_1, \ldots, \tilde{\mathbf{Z}}_n\}$. In the case $\mathcal{Z}$ and $\tilde{\mathcal{Z}}$ are Gaussian, existence of $\mathcal{Z}$ and $\tilde{\mathcal{Z}}$ is automatic when $\mathbf{Z}_i$ and $\tilde{\mathbf{Z}}_i$ are allowed to take values in $\mathbb{R}^d$ and the only constraints are their respective mean and variance conditions (6.1) and (6.6). Therefore, we omit existence proof for all examples except for the special case of ridge regression in Appendix D.6.3. Finally, for functions $f : \mathcal{D}^{nk} \to \mathbb{R}^q$ and $g : \mathcal{D} \to \mathbb{R}^q$, and for any $s \leq q$, we use $f_s : \mathcal{D}^{nk} \to \mathbb{R}$ and $g_s : \mathcal{D} \to \mathbb{R}$ to denote the s-th coordinate of f and g respectively.

D.6.1. Empirical averages

In this section, we first prove Proposition 6.7 by verifying that for the empirical average, the bounds in Lemma D.2 and D.3 decay, and by computing the relevant variances and confidence intervals.

Proof of Proposition 6.7. We first apply Lemma D.3 to compare the distance in d_H of $f(\Phi\mathcal{X})$ to $f(\mathcal{Z})$. To do so, we need to compute the noise stability terms for $f(\mathbf{x}_{11:nk}) = \frac{1}{nk} \sum_{i=1}^n \sum_{j=1}^k \mathbf{x}_{ij}$. We first compute the derivatives: for any $\mathbf{v} \in \mathbb{R}^{dk}$, almost surely,

$$D_i f(\mathbf{W}_i(\mathbf{v})) = \frac{1}{nk} (\mathbf{I}_d, \ldots, \mathbf{I}_d)^\top \in \mathbb{R}^{dk \times d}, \qquad \text{and} \qquad D_i^2 f(\mathbf{W}_i(\mathbf{v})) = \mathbf{0}.$$

Then, for all $m \in \mathbb{N}$ we have

$$\alpha_{1;m} \coloneqq \sum_{s\le d} \max_{i\le n} \max\Big\{ \Big\| \sup_{\mathbf{w}\in[\mathbf{0},\Phi_i\mathbf{X}_i]} \|D_i f_s(\mathbf{W}_i(\mathbf{w}))\| \Big\|_{L_m}, \Big\| \sup_{\mathbf{w}\in[\mathbf{0},\mathbf{Z}_i]} \|D_i f_s(\mathbf{W}_i(\mathbf{w}))\| \Big\|_{L_m} \Big\}$$
$$= \sum_{s\le d} \frac{1}{nk} \big\| (\mathbf{I}_d, \dots, \mathbf{I}_d)^\top \mathbf{e}_s \big\| \;=\; \frac{d}{nk^{1/2}},$$

and the noise stability terms associated with higher derivatives are $\alpha_{2;m} = \alpha_{3;m} = 0$. Since d is fixed and $\phi_{11}\mathbf{X}_1$ and $\mathbf{Z}_1$ have bounded 4th moments, we get

$$c_X \;=\; \frac{1}{6}\sqrt{\mathbb{E}\|\phi_{11}\mathbf{X}_1\|^6} = O(1)\,, \quad c_Z \;=\; \frac{1}{6}\sqrt{\mathbb{E}\Big[\Big(\frac{1}{k}\sum\nolimits_{j\le k, s\le d} |Z_{1js}|^2\Big)^3\Big]} \;=\; O(1)\,.$$

Therefore, the bounds in Lemma D.3 (concerning weak convergence) with δ set to 0 become, respectively,

$$(nk)^{3/2}(n(\alpha_{1;6})^3 + 3n^{1/2}\alpha_{1;4}\alpha_{2;4} + \alpha_{3;2})(c_X + c_Z) \;=\; O(n^{-1/2})\,. \tag{D.51}$$

Note that while the above calculation uses $\Phi_i\mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i$ in the case of augmentation, the same calculation holds for $\tilde{\mathbf{X}}_i, \tilde{\mathbf{Z}}_i, \tilde{\mathbf{W}}_i$ in the case of no augmentation. Therefore, (D.51) and Lemma D.3 lead to the required convergence in (i) that as $n \to \infty$,

$$d_{\mathcal{H}}(\sqrt{n}f(\Phi\mathcal{X}), \sqrt{n}f(\mathcal{Z})) \xrightarrow{d} 0\,, \qquad d_{\mathcal{H}}(\sqrt{n}f(\tilde{\mathcal{X}}), \sqrt{n}f(\tilde{\mathcal{Z}})) \xrightarrow{d} 0\,.$$

To prove the statements on variances and confidence intervals, we first note that the equality in variance can be directly obtained by noting that moments of $\mathbf{Z}_i$ match moments of $\Phi_i\mathbf{X}_i$, which implies

$$\mathrm{Var} f(\Phi\mathcal{X}) \;=\; \frac{1}{n}\mathrm{Var}\Big[\frac{1}{k}\sum\nolimits_{j=1}^k \phi_{1j}\mathbf{X}_1\Big] \;=\; \frac{1}{n}\mathrm{Var}\Big[\frac{1}{k}\sum\nolimits_{j=1}^k \mathbf{Z}_{1j}\Big] \;=\; \mathrm{Var} f(\mathcal{Z})\,.$$

The same argument implies $\mathrm{Var} f(\tilde{\mathcal{X}}) = \mathrm{Var} f(\tilde{\mathcal{Z}})$. The next step is to obtain the formula for variances and asymptotic confidence intervals. Since $\mathbf{Z}_i$ is Gaussian in $\mathbb{R}^{dk}$ with mean $\mathbf{1}_{k\times 1} \otimes \mu$ and variance $\mathbf{I}_k \otimes \mathrm{Var}[\phi_{11}\mathbf{X}_1] + (\mathbf{1}_{k\times k} - \mathbf{I}_k) \otimes \mathrm{Cov}[\phi_{11}\mathbf{X}_1, \phi_{12}\mathbf{X}_1]$, we have

$$\frac{1}{k}\sum\nolimits_{j=1}^k \mathbf{Z}_{ij} = \frac{1}{k}\underbrace{(\mathbf{I}_d \dots \mathbf{I}_d)}_{k \text{ copies of } \mathbf{I}_d}\mathbf{Z}_i \sim \mathcal{N}\big(\mathbb{E}[\phi_{11}\mathbf{X}_1], V\big)\,.$$

where

$$V \coloneqq \frac{1}{k}\mathrm{Var}[\phi_{11}\mathbf{X}_1] + \frac{k-1}{k}\mathrm{Cov}[\phi_{11}\mathbf{X}_1, \phi_{12}\mathbf{X}_1].$$

We also remark that as the Gaussian vectors $(\mathbf{Z}_1, \dots, \mathbf{Z}_n)$ are independent, the empirical averages $\frac{1}{k}\sum_{j=1}^k \mathbf{Z}_{1j}, \dots, \frac{1}{k}\sum_{j=1}^k \mathbf{Z}_{nj}$ are also independent. This directly implies that

$$f(\mathcal{Z}) \;=\; \frac{1}{nk}\sum\nolimits_{i=1}^n\sum\nolimits_{j=1}^k \mathbf{Z}_{ij} \sim \mathcal{N}\big(\mathbb{E}[\phi_{11}\mathbf{X}_1], \frac{1}{n}V\big)\,. \tag{D.52}$$

This gives the desired variance for $f(\mathcal{Z})$. On the other hand, since each $\tilde{\mathbf{Z}}_i$ is a Gaussian in $\mathbb{R}^{dk}$ with mean $\mathbf{1}_{k\times 1} \otimes \mathbb{E}[\mathbf{X}_1]$ and variance $\mathbf{1}_{k\times k} \otimes \mathrm{Var}[\mathbf{X}_1]$, it can be viewed as a k-replicate of a Gaussian vector $\tilde{\mathbf{V}}_i$ in $\mathbb{R}^d$ with mean $\mathbb{E}[\mathbf{X}_1]$ and $\mathrm{Var}[\mathbf{X}_1]$. By independence

of $\tilde{\mathbf{Z}}_i$'s, $\mathbf{V}_i$'s are also independent and therefore

$$f(\tilde{\mathcal{Z}}) \;=\; \tfrac{1}{nk}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{k}\tilde{\mathbf{Z}}_{i1} \;\overset{d}{=}\; \tfrac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i \sim \mathcal{N}\big(\mathbb{E}[\mathbf{X}_1], \tfrac{1}{n}\mathrm{Var}[\mathbf{X}_1]\big)\;, \tag{D.53}$$

giving the variance expression for $f(\tilde{\mathcal{Z}})$. Finally, for $d = 1$, the normal distributions given in (D.52) and (D.53) imply that the lower and upper $\alpha/2$-th quantiles for $f(\mathcal{Z})$ and $f(\tilde{\mathcal{Z}})$ are given respectively as

$$\begin{aligned}
&\mathbb{E}[\phi_{11}\mathbf{X}_1] \pm \tfrac{1}{\sqrt{n}}z_{\alpha/2}\sqrt{V} \;=\; \mathbb{E}[\phi_{11}\mathbf{X}_1] \pm \tfrac{1}{\sqrt{\vartheta(f)^2 n}}z_{\alpha/2}\sqrt{\mathrm{Var}[\mathbf{X}_1]}\;,\\
&\mathbb{E}[\phi_{11}\mathbf{X}_1] \pm \tfrac{1}{\sqrt{n}}z_{\alpha/2}\sqrt{\mathrm{Var}[\mathbf{X}_1]}\;.
\end{aligned}$$

These quantiles are asymptotically valid for $f(\Phi\mathcal{X})$ and $f(\tilde{\mathcal{X}})$ respectively since convergence in $d_{\mathcal{H}}$ implies convergence in distribution by Lemma 6.3, which finishes the proof. □

D.6.2. Exponential of negative chi-squared statistic

In this section, we prove Proposition D.13 for the one-dimensional statistic defined in (6.13):

$$f(x_{11},\ldots,x_{nk}) \coloneqq \exp\big(-\big(\tfrac{1}{\sqrt{nk}}\textstyle\sum_{i\le n}\sum_{j\le k}x_{ij}\big)^2\big)\;.$$

We also state a 2d generalisation of this statistic used in our simulation and prove an analogous lemma that justifies convergences and analytical formula for its confidence regions.

Proof of Proposition D.13. For convergence in $d_{\mathcal{H}}$ and variance, define

$$g(x) \coloneqq \tfrac{1}{\sqrt{n}}\exp(-nx^2) \text{ and } \tilde{f}(x_{11:nk}) \coloneqq g\big(\tfrac{1}{nk}\textstyle\sum_{i\le n, j\le k}x_{ij}\big)\;.$$

Then, the required statistic in (6.13) satisfies $f(x_{11:nk}) = \sqrt{n}\tilde{f}(x_{11:nk})$, and applying Lemma D.7(ii) with δ set to 0 to $\tilde{f}$ and g will recover the convergences

$$\begin{aligned}
d_{\mathcal{H}}(\sqrt{n}\tilde{f}(\Phi\mathcal{X}), \sqrt{n}\tilde{f}(\mathcal{Z})) \;&=\; d_{\mathcal{H}}(f(\Phi\mathcal{X}), f(\mathcal{Z}))\;,\\
n(\mathrm{Var}[\tilde{f}(\Phi\mathcal{X})] - \mathrm{Var}[\tilde{f}(\mathcal{Z})]) \;&=\; \mathrm{Var}[f(\Phi\mathcal{X})] - \mathrm{Var}[f(\mathcal{Z})]\;.
\end{aligned}$$

It now suffices to compute the noise stability terms $\nu_{r;m}(g)$ used in Lemma D.7(ii) defined for g. The derivatives for g can be bounded by

$$\begin{aligned}
\partial g(x) \;&=\; -2n^{1/2}x\exp(-nx^2)\;,\quad \partial^2 g(x) = -2n^{1/2}\exp(-nx^2) + 4n^{3/2}x^2\exp(-nx^2),\\
\partial^3 g(x) \;&=\; 12n^{3/2}x\exp(-nx^2) - 8n^{5/2}x^3\exp(-nx^2)\;.
\end{aligned}$$

Note that $\exp(-nx^2) \in [0,1]$ for all $x \in \mathbb{R}$, so only x, x^2 and x^3 play a role in the bound for $\nu_{1;m}$. The noise stability terms can now be bounded by

$$\nu_{1;m} \;=\; \max_{i\le n}\max\Big\{\big\|\sup_{\mathbf{w}\in[\mathbf{0},\Phi_i\mathbf{X}_i]}|\partial g(\overline{\mathbf{W}}_i(\mathbf{w}))|\big\|_{L_m},\; \big\|\sup_{\mathbf{w}\in[\mathbf{0},\mathbf{Z}_i]}|\partial g(\overline{\mathbf{W}}_i(\mathbf{w}))|\big\|_{L_m}\Big\}$$

$$\leq\ 2n^{1/2}\max_{i\leq n}\max\left\{\left\|\sup_{\mathbf{w}\in[\mathbf{0},\Phi_i\mathbf{X}_i]}|\overline{\mathbf{W}}_i(\mathbf{w})|\right\|_{L_m},\ \left\|\sup_{\mathbf{w}\in[\mathbf{0},\mathbf{Z}_i]}|\overline{\mathbf{W}}_i(\mathbf{w})|\right\|_{L_m}\right\}$$
$$\leq\ 2n^{1/2}\max_{i\leq n}\left\|\sup_{\mathbf{w}\in[\mathbf{0},\Phi_i\mathbf{X}_i]\cup[\mathbf{0},\mathbf{Z}_i]}|\overline{\mathbf{W}}_i(\mathbf{w})|\right\|_{L_m}. \tag{D.54}$$

We need to bound the absolute value of $\overline{\mathbf{W}}_i(\mathbf{w})$. Define $\mathbf{A}_{i'} \coloneqq \sum_{j=1}^k \phi_{i'j}\mathbf{X}_{i'}$ and $\mathbf{B}_{i'} \coloneqq \sum_{j=1}^k \mathbf{Z}_{i'j}$, and write $\mathcal{I} \coloneqq [\mathbf{0}, \Phi_i\mathbf{X}_i] \cup [\mathbf{0}, \mathbf{Z}_i]$. Then by triangle inequality,

$$\left\|\sup_{\mathbf{w}\in\mathcal{I}}|\overline{\mathbf{W}}_i(\mathbf{w})|\right\|_{L_m} \tag{D.55}$$
$$=\frac{1}{nk}\left\|\sup_{\mathbf{w}\in\mathcal{I}}\left|\sum_{i'=1}^{i-1}\sum_{j=1}^k\phi_{i'j}\mathbf{X}_{i'}+\sum_{j=1}^k\mathbf{w}_j+\sum_{i'=i+1}^n\sum_{j=1}^k\mathbf{Z}_{i'j}\right|\right\|_{L_m}$$
$$=\frac{1}{nk}\left\|\sup_{\mathbf{w}\in\mathcal{I}}\left|\sum_{i'=1}^{i-1}\mathbf{A}_{i'}+\sum_{j=1}^k\mathbf{w}_j+\sum_{i'=i+1}^n\mathbf{B}_{i'}\right|\right\|_{L_m}$$
$$\leq\frac{1}{nk}\left\|\left|\sum_{i'=1}^{i-1}\mathbf{A}_{i'}\right|+\max\{|\mathbf{A}_i|,|\mathbf{B}_i|\}+\left|\sum_{i'=i+1}^n\mathbf{B}_{i'}\right|\right\|_{L_m}. \tag{D.56}$$

Note that $\mathbf{A}_1,\ldots,\mathbf{A}_{i-1}$ are i.i.d. random variables with zero mean and finite 12th moments by assumption. Also, for $m \leq 12$, by triangle inequality,

$$\|\mathbf{A}_{i'}\|_{L_m} \leq \sum_{j\leq k}\|\phi_{i'j}\mathbf{X}_i\|_{L_m}\ =\ O(k)\,.$$

Rosenthal's inequality from Lemma D.21 implies, for $m \leq 12$, there exists a constant K_m depending only on m such that

$$\left\|\sum_{i'<i}\mathbf{A}_{i'}\right\|_{L_m}\ \leq\ K_m\max\left\{i^{1/m}\|\mathbf{A}_1\|_{L_m}, i^{1/2}\|\mathbf{A}_1\|_{L_2}\right\}\ =\ O(n^{1/2}k)\,.$$

The exact same argument applies to $\mathbf{B}_{i+1},\ldots,\mathbf{B}_n$, implying that

$$\|\mathbf{B}_i\|_{L_m}\ =\ O(k)\,,\qquad\qquad \left\|\sum_{i'>i}\mathbf{B}_{i'}\right\|_{L_m}\ =\ O(n^{1/2}k)\,.$$

Substituting these results into (D.56) gives the following control on $\overline{\mathbf{W}}_i(\mathbf{w})$:

$$\left\|\sup_{\mathbf{w}\in\mathcal{I}}\left|\overline{\mathbf{W}}_i(\mathbf{w})\right|\right\|_{L_m}\ =\ O(n^{-1/2})\,,$$

and finally substituting the bound into (D.54) gives, for $m \leq 12$,

$$\nu_{1;m}\ =\ O(1)\,.$$

The arguments for $\nu_{2;m}$ and $\nu_{3;m}$ are similar, except that $\nu_{2;m}$ involves x^2 and $\nu_{3;m}$ involves x^3. $\nu_{2;m}$ then requires bounding terms of the form

$$\left\|\sup_{\mathbf{w}\in\mathcal{I}}\left|\overline{\mathbf{W}}_i(\mathbf{w})\right|^2\right\|_{L_m}\leq\frac{1}{n^2k^2}\left\|\left(\left|\sum_{i'=1}^{i-1}\mathbf{A}_{i'}\right|+\max\{|\mathbf{A}_i|,|\mathbf{B}_i|\}+\left|\sum_{i'=i+1}^n\mathbf{B}_{i'}\right|\right)^2\right\|_{L_m}$$
$$=\frac{1}{n^2k^2}\left\|\left|\sum_{i'=1}^{i-1}\mathbf{A}_{i'}\right|+\max\{|\mathbf{A}_i|,|\mathbf{B}_i|\}+\left|\sum_{i'=i+1}^n\mathbf{B}_{i'}\right|\right\|_{L_{2m}}^2=O(n^{-1})\,,$$

where the argument proceeds as before but now hold only for $m \leq 6$. $\nu_{3;m}$ similarly requires controlling

$$\left\|\sup_{\mathbf{w}\in\mathcal{I}}\left|\overline{\mathbf{W}}_i(\mathbf{w})\right|^3\right\|_{L_m}\leq\frac{1}{n^3k^3}\left\|\left|\sum_{i'=1}^{i-1}\mathbf{A}_{i'}\right|+\max\{|\mathbf{A}_i|,|\mathbf{B}_i|\}+\left|\sum_{i'=i+1}^n\mathbf{B}_{i'}\right|\right\|_{L_{3m}}^3$$

$$= O(n^{-3/2}) \,,$$

which holds now for $m \leq 4$. Therefore,

$$\begin{aligned} \nu_{2;m} &= O(n^{1/2} + n^{3/2} \times n^{-1}) = O(n^{1/2}) && \text{for } m \leq 6 \,, \\ \nu_{3;m} &= O(n^{3/2} \times n^{-1/2} + n^{5/2} \times n^{-3/2}) = O(n) && \text{for } m \leq 4 \,. \end{aligned}$$

Note also that the moment terms $c_X = O(1)$ by assumption and $c_Z = O(1)$ since the 4th moment of a Gaussian random variable with finite mean and variance is bounded. Moreover, $g(x) = \frac{1}{\sqrt{n}} \exp(-nx^2) \in [0, n^{-1/2}]$ and therefore $\nu_{0;m} = O(n^{-1/2})$ for all $m \in \mathbb{N}$. The two bounds in Lemma D.7(ii) then become:

$$\begin{aligned} \big(n^{-1/2}\nu_{1;6}^3 + n^{-1}\nu_{1;4}\nu_{2;4} + n^{-3/2}\nu_{3;2}\big)(c_X + c_Z) &= O(n^{-1/2}) \,, \\ n^{-1}(\nu_{0;4}(g)\nu_{3;4}(g) + \nu_{1;4}(g)\nu_{2;4}(g))(c_X + c_Z) &= O(n^{-1/2}) \,, \end{aligned}$$

both of which go to zero as $n \to \infty$. Applying Lemma D.7(ii) to $\tilde{f}$ then gives the desired convergences that

$$\begin{aligned} f(\Phi\mathcal{X}) - f(\mathcal{Z}) &= \sqrt{n}(\tilde{f}(\Phi\mathcal{X}) - \tilde{f}(\mathcal{Z})) \xrightarrow{d} 0 \,, \\ \mathrm{Var}[f(\Phi\mathcal{X})] - \mathrm{Var}[f(\mathcal{Z})] &= n(\mathrm{Var}[\tilde{f}(\Phi\mathcal{X})] - \mathrm{Var}[\tilde{f}(\mathcal{Z})]) \xrightarrow{d} 0 \,. \end{aligned}$$

The exact same argument works for $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Z}}$ by setting ϕ_{ij} to identity almost surely and by invoking boundedness of 8th moments of $\mathbf{X}_i$ and $\mathbb{E}[\mathbf{X}_i] = 0$. Therefore, the same convergences hold with $(\Phi\mathcal{X}, \mathcal{Z})$ above replaced by $(\tilde{\mathcal{X}}, \tilde{\mathcal{Z}})$.

Next, we prove the formulas for variance and quantiles. Recall the function $V(s) \coloneqq (1+4s^2)^{-1/2} - (1+2s^2)^{-1}$ and the standard deviation terms

$$\tilde{\sigma} \coloneqq \sqrt{\mathrm{Var}[\mathbf{X}_1]} \,, \qquad \sigma \coloneqq \sqrt{\frac{1}{k}\mathrm{Var}[\phi_{11}\mathbf{X}_1] + \frac{k-1}{k}\mathrm{Cov}[\phi_{11}\mathbf{X}_1, \phi_{12}\mathbf{X}_1]} \,.$$

Recall from (D.52) and (D.53) in the proof of Proposition 6.7 (empirical averages) that

$$\begin{aligned} \frac{1}{nk}\sum_{i=1}^n \sum_{j=1}^k \mathbf{Z}_{ij} &\sim \mathcal{N}\big(\mathbb{E}[\phi_{11}\mathbf{X}_1], \tfrac{1}{n}\sigma^2\big) \equiv \mathcal{N}\big(0, \tfrac{1}{n}\sigma^2\big) \,, \\ \frac{1}{nk}\sum_{i=1}^n \sum_{j=1}^k \tilde{\mathbf{Z}}_{ij} &\sim \mathcal{N}\big(\mathbb{E}[\mathbf{X}_1], \tfrac{1}{n}\tilde{\sigma}^2\big) \equiv \mathcal{N}\big(0, \tfrac{1}{n}\tilde{\sigma}^2\big) \,. \end{aligned}$$

Thus, the following quantities are both chi-squared distributed with 1 degree of freedom:

$$\begin{aligned} -\frac{1}{\sigma^2}\log f(\mathcal{Z}) &= \frac{1}{\sigma^2}\big(\frac{1}{\sqrt{nk}}\sum_{i=1}^n \sum_{j=1}^k \mathbf{Z}_{ij}\big)^2, \\ -\frac{1}{\tilde{\sigma}^2}\log f(\tilde{\mathcal{Z}}) &= \frac{1}{\tilde{\sigma}^2}\big(\frac{1}{\sqrt{nk}}\sum_{i=1}^n \sum_{j=1}^k \tilde{\mathbf{Z}}_{ij}\big)^2. \end{aligned} \tag{D.57}$$

Let $\mathbf{U}$ be a chi-squared distributed random variable with 1 degree of freedom. We can now use the formula of moment generating functions of χ_1^2 to get

$$\begin{aligned} \mathrm{Var}[f(\mathcal{Z})] = \mathrm{Var}[\exp(-\sigma^2\mathbf{U})] &= \mathbb{E}[\exp(-2\sigma^2\mathbf{U})] - \mathbb{E}[\exp(-\sigma^2\mathbf{U})]^2 \\ &= (1+4\sigma^2)^{-1/2} - (1+2\sigma^2)^{-1} = V(\sigma) \,, \end{aligned}$$

as desired. The same argument gives the desired variance for the unaugmented case:

$$\mathrm{Var}[f(\tilde{\mathcal{Z}})] \;=\; V(\tilde{\sigma})\;,$$

and the ratio $\vartheta(f)$ defined in (6.8) can be computed by:

$$\vartheta(f) \;=\; \sqrt{\mathrm{Var}[f(\tilde{\mathcal{Z}})]/\mathrm{Var}[f(\mathcal{Z})]} \;=\; \sqrt{V(\tilde{\sigma})/V(\sigma)}\;.$$

Finally, notice that (π_l, π_u) are the lower and upper $\alpha/2$-th quantiles for the quantities in (D.57). The corresponding quantiles for $f(\mathcal{Z})$ and $f(\tilde{\mathcal{Z}})$ then follow by monotonicity of the transforms $x \mapsto \exp(-\sigma^2 x)$ and $x \mapsto \exp(-\tilde{\sigma}^2 x)$: They are given by

$$\big(\exp\big(-\sigma^2\pi_u\big),\, \exp\big(-\sigma^2\pi_l\big)\big) \qquad \text{and} \qquad \big(\exp(-\tilde{\sigma}^2\pi_u),\, \exp(-\tilde{\sigma}^2\pi_l)\big)\;,$$

as required, and are asymptotically valid for $f(\Phi\mathcal{X})$ and $f(\tilde{\mathcal{X}})$ respectively since convergence in $d_{\mathcal{H}}$ implies convergence in distribution by Lemma 6.3. □

We next prove Lemma D.14 concerning the 2d generalisation of the toy statistic (6.13):

$$f_2(\mathbf{x}_{11:nk}) \;\coloneqq\; \textstyle\sum_{s=1}^{2} \exp\big(-\big(\frac{1}{\sqrt{nk}}\sum_{i=1}^{n}\sum_{j=1}^{k} x_{ijs}\big)^2\big)\;.$$

Proof of Lemma D.14. The proof for (i) is similar to the 1d case. Recall that in the proof of Proposition D.13, we have defined $g(x) \coloneqq \frac{1}{\sqrt{n}}\exp(-nx^2)$. Define $g_2 : \mathbb{R}^2 \to \mathbb{R}$ and $\tilde{f}_2 : \mathbb{R}^{2nk} \to \mathbb{R}$ as

$$g_2(\mathbf{x}) \;\coloneqq\; \textstyle\sum_{s=1}^{2} g(x_s)\;, \qquad \text{and} \qquad \tilde{f}_2(\mathbf{x}_{11:nk}) \;\coloneqq\; g_2\big(\frac{1}{nk}\sum_{i\le n, j\le k}\mathbf{x}_{ij}\big)\;.$$

Then as before, $\tilde{f}_2(\mathbf{x}_{11:nk}) = \sqrt{n} f_2(\mathbf{x}_{11:nk})$, and applying Lemma D.7(ii) to $\tilde{f}_2$ and g_2 will recover convergences for

$$d_{\mathcal{H}}(\sqrt{n}\tilde{f}_2(\Phi\mathcal{X}), \sqrt{n}\tilde{f}_2(\mathcal{Z})) \;=\; d_{\mathcal{H}}(f_2(\Phi\mathcal{X}), f_2(\mathcal{Z}))\;, \tag{D.58}$$

$$n(\mathrm{Var}[\tilde{f}_2(\Phi\mathcal{X})] - \mathrm{Var}[\tilde{f}_2(\mathcal{Z})]) \;=\; \mathrm{Var}[f_2(\Phi\mathcal{X})] - \mathrm{Var}[f_2(\mathcal{Z})]\;. \tag{D.59}$$

To compute the noise stability terms for g_2, recall from the definition in (D.3) that

$$\overline{\mathbf{W}}_i(\mathbf{w}) \;\coloneqq\; \frac{1}{nk}\big(\textstyle\sum_{i'=1}^{i-1}\sum_{j=1}^{k}\phi_{i'j}\mathbf{X}_{i'} + \sum_{j=1}^{k}\mathbf{w}_j + \sum_{i'=i+1}^{n}\sum_{j=1}^{k}\mathbf{Z}_{i'j}\big) \in \mathbb{R}^2\;.$$

Denote its two coordinates by $\overline{\mathbf{W}}_{i1}(\mathbf{w})$ and $\overline{\mathbf{W}}_{i2}(\mathbf{w})$. Then by linearity of differentiation followed by triangle inequality of $\zeta_{i;m}$ from Lemma D.20,

$$\begin{aligned}
\nu_{r;m}(g_2) \;&=\; \max_{i\le n}\zeta_{i;m}\big(\big\|\partial^r g_2\big(\overline{\mathbf{W}}_i(\bullet)\big)\big\|\big)\\
&=\; \max_{i\le n}\zeta_{i;m}\big(\big\|\partial^r g\big(\overline{\mathbf{W}}_{i1}(\bullet)\big) + \partial^r g\big(\overline{\mathbf{W}}_{i2}(\bullet)\big)\big\|\big)\\
&\le\; \max_{i\le n}\zeta_{i;m}\big(\big\|\partial^r g\big(\overline{\mathbf{W}}_{i1}(\bullet)\big)\big\|\big) + \max_{i\le n}\zeta_{i;m}\big(\big\|\partial^r g\big(\overline{\mathbf{W}}_{i2}(\bullet)\big)\big\|\big)\\
&=:\; \nu^{(1)}_{r;m}(g) + \nu^{(2)}_{r;m}(g)\;.
\end{aligned}$$

Note that $\nu^{(1)}_{r;m}(g)$ is $\nu_{r;m}(g)$ defined with respect to the sets of 2d data $\Phi\mathcal{X}$ and $\mathcal{Z}$ but restricted to their first coordinates, and $\nu^{(2)}_{r;m}(g)$ with respect to the data restricted to their second. The model (D.6) ensures existence of all moments, so the same bounds computed for $\nu_{r;m}(g)$ in the 1d case in the proof of Proposition D.13 directly apply to $\nu^{(1)}_{r;m}(g)$, $\nu^{(2)}_{r;m}(g)$ and consequently $\nu_{r;m}(g_2)$. Since we also have $c_x, c_Z = O(1)$, the bounds on (D.58) and (D.59) are $O(n^{-1/2})$, exactly the same as the 1d case. Applying Lemma D.7(ii) proves the required convergences in (i) as $n \to \infty$ as before.

For (ii), by Lemma D.19 and linearity of ϕ_{11}, ϕ_{12},

$$\begin{aligned}\mathrm{Cov}[\phi_{11}\mathbf{X}_1, \phi_{12}\mathbf{X}_2] &= \mathbb{E}\mathrm{Cov}[\phi_{11}\mathbf{X}_1, \phi_{12}\mathbf{X}_2|\phi_{11}, \phi_{12}] \\ &= \mathbb{E}[\phi_{11}]\mathrm{Var}[\mathbf{X}_{11}]\mathbb{E}[\phi_{12}] = \frac{(1+\rho)\sigma^2}{2}\mathbf{1}_{2\times 2} .\end{aligned}$$

Meanwhile, note that $\phi_{ij}\mathbf{X}_i \overset{d}{=} \mathbf{X}_i$, which implies that $\mathrm{Var}[\phi_{11}\mathbf{X}_1] = \mathrm{Var}[\mathbf{X}_1] = \sigma^2\left(\begin{smallmatrix}1 & \rho\\ \rho & 1\end{smallmatrix}\right)$ and $\mathbb{E}[\phi_{11}\mathbf{X}_1] = \mathbb{E}[\mathbf{X}_1] = \mathbf{0}$. Substituting these into the formula for moments of $\mathbf{Z}_i$ from (6.1) gives the mean and variance required:

$$\mathbb{E}\mathbf{Z}_i = \mathbf{0} , \qquad \mathrm{Var}\mathbf{Z}_i = \sigma^2\mathbf{I}_k \otimes \left(\begin{smallmatrix}1 & \rho\\ \rho & 1\end{smallmatrix}\right) + \frac{(1+\rho)\sigma^2}{2}(\mathbf{1}_{k\times k} - \mathbf{I}_k) \otimes \mathbf{1}_{2\times 2} .$$

Similarly, substituting the calculations into the formula for moments of $\tilde{\mathbf{Z}}_i$ from (6.6) gives $\mathbb{E}[\tilde{\mathbf{Z}}_i] = \mathbf{0}$ and $\mathrm{Var}[\tilde{\mathbf{Z}}_i] = \sigma^2\mathbf{1}_{k\times k} \otimes \left(\begin{smallmatrix}1 & \rho\\ \rho & 1\end{smallmatrix}\right)$.

To compute (iii), first re-express the variance of $\mathbf{Z}_i$ above as

$$\begin{aligned}\mathrm{Var}\mathbf{Z}_i &= \sigma^2\mathbf{I}_k \otimes \begin{pmatrix}\frac{1-\rho}{2} & \frac{\rho-1}{2}\\ \frac{\rho-1}{2} & \frac{1-\rho}{2}\end{pmatrix} + \frac{(1+\rho)\sigma^2}{2}\mathbf{1}_{2k\times 2k} \\ &= \frac{(1-\rho)\sigma^2}{2}\mathbf{I}_k \otimes \left(\begin{smallmatrix}1 & -1\\ -1 & 1\end{smallmatrix}\right) + \frac{(1+\rho)\sigma^2}{2}\mathbf{1}_{2k\times 2k} = \sigma_-^2\mathbf{I}_k \otimes \left(\begin{smallmatrix}1 & -1\\ -1 & 1\end{smallmatrix}\right) + \sigma_+^2\mathbf{1}_{2k\times 2k} ,\end{aligned}$$

Notice that the structure in mean and variance of $\mathbf{Z}_i$ allows us to rewrite it as a combination of simple 1d Gaussian random variables. Consider $\mathbf{U}_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_-^2)$ for $i \le n, j \le k$ and $\mathbf{V}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_+^2)$ independent of $\mathbf{U}_{ij}$'s. Define the random vector in $\mathbb{R}^{2k}$ as

$$\xi_i := (\mathbf{U}_{i;1} + \mathbf{V}_i, -\mathbf{U}_{i;1} + \mathbf{V}_i, \mathbf{U}_{i;2} + \mathbf{V}_i, -\mathbf{U}_{i;2} + \mathbf{V}_i, \ldots, \mathbf{U}_{i;k} + \mathbf{V}_i, -\mathbf{U}_{i;k} + \mathbf{V}_i)^\top .$$

Since $\mathbb{E}\mathbf{Z}_i = \mathbb{E}\xi_i$ and $\mathrm{Var}\mathbf{Z}_i = \mathrm{Var}\xi_i$, we have $\xi_i \overset{d}{=} \mathbf{Z}_i$, which implies

$$\begin{aligned}f_2(\mathcal{Z}) &\overset{d}{=} f_2(\xi_1, \ldots, \xi_n) \\ &= \exp\left(-\left(\frac{1}{\sqrt{nk}}\textstyle\sum_{i,j}(\mathbf{U}_{ij} + \mathbf{V}_i)\right)^2\right) + \exp\left(-\left(\frac{1}{\sqrt{nk}}\textstyle\sum_{i,j}(-\mathbf{U}_{ij} + \mathbf{V}_i)\right)^2\right) \\ &=: \exp(-\mathbf{S}_+) + \exp(-\mathbf{S}_-) ,\end{aligned}$$

and therefore

$$\text{Var}[f_2(\mathcal{Z})] \;=\; \text{Var}[\exp(-\mathbf{S}_+)] + \text{Var}[\exp(-\mathbf{S}_-)] + 2\text{Cov}[\exp(-\mathbf{S}_+), \exp(-\mathbf{S}_-)] \;. \tag{D.60}$$

Notice that $\mathbf{S}_+ \coloneqq \frac{1}{\sqrt{nk}} \sum_{i,j} (\mathbf{U}_{ij} + \mathbf{V}_i)$ and $\mathbf{S}_- \coloneqq \frac{1}{\sqrt{nk}} \sum_{i,j} (-\mathbf{U}_{ij} + \mathbf{V}_i)$ are both normally distributed with mean 0 and variance $\sigma_S^2 \coloneqq \frac{\sigma_-^2}{k} + \sigma_+^2$. This means $\frac{\mathbf{S}_+}{\sigma_S^2}$ and $\frac{\mathbf{S}_-}{\sigma_S^2}$ are both chi-squared distributed with 1 degree of freedom, and the formula for moment generating function of chi-squared distribution again allows us to compute

$$\begin{aligned}
\mathbb{E}[\exp(-\mathbf{S}_+)] \;&=\; \mathbb{E}[\exp(-\mathbf{S}_-)] \;=\; (1 + 2\sigma_S^2)^{-1/2} \;, \\
\text{Var}[\exp(-\mathbf{S}_+)] \;&=\; \text{Var}[\exp(-\mathbf{S}_-)] \;=\; (1 + 4\sigma_S^2)^{-1/2} - (1 + 2\sigma_S^2)^{-1} \;.
\end{aligned}$$

Moreover, write $\bar{\mathbf{U}} \coloneqq \frac{1}{\sqrt{nk}} \sum_{i,j} \mathbf{U}_{ij} \sim \mathcal{N}\big(0, \frac{\sigma_-^2}{k}\big)$ and $\bar{\mathbf{V}} \coloneqq \frac{1}{\sqrt{n}} \sum_{i \le n} \mathbf{V}_i \sim \mathcal{N}(0, \sigma_+^2)$. We have

$$\begin{aligned}
\mathbb{E}[\exp(-\mathbf{S}_+ - \mathbf{S}_-)] \;&= \mathbb{E}[\exp(-(\bar{\mathbf{U}} + \bar{\mathbf{V}})^2 - (-\bar{\mathbf{U}} + \bar{\mathbf{V}})^2] \\
&= \mathbb{E}[\exp(-2\bar{\mathbf{U}}^2 - 2\bar{\mathbf{V}}^2)] \;=\; \mathbb{E}[\exp(-2\bar{\mathbf{U}}^2)]\mathbb{E}[\exp(-2\bar{\mathbf{V}}^2)] \\
&= \Big(1 + \frac{4\sigma_-^2}{k}\Big)^{-1/2} (1 + 4\sigma_+^2)^{-1/2} \;,
\end{aligned}$$

which implies

$$\begin{aligned}
\text{Cov}[\exp(-\mathbf{S}_+), \exp(-\mathbf{S}_-)] \;&=\; \mathbb{E}[\exp(-\mathbf{S}_+ - \mathbf{S}_-)] - \mathbb{E}[\exp(-\mathbf{S}_+)]\mathbb{E}[\exp(-\mathbf{S}_-)] \\
&=\; \Big(1 + \frac{4\sigma_-^2}{k}\Big)^{-1/2} (1 + 4\sigma_+^2)^{-1/2} - (1 + 2\sigma_S^2)^{-1} \;.
\end{aligned}$$

Substituting the calculations for variances and covariance into (D.60), we obtain

$$\begin{aligned}
&\text{Var}[f_2(\mathcal{Z})] \\
&=\; 2\big((1 + 4\sigma_S^2)^{-1/2} - (1 + 2\sigma_S^2)^{-1}\big) + 2\Big(\Big(1 + \frac{4\sigma_-^2}{k}\Big)^{-1/2} (1 + 4\sigma_+^2)^{-1/2} - (1 + 2\sigma_S^2)^{-1}\Big) \\
&=\; 2(1 + 4\sigma_S^2)^{-1/2} + 2\Big(1 + \frac{4\sigma_-^2}{k}\Big)^{-1/2} (1 + 4\sigma_+^2)^{-1/2} - 4(1 + 2\sigma_S^2)^{-1} \\
&=\; 2\Big(1 + \frac{4\sigma_-^2}{k} + 4\sigma_+^2\Big)^{-1/2} + 2\Big(1 + \frac{4\sigma_-^2}{k}\Big)^{-1/2} (1 + 4\sigma_+^2)^{-1/2} - 4(1 + \frac{2\sigma_-^2}{k} + 2\sigma_+^2)^{-1} \;,
\end{aligned}$$

which is the required formula. □

D.6.3. Ridge regression

In this section, it is useful to define the function $g_B : \mathbb{M}^d \times \mathbb{R}^{d \times b} \to \mathbb{R}^{d \times b}$:

$$g_B(\Sigma, A) \;\coloneqq\; \tilde{\Sigma}^{-1} A \;, \tag{D.61}$$

which allows the ridge estimator to be written as

$$\hat{B}^{\Phi\mathcal{X}} \;\coloneqq\; \hat{B}(\Phi\mathcal{X}) \;=\; g_B\Big(\frac{1}{nk} \sum_{i,j} (\pi_{ij}\mathbf{V}_i)(\pi_{ij}\mathbf{V}_i)^\top, \frac{1}{nk} \sum_{i,j} (\pi_{ij}\mathbf{V}_i)(\tau_{ij}\mathbf{Y}_i)^\top\Big) \;.$$

Similarly, we can use g_B to rewrite the estimator with surrogate variables considered in Theorem 6.1 and the truncated first-order Taylor version in Lemma D.7:

$$\hat{B}^Z \coloneqq g_B\big(\tfrac{1}{nk}\textstyle\sum_{i,j}\mathbf{Z}_{ij}\big) \quad \text{and} \quad \hat{B}^T \coloneqq g_B(\mu) + \partial g_B(\mu)\big(\tfrac{1}{nk}\textstyle\sum_{i,j}\mathbf{Z}_{ij} - \mu\big)\ ,$$

where $\mu \coloneqq (\mu_1, \mu_2) \coloneqq \big(\mathbb{E}[(\pi_{11}\mathbf{V}_1)(\pi_{11}\mathbf{V}_1)^\top], \mathbb{E}[(\pi_{11}\mathbf{V}_1)(\tau_{11}\mathbf{V}_1)^\top]\big)$. Similarly, consider the function $g_R : \mathbb{M}^d \times \mathbb{R}^{d\times b} \to \mathbb{R}$ defined by

$$g_R(\Sigma, A) \coloneqq \mathbb{E}[\|\mathbf{Y}_{new} - (\tilde{\Sigma}^{-1}A)^\top \mathbf{V}_{new}\|_2^2]\ . \tag{D.62}$$

This allows us to write the risk as

$$R^{\Phi\mathcal{X}} = g_R\Big(\tfrac{1}{nk}\textstyle\sum_{i,j}(\pi_{ij}\mathbf{V}_i)(\pi_{ij}\mathbf{V}_i)^\top, \tfrac{1}{nk}\sum_{i,j}(\pi_{ij}\mathbf{V}_i)(\tau_{ij}\mathbf{Y}_i)^\top\Big)\ ,$$

while the estimator considered in Theorem 6.1 and the first-order Taylor version in Lemma D.7 become

$$R^Z \coloneqq g_R\big(\tfrac{1}{nk}\textstyle\sum_{i,j}\mathbf{Z}_{ij}\big)\ , \quad \text{and} \quad R^T \coloneqq g_R(\mu) + \partial g_R(\mu)\big(\tfrac{1}{nk}\textstyle\sum_{i,j}\mathbf{Z}_{ij} - \mu\big)\ ,$$

In this section, we first prove

(i) the convergence of $\hat{B}^{\Phi\mathcal{X}}$ to $\hat{B}^Z$ and $\hat{B}^T$, and the convergence of $R^{\Phi\mathcal{X}}$ to R^Z and R^T, with each convergence rate specified, and

(ii) existence of surrogate variables satisfying those convergences.

The proof for (i) follows an argument analogous to previous examples: we compute derivatives of the estimator of interest, and apply variants of Theorem 6.1 to obtain convergences. The results are collected in Lemma D.32 in Appendix D.6.3. The comment on different convergence rates in Remark 6.3 is also clear from Lemma D.32.

(ii) is of concern in this setup because the surrogate variables can no longer be Gaussian. Appendix D.6.3 states one possible choice from an approximate maximum entropy principle. Combining (i) and (ii) gives the statement in Proposition 6.8.

Finally, Appendix D.6.3 focuses on the toy model in (6.15). We prove Lemma 6.9, which discusses the non-monotonicity of variance of risk as a function of data variance. We also prove Lemma D.35, a formal statement of Remark 6.3 that $\mathrm{Var}[R^{\Phi\mathcal{X}}]$ does not converge to $\mathrm{Var}[R^T]$ for sufficiently high dimensions under a toy model.

Proof for convergence of variance and weak convergence

Lemma D.32. *Assume that* $\max_{l\le d}\max\{(\pi_{11}\mathbf{V}_1)_l, (\tau_{11}\mathbf{Y}_1)_l\}$ *is a.s. bounded by* $C\tau$ *for some* τ *to be specified and some absolute constant* $C > 0$*, and that* $b = O(d)$*. Then, for any i.i.d. surrogate variables* $\{\mathbf{Z}_i\}_{i\le n}$ *taking values in* $(\mathbb{M}^d \times \mathbb{R}^{d\times b})^k$ *matching the first moments of* $\Phi_1\mathbf{X}_1$ *with all coordinates uniformly bounded by* $C'\tau^2$ *a.s. for some absolute constant* $C' > 0$*, we have:*

(i) assuming $\tau = O(1)$ and fixing $r \le d$, $s \le b$, then the (r,s)-the coordinate of $\hat{B}^{\Phi\mathcal{X}}$ satisfies

$$d_{\mathcal{H}}\big(\sqrt{n}\big(\hat{B}^{\Phi\mathcal{X}}\big)_{r,s}, \sqrt{n}\big(\hat{B}^{T}\big)_{r,s}\big) \;=\; O(n^{-1/2}d^{9}) \;,$$
$$d_{\mathcal{H}}\big(\sqrt{n}\big(\hat{B}^{\Phi\mathcal{X}}\big)_{r,s}, \sqrt{n}\big(\hat{B}^{Z}\big)_{r,s}\big) \;=\; O(n^{-1/2}d^{9}) \;;$$

(ii) assuming $\tau = O(d^{-1/2})$, then $\hat{B}^{\Phi\mathcal{X}}$ satisfies

$$n\|\mathrm{Var}[\hat{B}^{\Phi\mathcal{X}}] - \mathrm{Var}[\hat{B}^{T}]\| \;=\; O\big(n^{-1/2}d^{7} + n^{-1}d^{8}\big) \;,$$
$$n\|\mathrm{Var}[\hat{B}^{\Phi\mathcal{X}}] - \mathrm{Var}[\hat{B}^{Z}]\| \;=\; O\big(n^{-1}d^{7}\big) \;;$$

(iii) assuming $\tau = O(d^{-1/2})$, then $R^{\Phi\mathcal{X}}$ satisfies

$$d_{\mathcal{H}}(\sqrt{n}R^{\Phi\mathcal{X}},\sqrt{n}R^{T}) \;=\; O(n^{-1/2}d^{9}) \;, \quad d_{\mathcal{H}}(\sqrt{n}R^{\Phi\mathcal{X}},\sqrt{n}R^{\mathcal{Z}}) \;=\; O(n^{-1/2}d^{9}) \;,$$
$$n(\mathrm{Var}[R^{\Phi\mathcal{X}}] - \mathrm{Var}[R^{T}]) \;=\; O(n^{-1/2}d^{7} + n^{-1}d^{8}) \;,$$
$$n(\mathrm{Var}[R^{\Phi\mathcal{X}}] - \mathrm{Var}[R^{Z}]) \;=\; O(n^{-1}d^{7}) \;.$$

Remark D.5. In the statement of weak convergence of the estimator $\hat{B}^{\Phi\mathcal{X}}$, we only consider convergence of one coordinate of $\hat{B}^{\Phi\mathcal{X}}$ since we allow dimensions d, b to grow with n; this setting was discussed in more details in Lemma D.4. The assumption $\tau = O(1)$ for (i) is such that the coordinates we are studying do not go to zero as n grows, while the assumption $\tau = O(d^{-1/2})$ for (ii) and (iii) is such that $\|\pi_{11}\mathbf{V}_1\|$ and $\|\tau_{11}\mathbf{Y}_1\|$ are $O(1)$ as n grows, which keeps $\|\hat{B}^{\Phi\mathcal{X}}\|$ and $R^{\Phi\mathcal{X}}$ bounded.

Remark D.6. The difference between the convergence rate of $\mathrm{Var}[R^{\Phi\mathcal{X}}]$ towards $\mathrm{Var}[R^{Z}]$ and that towards $\mathrm{Var}[R^{T}]$ is clear in the additional factor $(n^{1/2} + d)$ in Lemma D.32(iii). If we take d to be $\Theta(n^{\alpha})$ for $\frac{1}{14} < \alpha < \frac{1}{7}$, we are guaranteed convergence of $\mathrm{Var}[R^{\Phi\mathcal{X}}]$ to $\mathrm{Var}[R^{Z}]$ but not necessarily convergence of $\mathrm{Var}[R^{\Phi\mathcal{X}}]$ to $\mathrm{Var}[R^{T}]$. Note that the bounds here are not necessarily tight in terms of dimensions, and we discuss this difference in convergence rate in more details in Appendix D.6.4.

Proof of Lemma D.32(i). We first prove the weak convergence statements for $(\hat{B}^{\Phi\mathcal{X}})_{r,s}$. Let $\mathbf{e}_r$ be the r-th basis vector of $\mathbb{R}^d$ and $\mathbf{o}_s$ be the s-th basis vector of $\mathbb{R}^b$. We define the function $g_{B;rs} : \mathbb{M}^d \times \mathbb{R}^{d\times b} \to \mathbb{R}$ as

$$g_{B;rs}(\Sigma, A) \;\coloneqq\; \mathbf{e}_r^{\top} g_B(\Sigma, A)\mathbf{o}_s \;=\; \mathbf{e}_r^{\top}\tilde{\Sigma}^{-1}A\mathbf{o}_s \;,$$

i.e. the (r,s)-th coordinate of g_B. The (r,s)-th coordinate of $\hat{B}^{\Phi\mathcal{X}}$, $\hat{B}^{Z}$ and $\hat{B}^{T}$ can then be expressed in terms of $g_{B;rs}$ similar to before:

$$\big(\hat{B}^{\Phi\mathcal{X}}\big)_{r,s} \;=\; g_{B;rs}\Big(\tfrac{1}{nk}\textstyle\sum_{i,j}(\pi_{ij}\mathbf{V}_i)(\pi_{ij}\mathbf{V}_i)^{\top}, \tfrac{1}{nk}\sum_{i,j}(\pi_{ij}\mathbf{V}_i)(\tau_{ij}\mathbf{Y}_i)^{\top}\Big) \;,$$
$$\big(\hat{B}^{Z}\big)_{r,s} \;=\; g_{B;rs}\big(\tfrac{1}{nk}\textstyle\sum_{i,j}\mathbf{Z}_{ij}\big) \;, \quad \big(\hat{B}^{T}\big)_{r,s} \;=\; g_{B;rs}(\mu) + \partial g_{B;rs}(\mu)\big(\tfrac{1}{nk}\textstyle\sum_{i,j}\mathbf{Z}_{ij} - \mu\big) \;.$$

To obtain weak convergence of $(\hat{B}^{\Phi\mathcal{X}})_{r,s}$ to $(\hat{B}^{Z})_{r,s}$ and $(\hat{B}^{T})_{r,s}$, it suffices to apply the result for the plug-in estimates from Lemma D.7 with $\delta = 0$ to the function $g_{B;rs}$ with respect to the transformed data $\phi_{ij}\mathbf{X}_i^* := \big((\pi_{ij}\mathbf{V}_i)(\pi_{ij}\mathbf{V}_i)^\top, (\pi_{ij}\mathbf{V}_i)(\tau_{11}\mathbf{Y}_i)^\top\big)$.

As before, we start with computing the partial derivatives of $g_{B;rs}(\Sigma, A)$, which can be expressed using $\tilde{\Sigma} := \Sigma + \lambda\mathbf{I}_d$ and A as:

$$\begin{aligned}
g_{B;rs}(\Sigma, A) &= \mathbf{e}_r^\top\tilde{\Sigma}^{-1}A\mathbf{o}_s,\\
\frac{\partial g_{B;rs}(\Sigma, A)}{\partial\Sigma_{r_1s_1}} &= -\mathbf{e}_r^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_1}\mathbf{e}_{s_1}^\top\tilde{\Sigma}^{-1}A\mathbf{o}_s, \quad \frac{\partial g_{B;rs}(\Sigma, A)}{\partial A_{r_1s_1}} = \mathbf{e}_r^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_1}\mathbb{I}_{\{s=s_1\}},\\
\frac{\partial^2 g_{B;rs}(\Sigma, A)}{\partial\Sigma_{r_1s_1}\partial\Sigma_{r_2s_2}} &= \sum\nolimits_{l_1,l_2\in\{1,2\};\, l_1\neq l_2}\mathbf{e}_r\tilde{\Sigma}^{-1}\mathbf{e}_{r_{l_1}}\mathbf{e}_{s_{l_1}}^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_{l_2}}\mathbf{e}_{s_{l_2}}^\top\tilde{\Sigma}^{-1}A\mathbf{o}_s,\\
\frac{\partial^2 g_{B;rs}(\Sigma, A)}{\partial\Sigma_{r_1s_1}\partial A_{r_2s_2}} &= -\mathbf{e}_r^\top\,\tilde{\Sigma}^{-1}\,\mathbf{e}_{r_1}\mathbf{e}_{s_1}^\top\,\tilde{\Sigma}^{-1}\,\mathbf{e}_{r_2}\mathbb{I}_{\{s=s_2\}}, \qquad \frac{\partial^2 g_{B;rs}(\Sigma, A)}{\partial A_{r_1s_1}\partial A_{r_2s_2}} = \mathbf{0},\\
\frac{\partial^3 g_{B;rs}(\Sigma, A)}{\partial\Sigma_{r_1s_1}\partial\Sigma_{r_2s_2}\partial\Sigma_{r_3s_3}} &= -\sum\nolimits_{\substack{l_1,l_2,l_3\in\{1,2,3\}\\ l_1,l_2,l_3\text{ distinct}}}\mathbf{e}_r^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_{l_1}}\mathbf{e}_{s_{l_1}}^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_{l_2}}\mathbf{e}_{s_{l_2}}^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_{l_3}}\mathbf{e}_{s_{l_3}}^\top\tilde{\Sigma}^{-1}A\mathbf{o}_s,\\
\frac{\partial^3 g_{B;rs}(\Sigma, A)}{\partial\Sigma_{r_1s_1}\partial\Sigma_{r_2s_2}\partial A_{r_3s_3}} &= \sum\nolimits_{l_1,l_2\in\{1,2\};\, l_1\neq l_2}\mathbf{e}_r^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_{l_1}}\mathbf{e}_{s_{l_1}}^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_{l_2}}\mathbf{e}_{s_{l_2}}^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_3}\mathbb{I}_{\{s=s_3\}}.
\end{aligned} \tag{D.63}$$

To bound the norm of the derivatives, it is useful to have controls over the norms of $\tilde{\Sigma}^{-1}$ and A. Suppose the coordinates of A are uniformly bounded by $C\tau^2$ for some absolute constant $C > 0$, which is the case when we compute the derivatives in $\nu_{r;m}$. Then since $b = O(d)$, we have

$$\begin{aligned}
\|A\|_{op} &\leq \|A\| = O(d\tau^2), \qquad \|A\mathbf{o}_s\| = O(d^{1/2}\tau^2),\\
\|\tilde{\Sigma}\|_{op} &= \|\Sigma + \lambda\mathbf{I}_d\|_{op} = \frac{1}{\sigma_1 + \lambda} = O(1)\,,
\end{aligned}$$

where $\sigma_1 \geq 0$ is the smallest eigenvalue of the positive semi-definite matrix A. We also note that for any matrix $M \in \mathbb{R}^{n_1\times n_2}$ and vectors $\mathbf{u} \in \mathbb{R}^{n_2}, \mathbf{v} \in \mathbb{R}^{n_3}$,

$$\|M\mathbf{u}\| \leq \|M\|_{op}\|\mathbf{u}\|\,, \qquad \|\mathbf{u}\mathbf{v}^\top\|_{op} \leq \|\mathbf{u}\|\|\mathbf{v}\|\,.$$

Making use of these bounds, we can bound the norms of partial derivatives of g as follows:

$$\Big\|\frac{\partial g_{B;rs}(\Sigma, A)}{\partial\Sigma_{r_1s_1}}\Big\| \leq \|\tilde{\Sigma}^{-1}\mathbf{e}_{r_1}\|\|\mathbf{e}_{s_1}^\top\tilde{\Sigma}^{-1}A\| \leq \|\Sigma^{-1}\|_{op}^2\|A\|_{op} = O(d\tau^2)\,.$$

We can perform a similar argument for the remaining derivatives. It suffices to count the number of A in each expression and use the bound $\|A\|_{op} \leq \|A\| = O(d\tau^2)$:

$$\begin{aligned}
\Big\|\frac{\partial^2 g_{B;rs}(\Sigma, A)}{\partial A_{r_1s_1}\partial A_{r_2s_2}}\Big\|,\ \Big\|\frac{\partial^3 g_{B;rs}(\Sigma, A)}{\partial\Sigma_{r_1s_1}\partial A_{r_2s_2}\partial M_{r_3s_3}}\Big\|,\ \Big\|\frac{\partial^3 g_{B;rs}(\Sigma, A)}{\partial A_{r_1s_1}\partial A_{r_2s_2}\partial A_{r_3s_3}}\Big\| &= 0,\\
\Big\|\frac{\partial g_{B;rs}(\Sigma, A)}{\partial A_{rs}}\Big\|,\ \Big\|\frac{\partial^2 g_{B;rs}(\Sigma, A)}{\partial\Sigma_{r_1s_1}\partial A_{r_2s_2}}\Big\|,\ \Big\|\frac{\partial^3 g_{B;rs}(\Sigma, A)}{\partial\Sigma_{r_1s_1}\partial\Sigma_{r_2s_2}\partial A_{r_3s_3}}\Big\| &= O(1),
\end{aligned}$$

$$\|g_{B;rs}(\Sigma,A)\|,\ \Big\|\frac{\partial g_{B;rs}(\Sigma,A)}{\partial\Sigma_{rs}}\Big\|,\ \Big\|\frac{\partial^2 g_{B;rs}(\Sigma,A)}{\partial\Sigma_{r_1s_1}\partial\Sigma_{r_2s_2}}\Big\|,\ \Big\|\frac{\partial^3 g_{B;rs}(\Sigma,A)}{\partial\Sigma_{r_1s_1}\partial\Sigma_{r_2s_2}\partial\Sigma_{r_3s_3}}\Big\| \ = \ O(d\tau^2)\ .$$

This implies

$$\begin{aligned}
\big\|\partial g_{B;rs}(\Sigma,A)\big\| &= \sqrt{\textstyle\sum_{r_1,s_1=1}^{d}\Big\|\frac{\partial g_{B;rs}(\Sigma,A)}{\partial\Sigma_{r_1s_1}}\Big\|^2+\sum_{r_1=1}^{d}\sum_{s_1=1}^{b}\Big\|\frac{\partial g_{B;rs}(\Sigma,A)}{\partial A_{r_1s_1}}\Big\|^2}\\
&= O(d^2\tau^2+d)\ ,\\
\big\|\partial^2 g_{B;rs}(\Sigma,A)\big\| &= \sqrt{\sum_{\substack{r_1,r_2,\\ s_1,s_2=1}}^{d}\Big\|\frac{\partial^2 g_{B;rs}(\Sigma,A)}{\partial\Sigma_{r_1s_1}\partial\Sigma_{r_2s_2}}\Big\|^2+\sum_{r_1,s_1,r_2=1}^{d}\sum_{s_2=1}^{b}\Big\|\frac{\partial^2 g_{B;rs}(\Sigma,A)}{\partial\Sigma_{r_1s_1}\partial A_{r_2s_2}}\Big\|^2}\\
&= O(d^3\tau^2+d^2),\\
\big\|\partial^3 g_{B;rs}(\Sigma,A)\big\| &= \sqrt{\sum_{\substack{r_1,r_2,r_3,\\ s_1,s_2,s_3=1}}^{d}\Big\|\frac{\partial^3 g(\Sigma,A)}{\partial\Sigma_{r_1s_1}\partial\Sigma_{r_2s_2}\partial\Sigma_{r_3s_3}}\Big\|^2+\sum_{\substack{r_1,r_2,r_3,\\ s_1,s_2=1}}^{d}\sum_{s_3=1}^{b}\Big\|\frac{\partial^3 g(\Sigma,A)}{\partial\Sigma_{r_1s_1}\partial\Sigma_{r_2s_2}\partial A_{r_3s_3}}\Big\|^2}\\
&= O(d^4\tau^2+d^3)\ .
\end{aligned}$$

Recall that the noise stability terms in Lemma D.7 are defined by, for $\delta=0$,

$$\kappa_{t;m}(g)=\sum_{l\le q}\Big\|\sup_{\mathbf{w}\in[\mathbf{0},\bar{\mathbf{X}}]}\big\|\partial^t g_l\big(\mu+\mathbf{w}\big)\big\|\Big\|_{L_m},\ \nu_{t;m}(g)=\sum_{l\le q}\max_{i\le n}\zeta_{i;m}\big(\big\|\partial^t g_l\big(\overline{\mathbf{W}}_i(\,\bullet\,)\big)\big\|\big),$$

where $q=1$ in the case of $g_{B;rs}$, and the moment terms are defined by

$$\begin{aligned}
\bar{c}_m &= \Big(\textstyle\sum_{l=1}^{d^2+db}\max\big\{n^{\frac{2}{m}-1}\big\|\frac{1}{k}\sum_{j=1}^{k}[\phi_{1j}\mathbf{X}_1^*-\mu]_l\big\|_{L_m}^2,\ \big\|\frac{1}{k}\sum_{j=1}^{k}[\phi_{1j}\mathbf{X}_1^*-\mu]_l\big\|_{L_2}^2\big\}\Big)^{1/2},\\
c_X &= \frac{1}{6}\sqrt{\mathbb{E}[\|\phi_{11}\mathbf{X}_1^*\|^6}\ ,\quad c_Z \ = \ \frac{1}{6}\sqrt{\mathbb{E}\Big[\Big(\frac{|Z_{111}|^2+\ldots+|Z_{1k(d^2+db)}|^2}{k}\Big)^3\Big]}\ .
\end{aligned}$$

By the bounds on the derivatives of $g_{B;rs}$ from above, we get

$$\begin{aligned}
\kappa_{0;m}(g_{B;rs}),\ \nu_{0;m}(g_{B;rs}) &= O(d\tau^2)\ , & \kappa_{1;m}(g_{B;rs}),\ \nu_{1;m}(g_{B;rs}) &= O(d^2\tau^2+d)\ ,\\
\kappa_{2;m}(g_{B;rs}),\ \nu_{2;m}(g_{B;rs}) &= O(d^3\tau^2+d^2)\ , & \kappa_{3;m}(g_{B;rs}),\ \nu_{3;m}(g_{B;rs}) &= O(d^4\tau^2+d^3)\ ,
\end{aligned}$$

and since the coordinates of $\phi_{11}\mathbf{X}_1$ and $\mathbf{Z}_1$ are uniformly bounded by $C''\tau^2$ for $C''=\max\{C,C'\}$ almost surely, we get that

$$\bar{c}_m \ = \ O(d\tau^2)\ , \qquad\qquad c_X, c_Z \ = \ O(d^3\tau^6)\ .$$

Applying Lemma D.7(i) to $g_{B;rs}$ with $\delta=0$ and the assumption $\tau=O(1)$ then gives

$$\begin{aligned}
d_{\mathcal{H}}\big(\sqrt{n}\big(\hat{B}^{\Phi\mathcal{X}}\big)_{r,s},\sqrt{n}\big(\hat{B}^{T}\big)_{r,s}\big) &= O\big(n^{-1/2}\kappa_{2;3}(g_{B;rs})\,\bar{c}_3^2+n^{-1/2}\kappa_{1;1}(g_{B;rs})^3(c_X+c_Z)\big)\big)\\
&= O(n^{-1/2}d^5+n^{-1/2}d^6d^3) \ = \ O(n^{-1/2}d^9),
\end{aligned}$$

and applying Lemma D.7(ii) with δ set to 0 gives

$$\begin{aligned}
&d_{\mathcal{H}}\big(\sqrt{n}\big(\hat{B}^{\Phi\mathcal{X}}\big)_{r,s},\sqrt{n}\big(\hat{B}^{Z}\big)_{r,s}\big)\\
&= O\big(\big(n^{-1/2}\nu_{1;6}(g_{B;rs})^3+n^{-1}\nu_{1;4}(g_{B;rs})\nu_{2;4}(g_{B;rs})+n^{-3/2}\nu_{3;2}(g_{B;rs})\big)(c_X+c_Z)\big)
\end{aligned}$$

$$= O\big(\big(n^{-1/2}d^6 + n^{-1}d^5 + n^{-3/2}d^4\big)d^3\big) \;=\; O(n^{-1/2}d^9)\;.$$

These are the desired bounds concerning weak convergence of $\big(\hat{B}^{\Phi\mathcal{X}}\big)_{r,s}$. d_H indeed metrizes weak convergence here, since $\big(\hat{B}^{\Phi\mathcal{X}}\big)_{r,s} \in \mathbb{R}$ and Lemma 6.3 applies. □

Proof of Lemma D.32(ii). For convergence of variance of $\hat{B}^{\Phi\mathcal{X}}$, we need to apply Lemma D.7 to g_B instead of $g_{B;rs}$. Notice that the noise stability terms of g_B can be computed in terms of those for $g_{B;rs}$ already computed in the proof of (i):

$$\kappa_{t;m}(g_B) \;=\; \sum\nolimits_{r=1}^{d}\sum\nolimits_{s=1}^{b}\kappa_{t;m}(g_{B;rs})\;,\qquad \nu_{t;m}(g_B) \;=\; \sum\nolimits_{r=1}^{d}\sum\nolimits_{s=1}^{b}\nu_{t;m}(g_{B;rs})\;.$$

This suggests that

$$\begin{aligned}
\kappa_{0;m}(g_B),\ \nu_{0;m}(g_B) \;&=\; O(d^3\tau^2)\;, & \kappa_{1;m}(g_B),\ \nu_{1;m}(g_B) \;&=\; O(d^4\tau^2 + d^3)\;,\\
\kappa_{2;m}(g_B),\ \nu_{2;m}(g_B) \;&=\; O(d^5\tau^2 + d^4)\;, & \kappa_{3;m}(g_B),\ \nu_{3;m}(g_B) \;&=\; O(d^6\tau^2 + d^5)\;,
\end{aligned}$$

The moment terms are bounded as before: $\bar{c}_m = O(d\tau^2)$ and $c_X, c_Z = O(d^3\tau^6)$. Applying Lemma D.7 with $\delta = 0$ and the assumption $\tau = O(d^{-1/2})$ gives

$$\begin{aligned}
n\|\mathrm{Var}[\hat{B}^{\Phi\mathcal{X}}] - \mathrm{Var}[\hat{B}^{T}]\| \;&=\; O\big(n^{-1/2}\kappa_{1;1}(g_B)\kappa_{2;4}(g_B)\bar{c}_4^3 + n^{-1}\kappa_{2;6}(g_B)\kappa_{2;6}(g_B)\,\bar{c}_6^4\big)\\
&=\; O\big(n^{-1/2}d^7 + n^{-1}d^8\big)\;,\\
n\|\mathrm{Var}[\hat{B}^{\Phi\mathcal{X}}] - \mathrm{Var}[\hat{B}^{Z}]\| \;&=\; O\big(n^{-1}(\nu_{0;4}(g_B)\nu_{3;4}(g_B) + \nu_{1;4}(g_B)\nu_{2;4}(g_B))(c_X + c_Z)\big)\\
&=\; O\big(n^{-1}d^7\big)\;,
\end{aligned}$$

which are the desired bounds for convergence of variance of $\hat{B}^{\Phi\mathcal{X}}$. □

Proof of Lemma D.32(iii). We seek to apply Lemma D.7 to g_R. Define

$$c^Y \;:=\; \mathbb{E}[\|\mathbf{Y}_{new}\|_2^2]\;,\quad C_{rs}^{VY} \;:=\; \big(\mathbb{E}[\mathbf{V}_{new}\mathbf{Y}_{new}^\top]\big)_{rs}\;,\quad C_{rs}^{V} \;:=\; \big(\mathbb{E}[\mathbf{V}_{new}\mathbf{V}_{new}^\top]\big)_{rs}\;,$$

This allows us to rewrite g_R as

$$\begin{aligned}
g_R(\Sigma, A) \;&=\; \mathbb{E}[\|\mathbf{Y}_{new} - g_B(\Sigma,A)^\top\mathbf{V}_{new}\|_2^2]\\
&= \mathbb{E}[\|\mathbf{Y}_{new}\|_2^2] - 2\mathrm{Tr}\big(\mathbb{E}[\mathbf{V}_{new}\mathbf{Y}_{new}^\top]g_B(\Sigma,A)^\top\big)\\
&\quad + \mathrm{Tr}\big(\mathbb{E}[\mathbf{V}_{new}\mathbf{V}_{new}^\top]g_B(\Sigma,A)g_B(\Sigma,A)^\top\big)\\
&= c^Y - 2\sum\nolimits_{r=1}^{d}\sum\nolimits_{s=1}^{b} C_{rs}^{VY}\, g_{B;rs}(\Sigma,A) + \sum\nolimits_{rs,t=1}^{d} C_{rs}^{V} g_{B;rt}(\Sigma,A)g_{B;ts}(\Sigma,A)\;.
\end{aligned}$$

As before, we first consider expressing derivatives of g_R in terms of those of $g_{B;rs}$. Omitting the (Σ, A)-dependence temporarily, we get

$$\begin{aligned}
\partial g_R \;&=\; -2\sum\nolimits_{r=1}^{d}\sum\nolimits_{s=1}^{b} C_{rs}^{VY}\,\partial g_{B;rs} + \sum\nolimits_{rs,t=1}^{d} C_{rs}^{V}\big(\partial g_{B;rt}g_{B;ts} + g_{B;rt}\partial g_{B;ts}\big),\\
\partial^2 g_R \;&=\; -2\sum\nolimits_{r=1}^{d}\sum\nolimits_{s=1}^{b} C_{rs}^{VY}\,\partial^2 g_{B;rs}\\
&\quad + \sum\nolimits_{rs,t=1}^{d} C_{rs}^{V}\big(\partial^2 g_{B;rt}g_{B;ts} + 2\partial g_{B;rt}\partial g_{B;ts} + g_{B;rt}\partial^2 g_{B;ts}\big),\\
\partial^3 g_R \;&=\; -2\sum\nolimits_{r=1}^{d}\sum\nolimits_{s=1}^{b} C_{rs}^{VY}\,\partial^3 g_{B;rs}
\end{aligned}$$

$$+\sum_{rs,t=1}^{d} C_{rs}^{V}\left(\partial^3 g_{B;rt} g_{B;ts} + 3\partial^2 g_{B;rt}\partial g_{B;ts} + 3\partial_{B;rt}^{g}\partial^2 g_{B;ts} + g_{B;rt}\partial^3 g_{B;ts}\right) .$$

Since the noise stability terms of g_R are given by

$$\kappa_{t;m}(g_R) \;=\; \Big\| \sup_{\mathbf{w}\in[\mathbf{0},\bar{\mathbf{X}}]} \big\|\partial^t g_R(\mu+\mathbf{w})\big\|\Big\|_{L_m}, \nu_{t;m}(g_R) \;=\; \max_{i\le n}\zeta_{i;m}\big(\big\|\partial^t g_R\big(\overline{\mathbf{W}}_i(\,\bullet\,)\big)\big\|\big),$$

they can be bounded in terms of those of $g_{B;rs}$ computed in the proof of (i). With the assumption $\tau = O(d^{-1/2})$, the noise stability terms of $g_{B;rs}$ become

$$\begin{aligned}
\kappa_{0;m}(g_{B;rs}),\ \nu_{0;m}(g_{B;rs}) \;&=\; O(1)\ , \qquad & \kappa_{1;m}(g_{B;rs}),\ \nu_{1;m}(g_{B;rs}) \;&=\; O(d)\ ,\\
\kappa_{2;m}(g_{B;rs}),\ \nu_{2;m}(g_{B;rs}) \;&=\; O(d^2)\ , \qquad & \kappa_{3;m}(g_{B;rs}),\ \nu_{3;m}(g_{B;rs}) \;&=\; O(d^3)\ .
\end{aligned}$$

Also note that $c^Y = O(d\tau) = O(1)$ and $C_{r,s}^{VY}, C_{r,s}^{V} = O(\tau) = O(d^{-1})$ by assumption. Then, by the triangle inequality followed by Hölder's inequality,

$$\begin{aligned}
\kappa_{0;m}(g_R) \;&\le c^Y + 2\sum_{r=1}^{d}\sum_{s=1}^{b} C_{r,s}^{VY}\kappa_{0;m}(g_{B;rs}) + \sum_{r,s,t=1}^{d} C_{r,s}^{V}\kappa_{0;m}(g_{B;rt}g_{B;ts})\\
&\le c^Y + 2\sum_{r=1}^{d}\sum_{s=1}^{b} C_{r,s}^{VY}\kappa_{0;m}(g_{B;rs}) + \sum_{r,s,t=1}^{d} C_{r,s}^{V}\kappa_{0;2m}(g_{B;rt})\kappa_{0;2m}(g_{B;ts})\\
&= O(1+d+d^2) \;=\; O(d^2)\ .
\end{aligned}$$

Similarly, by triangle inequality and Hölder's inequality of $\zeta_{i;m}$ in Lemma D.20,

$$\begin{aligned}
\nu_{0;m}(g_R) \;&\le c^Y + 2\sum_{r=1}^{d}\sum_{s=1}^{b} C_{r,s}^{VY}\nu_{0;m}(g_{B;rs}) + \sum_{r,s,t=1}^{d} C_{r,s}^{V}\nu_{0;2m}(g_{B;rt})\nu_{0;2m}(g_{B;ts})\\
&= O(1+d+d^2) \;=\; O(d^2)\ .
\end{aligned}$$

The same reasoning allows us to read out other noise stability terms of g_R directly in terms of those of $g_{R;rs}$ and bounds on $C_{r,s}^{VY}$ and $C_{r,s}^{V}$:

$$\begin{aligned}
\kappa_{1;m}(g_R),\ \nu_{1;m}(g_R) \;&=\; O(d^2+d^3) \;=\; O(d^3)\ ,\\
\kappa_{2;m}(g_R),\ \nu_{2;m}(g_R) \;&=\; O(d^4)\ , \qquad & \kappa_{3;m}(g_R),\ \nu_{3;m}(g_R) \;&=\; O(d^5)\ .
\end{aligned}$$

The moment terms are bounded as before: $\bar{c}_m = O(d\tau) = O(1)$ and $c_X, c_Z = O(d^3\tau^3) = O(1)$. By Lemma D.7 with δ set to 0, we have

$$\begin{aligned}
d_{\mathcal{H}}(\sqrt{n}R^{\Phi\mathcal{X}}, \sqrt{n}R^{T}) \;&= O\big(n^{-1/2}\kappa_{2;3}(g_R)\,\bar{c}_3^2 + n^{-1/2}\kappa_{1;1}(g_R)^3(c_X+c_Z)\big)\big)\ ,\\
&= O(n^{-1/2}d^4 + n^{-1/2}d^9) \;=\; O(n^{-1/2}d^9)\ ,\\
d_{\mathcal{H}}(\sqrt{n}R^{\Phi\mathcal{X}}, \sqrt{n}R^{\mathcal{Z}}) \;&= O\big(\big(n^{-1/2}\nu_{1;6}(g_R)^3 + 3n^{-1}\nu_{1;4}(g_R)\nu_{2;4}(g_R) + n^{-3/2}\nu_{3;2}(g_R)\big)\\
&\qquad \times (c_X+c_Z)\big)\\
&= O(n^{-1/2}d^9 + n^{-1}d^7 + n^{-3/2}d^5) \;=\; O(n^{-1/2}d^9)\ ,
\end{aligned}$$

which are the desired bounds in d_H, and by Lemma D.7 with $\delta = 0$ again, we have

$$\begin{aligned}
n(\mathrm{Var}[R^{\Phi\mathcal{X}}] - \mathrm{Var}[R^{T}]) \;&= O\big(n^{-1/2}\kappa_{1;1}(g_R)\kappa_{2;4}(g_R)\bar{c}_4^3 + n^{-1}\kappa_{2;6}(g_R)\kappa_{2;6}(g_R)\,\bar{c}_6^4\big)\\
&= O\big(n^{-1/2}d^7 + n^{-1}d^8\big)\ ,\\
n(\mathrm{Var}[R^{\Phi\mathcal{X}}] - \mathrm{Var}[R^{\mathcal{Z}}]) \;&= O\big(n^{-1}(\nu_{0;4}\nu_{3;4} + \nu_{1;4}\nu_{2;4})(c_X+c_Z)\big)
\end{aligned}$$

$$= O\big(n^{-1}d^7\big) \ ,$$

which are again the desired bounds for variance. □

Existence of surrogate variables from a maximum entropy principle As discussed after Proposition 6.8, the surrogate variables $\mathbf{Z}_i \coloneqq \{\mathbf{Z}_{ij}\}_{j\le k} = \{(\mathbf{Z}_{ij1}, \mathbf{Z}_{ij2})\}_{j\le k}$ cannot be Gaussian since they take values in $(\mathbb{M}^d \times \mathbb{R}^{d\times b})^k$. Recall that the only restriction we have on $\mathbf{Z}_i$ is from (6.1): $\mathbf{Z}_i$ should match the first two moments of $\Phi_i\mathbf{X}_i^*$. A trivial choice is $\Phi_i\mathbf{X}_i^*$ itself, but is not meaningful because the key of the theorem is that only the first two moments of $\Phi_i\mathbf{X}_i^*$ matter in the limit.

The main difficulty is finding a distribution $p_{\mathbb{M}}$ on $\mathbb{M}^d$, the set of $d \times d$ positive semi-definite matrices, such that for $\mathbf{Z}_{ij1} \sim p_{\mathbb{M}}$,

$$\mathbb{E}[\mathbf{Z}_{ij1}] \ = \ \mathbb{E}\big[(\pi_{11}\mathbf{V}_1)(\pi_{11}\mathbf{V}_1)^\top\big] \quad \text{and} \quad \mathrm{Var}[\mathbf{Z}_{ij1}] \ = \ \mathrm{Var}\big[(\pi_{11}\mathbf{V}_1)(\pi_{11}\mathbf{V}_1)^\top\big] \ . \tag{D.64}$$

When $d = 1$, the problem reduces to finding a distribution on non-negative reals given the first two moments, and one can choose the gamma distribution. When $d > 1$, a natural guess of a distribution on non-negative matrices is the non-central Wishart distribution. Unfortunately, one cannot form a non-central Wishart distribution given any mean and variance on $\mathbb{M}^d$, as illustrated in Lemma D.33.

Lemma D.33. *Let $d = 1$. There exists random variable V with $\mathbb{E}V^2 = 1$ and $\mathrm{Var}V^2 = 4$, but there is no non-central Wishart random variable W with $\mathbb{E}W = 1$ and $\mathrm{Var}W = 5$.*

Proof. Recall that $V \sim \Gamma(\alpha, \nu)$ has $\mathbb{E}V^2 = \frac{\alpha(\alpha+1)}{\nu^2}$ and $\mathbb{E}V^4 = \frac{\alpha(\alpha+1)(\alpha+2)(\alpha+3)}{\nu^4}$. Choose $\alpha = \frac{\sqrt{6}}{2}$ and $\nu = \sqrt{\frac{3+\sqrt{6}}{2}}$ gives

$$\mathbb{E}V^2 \ = \ \frac{\sqrt{6}(\sqrt{6}+2)/4}{(3+\sqrt{6})/2} \ = \ 1 \ , \qquad \mathbb{E}V^4 \ = \ \frac{(\sqrt{6}+4)(\sqrt{6}+6)/4}{(3+\sqrt{6})/2} \ = \ 5 \ ,$$

which gives the desired mean and variance for V^2. On the other hand, when $d = 1$, the non-central Wishart distribution is exactly non-central chi-squared distribution parametrised by the degree of freedom m and mean μ and variance σ^2 of the individual Gaussians. We can form the non-central Wishart random variable W by drawing $Z_1, \ldots, Z_m \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and defining

$$W \ \coloneqq \ \textstyle\sum_{l=1}^m (\mu + \sigma Z_l)^2 \ .$$

Suppose $\mathbb{E}[W] = 1$ and $\mathrm{Var}[W] = 4$. This implies

$$m(\mu^2 + \sigma^2) \ = \ 1 \ , \qquad m(4\mu^2\sigma^2 + 2\sigma^4) \ = \ 4 \ .$$

Write $x = \sigma^2$ and $\mu^2 = \frac{1}{m} - x$, we get $m\big(4\big(\frac{1}{m} - x\big)x + 2x^2\big) = 4$, which rearranges to

$$x^2 - \frac{2}{m}x + \frac{2}{m} \;=\; 0\,. \tag{D.65}$$

LHS equals $(x - \frac{1}{m})^2 + \frac{2m-1}{m^2}$, which is strictly positive since m is a positive integer. Therefore there is no solution to (D.65) and hence no non-central Wishart random variable W with $\mathbb{E}W = 1$ and $\mathrm{Var}W = 5$. This finishes the proof. □

The choice $d = 1$ for the proof above is for simplicity and not necessity. Wishart distribution fails because of specific structure in its first two moments arisen from the outer product of Gaussian vectors, which may not satisfy the mean and variance required by (D.64). A different approach is to show existence of solution to the problem of moments via maximum entropy principle. In the case $\mathcal{D} = \mathbb{R}^d$, Gaussian distribution is a max entropy distribution that solves the problem of moments given mean and variance. In the case $\mathcal{D}$ is a closed subset of $\mathbb{R}^d$, the following result adapted from Ambrozie (2013) studies the problem of moments from an approximate maximum entropy principle:

Lemma D.34. [Adapted from Corollary 6(a-b) of Ambrozie (2013)] *Fix $\epsilon > 0$. Let $T \subseteq \mathbb{R}^d$ be a closed subset and define the multi-index set $I := \{i \in \mathbb{Z}_+^d \,|\, i_1 + \ldots + i_d \le 2\}$. Let $(g_i)_{i\in I}$ be a set of reals with $g_{\mathbf{0}} = 1$. Assume that there exist a probability measure p_U with Lebesgue density function f_U supported on T such that, for every $(i_1, \ldots, i_d) \in I$,*

$$\mathbb{E}_{\mathbf{U}\sim p_U}[|U_1^{i_1} \ldots U_d^{i_d}|] \;<\; \infty \qquad \textit{and} \qquad \mathbb{E}_{\mathbf{U}\sim p_U}[U_1^{i_1} \ldots U_d^{i_d}] \;=\; g_i\,. \tag{D.66}$$

Then, there exists a particular solution p_U^* of (D.66) with Lebesgue density f_U^* that maximizes the ϵ-entropy over all measures p with Lebesgue density f,

$$H_\epsilon(p, f) \;=\; -\mathbb{E}_{\mathbf{U}\sim p}[\log(f)] - \epsilon\mathbb{E}_{\mathbf{U}\sim p}\big[\|\mathbf{U}\|^3\big]\,.$$

We can now use Lemma D.34 to construct the surrogate variables $\mathbf{Z}_i$ in Proposition 6.8 if the distribution of $\phi_{11}\mathbf{X}_1^*$ admits a Lebesgue density function.

Proof for Proposition 6.8. Assume first that the distribution of $\phi_{11}\mathbf{X}_1^*$ admits a Lebesgue density function. Fix d, b. Note that $\mathcal{D}^k$ is closed since $\mathcal{D} = \mathbb{M}^d \times \mathbb{R}^{d\times b}$ is a product of two closed sets and therefore closed in $\mathbb{R}^{d\times d} \times \mathbb{R}^{d\times b}$. The distribution $p_{X;d,b}$ of $\Phi_1\mathbf{X}_1^*$ and its Lebesgue density $f_{X;d,b}$ then satisfy the assumption of Lemma D.34 with $T = \mathcal{D}^k$ and the condition (D.66) becoming a bounded moment condition together with

$$\mathbb{E}_{\mathbf{U}\sim p}[\mathbf{U}] \;=\; \mathbb{E}[\Phi_1\mathbf{X}_1^*] \qquad \text{and} \qquad \mathbb{E}_{\mathbf{U}\sim p}[\mathbf{U}^{\otimes 2}] \;=\; \mathbb{E}[(\Phi_1\mathbf{X}_1^*)^{\otimes 2}]\,. \tag{D.67}$$

Then by Lemma D.34, there exists a distribution $p_{Z;d,b}$ with Lebesgue density function $f_{Z;d,b}$ which maximizes the ϵ-entropy in Lemma D.34 while satisfying (D.67). For each fixed (d, b), taking $\mathbf{Z}_{i;d,b} \sim p_{Z;d,b}$ then gives a choice of the surrogate variables. If the

coordinates of $\mathbf{Z}_{i;d,b}$ are uniformly bounded as $O(d^{-1})$ almost surely as d grows with $b = O(d)$, we can apply Lemma D.32(iii) to yield the desired convergences, which finishes the proof. If either $\phi_{11}\mathbf{X}_1^*$ does not admit a Lebesgue density function or if there is no uniform bound over the coordinates of $\mathbf{Z}_{i;d,b}$ as $O(d^{-1})$, we take $\mathbf{Z}_i$ to be an i.i.d. copy of $\Phi_i\mathbf{X}_i^*$ which again gives the desired convergences but in a trivial manner. □

Simulation and proof for toy example In this section we focus on the toy model stated in Lemma 6.9, where $d = 1$ and

$$\mathbf{Y}_i \coloneqq \mathbf{V}_i \text{ where } \mathbf{V}_i \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2), \text{ and } \pi_{ij} = \tau_{ij} \text{ a.s. .} \tag{D.68}$$

Recall that we have taken the surrogate variables to be Gamma random variables. We now prove the convergence of variance and dependence of variance of estimate on the variance of data for the toy example in Lemma 6.9.

Proof of Lemma 6.9. To prove the first convergence statement, note that in 1d, $\mathbb{M}^1$ is the set of non-negative reals, and $\mathbf{Z}_i = \{\mathbf{Z}_{ij1}, \mathbf{Z}_{ij2}\}_{j\le k}$ takes values in $(\mathbb{M}^1\times\mathbb{R})^k = \mathcal{D}^k$ which agrees with the domain of data. Moreover, denoting $\mu_{V^2} \coloneqq \mathbb{E}[(\mathbf{V}_1)^2]$, the moments of $\mathbf{Z}_i$ satisfy

$$\begin{aligned}
\mathbb{E}[\mathbf{Z}_i] &= \mathbf{1}_{k\times 1} \otimes \begin{pmatrix} \mu_{V^2} \\ \mu_{V^2} \end{pmatrix}, = \mathbf{1}_{k\times 1} \otimes \begin{pmatrix} \mathbb{E}[(\pi_{11}\mathbf{V}_1)^2] \\ \mathbb{E}[(\tau_{11}\mathbf{Y}_1)^2] \end{pmatrix}, \\
\mathrm{Var}[\mathbf{Z}_i] &= \mathbf{1}_{k\times k} \otimes \begin{pmatrix} v_\pi & v_\pi \\ v_\pi & v_\pi \end{pmatrix} = \mathbf{1}_{k\times k} \otimes \begin{pmatrix} \mathrm{Cov}[(\pi_{11}\mathbf{V}_1)^2,(\pi_{12}\mathbf{V}_1)^2] & \mathrm{Cov}[(\pi_{11}\mathbf{V}_1)^2,(\tau_{12}\mathbf{Y}_1)^2] \\ \mathrm{Cov}[(\tau_{11}\mathbf{Y}_1)^2,(\pi_{12}\mathbf{V}_1)^2] & \mathrm{Cov}[(\tau_{11}\mathbf{Y}_1)^2,(\tau_{12}\mathbf{Y}_1)^2] \end{pmatrix}.
\end{aligned}$$

This corresponds to the mean and variance of $\mathbf{Z}_i^\delta$ in Lemma D.7 with δ set to 1. While the earlier result on ridge regression in Lemma D.32 does not apply directly, an analogous argument works by computing some additional mixed smoothness terms in Lemma D.7(ii). Recall from the proof of Lemma D.32 that for $d = 1$, $\nu_{r;m} = O(1)$ for $0 \le r \le 3$. Therefore by Lemma D.7(ii) with $\delta = 1$, the following convergences hold as $n, k \to \infty$:

$$\begin{aligned}
&d_{\mathcal{H}}(\sqrt{n}f(\Phi\mathcal{X}), \sqrt{n}f(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)) \\
&\quad = O\big((k^{-1/2} + n^{-1/2}k^{-1/2})c_1 + (n^{-1/2} + 3n^{-1} + n^{-3/2})(c_X + c_Z)\big) \;\to\; 0\,, \\
&\quad n\|\,\mathrm{Var}[f(\Phi\mathcal{X})] - \mathrm{Var}[f(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)]\| \;=\; O(k^{-1/2}c_1 + n^{-1}(c_X + c_Z)) \;\to\; 0\,.
\end{aligned}$$

For the second statement, we first note that

$$\begin{aligned}
S_Z \coloneqq \frac{1}{nk}\sum_{i=1}^n\sum_{j=1}^k \mathbf{Z}_{ij1} &= \frac{1}{nk}\sum_{i=1}^n\sum_{j=1}^k \mathbf{Z}_{ij2} \\
&= \frac{1}{n}\sum_{i=1}^n \mathbf{Z}_{i11} \sim \Gamma\big(\frac{n(\mu_{V^2})^2}{v_\pi}, \frac{n\mu_{V^2}}{v_\pi}\big)\,.
\end{aligned}$$

Then we can write the variance of R^Z in terms of S_Z:

$$\begin{aligned}
\mathrm{Var}[R^Z] &= \mathrm{Var}[\mathbb{E}[(\mathbf{V}_{new} - \hat{B}^Z\mathbf{V}_{new})^2|\hat{B}^Z]] \\
&= \mathrm{Var}[-2\mathbb{E}[\mathbf{V}_{new}^2]\hat{B}^Z + \mathbb{E}[\mathbf{V}_{new}^2](\hat{B}^Z)^2]
\end{aligned}$$

$$
\begin{aligned}
&= (\mu_{V^2})^2 \text{Var}[-2\hat{B}^Z + (\hat{B}^Z)^2] \\
&= (\mu_{V^2})^2 \text{Var}\Big[-2\frac{S_Z}{S_Z+\lambda} + \frac{(S_Z)^2}{(S_Z+\lambda)^2}\Big] \\
&= (\mu_{V^2})^2 \text{Var}\Big[\frac{-S_Z^2 - 2\lambda S_Z}{(S_Z+\lambda)^2}\Big] \\
&= (\mu_{V^2})^2 \text{Var}\Big[1 - \frac{S_Z^2 + 2\lambda S_Z}{(S_Z+\lambda)^2}\Big] \\
&= (\mu_{V^2})^2 \lambda^2 \text{Var}\Big[\frac{1}{(S_Z+\lambda)^2}\Big] \\
&= \mathbb{E}[\mathbf{V}_1^2]^2 \lambda^2 \text{Var}\Big[\frac{1}{(X_n(v)+\lambda)^2}\Big] \;=\; \sigma_n^2(v) \; .
\end{aligned}
$$

In the last line, we have denoted the random variable $X_n(v) \sim \Gamma(\frac{n(\mu_{V^2})^2}{v}, \frac{n\mu_{V^2}}{v})$ and recalled the definition of $\sigma_n(\nu)$, which is independent of k and the distribution of π_{ij}. This completes the proof. □

D.6.4. Departure from Taylor limit at higher dimensions

In Lemma D.32, we have shown convergences of the form

$$
\begin{aligned}
n(\text{Var}[R^{\Phi\mathcal{X}}] - \text{Var}[R^T]) &= O\big(n^{-1/2}d^7 + n^{-1}d^8\big) \; , \\
n(\text{Var}[R^{\Phi\mathcal{X}}] - \text{Var}[R^Z]) &= O\big(n^{-1}d^7\big) \; .
\end{aligned}
$$

While the bounds are not necessarily tight in terms of dimensions, they hint at different rates of convergences to the two limits. $\text{Var}[R^T]$ has a simple behavior under augmentations as discussed for plugin estimators in Section 6.4.3, and in particular is reduced when data is invariant under augmentations. On the other hand, $\text{Var}[R^Z]$ has a complex behavior under augmentations as discussed in Section 6.4.5. In the main text, the separation of convergence rates is illustrated by a simulation that shows complex dependence of variance of risk under augmentation at a moderately high dimension.

In this section we aim to find evidence for a non-trivial separation of the convergence rates by focusing on the following model: For positive constants $\sigma, \tilde{\lambda}$ independent of n and d, consider

$$
\mathbf{Y}_i \coloneqq \mathbf{V}_i \;\text{ where } \mathbf{V}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 \mathbf{1}_{d\times d}),\; \pi_{ij} = \tau_{ij} = \text{id a.s.},\; \text{ and } \lambda = d\tilde{\lambda} \; , \tag{D.69}
$$

where id is the identity map $\mathbb{R}^d \to \mathbb{R}^d$ and ψ is an increasing function describing the rate of growth as a function of d. The parameter λ is chosen to be $O(d)$ instead of $O(1)$ for this model so that the penalty does not vanish and the inverse in ridge regression stays well-defined as d grows to infinity. Focusing on a specific model allows us to have a tight bound in terms of dimensions. The following lemma characterizes the convergence behavior of $\text{Var}[R^{\Phi\mathcal{X}}]$ to $\text{Var}[R^T]$ and $\text{Var}[R^Z]$ in terms of a function depending on n.

Lemma D.35. *Assume (D.69). Let $\{Z_i\}_{i\le n}$ be i.i.d. non-negative random variables with mean 1, variance 2 and finite 6th moments, and define $\mathbf{Z}_i \coloneqq \{\sigma^2 Z_i \mathbf{1}_{d\times 1}, \sigma^2 Z_i \mathbf{1}_{d\times 1}\}_{j\le k}$. Then*

(i) for R^T defined on $\{\mathbf{Z}_i\}_{i\le n}$,

$$n|\mathrm{Var}[R^{\Phi\mathcal{X}}] - \mathrm{Var}[R^T]| \;=\; n\left|d^2\sigma^4\tilde{\lambda}^4\mathrm{Var}\left[\frac{1}{(\tilde{\lambda}+\sigma^2\chi_n^2/n)^2}\right] - \frac{8d^2\sigma^8\tilde{\lambda}^4}{n(\tilde{\lambda}+\sigma^2)^6}\right| ,$$

where χ_n^2 is a chi-squared distributed random variable with n degrees of freedom;

(ii) there exist a constant $C_1 > 0$ not depending on n and d and a quantity $C_2 = \Theta(1)$ as n, d grow such that

$$n|\mathrm{Var}[R^{\Phi\mathcal{X}}] - \mathrm{Var}[R^T]| \;\ge\; nd^2E(n)C_1 - n^{-1}d^2C_2 ,$$

where $E(n) \coloneqq \left|\mathbb{E}\left[\frac{(\chi_n^2-n)^3}{n^3(\tilde{\lambda}+\sigma^2\chi_\Delta^2/n)^4}\right]\right|$ and χ_Δ^2 is a random variable between χ_n^2 and n;

(iii) for R^Z defined on $\{\mathbf{Z}_i\}_{i\le n}$,

$$n|\mathrm{Var}[R^{\Phi\mathcal{X}}] - \mathrm{Var}[R^Z]| \;=\; O(n^{-1}d^2) .$$

In Lemma D.35, while $E(n)$ is a complicated function, if we compare it to $\mathbb{E}[n^{-3}(\chi_n^2-n)^3]$, we expect the term to be on the order $n^{-3/2}$ as n grows. A natural guess of the order of the first term in Lemma D.35 is $\Theta(n^{-1/2}d^2)$. This suggests that if $d = n^\alpha$ for some $\frac{1}{4} < \alpha < \frac{1}{2}$, we may have $n|\mathrm{Var}[R^{\Phi\mathcal{X}}] - \mathrm{Var}[R^T]|$ not converging to 0 while the convergence of $n|\mathrm{Var}[R^{\Phi\mathcal{X}}] - \mathrm{Var}[R^Z]|$ still holds due to Lemma D.35(iii). A simulation in Figure D.2 shows that this can indeed be the case in an example parameter regime: if $\{Z_i\}_{i\le n}$ in Lemma D.35 are Gamma random variables, $\mathrm{Var}[R^Z] = \mathrm{Var}[R^{\Phi\mathcal{X}}]$ exactly, whereas no matter how the distribution of $\{Z_i\}_{i\le n}$ are chosen, the gap between $\mathrm{Var}[R^{\Phi\mathcal{X}}]$ and $\mathrm{Var}[R^T]$ may not decay to zero as shown in Figure D.2. This suggests that for a moderately high dimension, it is most suitable to understand $\mathrm{Var}[R^{\Phi\mathcal{X}}]$ through $\mathrm{Var}[R^Z]$ instead of $\mathrm{Var}[R^T]$. This completes the discussion from Remark 6.3. It may be of interest to note that in Figure 6.5, the regime at which augmentation exhibits complex behavior despite invariance is when $d = 7$ and $n = 50$, i.e. when d is close to $n^{1/2}$.

The proof of Lemma D.35(i) is by a standard Taylor expansion argument followed by a careful lower bound. The essence of the proof of Lemma D.35(ii) is by applying Theorem 6.1 while considering the particular structure (D.69); we spell out the proof in full for clarity.

Proof of Lemma D.35(i). Denote $g_1(\Sigma) \coloneqq g_R(\Sigma, \Sigma)$ where g_R is as defined in (D.62) and $\mu_S \coloneqq \mathbb{E}[(\pi_{ij}\mathbf{V}_i)(\pi_{ij}\mathbf{V}_i)^\top] = \sigma^2\mathbf{1}_{d\times d}$. We first seek to simplify the expressions of

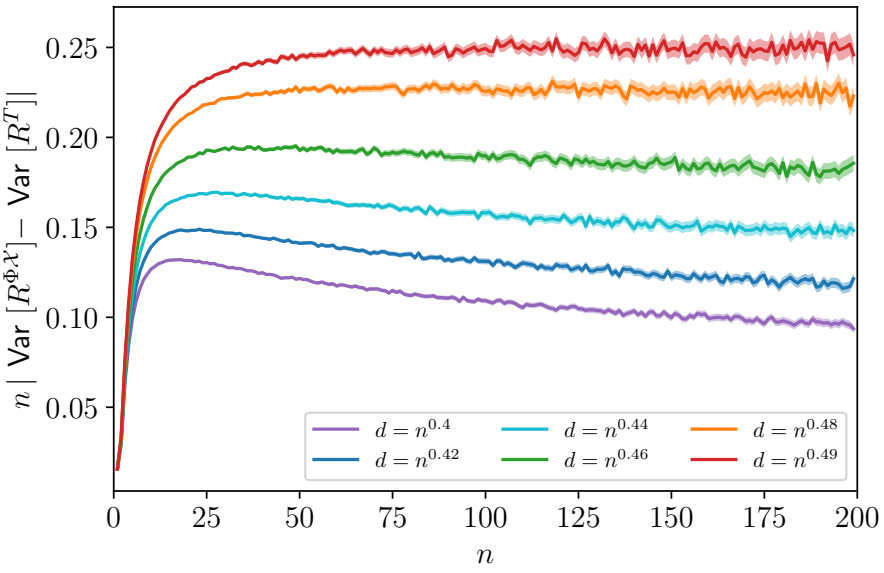


Figure D.2: Plot of difference in variances computed in Lemma D.35(i) against n for $\tilde{\lambda} = \sigma = 1$.

the variances:

$$\begin{aligned}\mathrm{Var}[R^T] &:= \mathrm{Var}\Big[g_R(\mu_S,\mu_S) + \partial g_R(\mu_S,\mu_S)\Big(\frac{1}{nk}\textstyle\sum_{i,j}\mathbf{Z}_{ij} - (\mu_S,\mu_S)\Big)\Big] \\ &= \mathrm{Var}\Big[\partial g_R(\mu_S,\mu_S)\Big(\frac{1}{nk}\textstyle\sum_{i,j}\mathbf{Z}_{ij} - (\mu_S,\mu_S)\Big)\Big] \,.\end{aligned}$$

Since $\mathbf{Z}_i$ matches the two moments of $\Phi_i\mathbf{X}_i = \{(\pi_{ij}\mathbf{V}_i)(\pi_{ij}\mathbf{V}_i)^\top, (\pi_{ij}\mathbf{V}_i)(\pi_{ij}\mathbf{V}_i)^\top\}_{j\le k}$ and $\{\mathbf{Z}_i\}_{i\le n}$ are i.i.d., we get that

$$\begin{aligned}\mathrm{Var}[R^T] &= \mathrm{Var}\Big[\partial g_R(\mu_S,\mu_S)\Big(\frac{1}{nk}\textstyle\sum_{i,j}(\pi_{ij}\mathbf{V}_i)(\pi_{ij}\mathbf{V}_i)^\top, (\pi_{ij}\mathbf{V}_i)(\pi_{ij}\mathbf{V}_i)^\top) - (\mu_S,\mu_S)\Big)\Big] \\ &= \mathrm{Var}\Big[\partial g_1(\mu_S)\Big(\frac{1}{nk}\textstyle\sum_{i,j}(\pi_{ij}\mathbf{V}_i)(\pi_{ij}\mathbf{V}_i)^\top - \mu_S)\Big)\Big] \,.\end{aligned}$$

Under (D.69), we can replace each $\pi_{ij}\mathbf{V}_i$ by $\sigma\xi_i\mathbf{1}_d$ where $\{\xi_i\}_{i\le n}$ are i.i.d. standard normal variables. Denote $\chi_n^2 := \sum_{i=1}^n \xi_i^2$. Then

$$\mathrm{Var}[R^T] = \mathrm{Var}\Big[\partial g_1(\mu_S)\Big(\frac{\sigma^2\chi_n^2}{n}\mathbf{1}_{d\times d}\Big)\Big] \,. \tag{D.70}$$

On the other hand,

$$R^{\Phi\mathcal{X}} = g_1\big(\frac{1}{nk}\textstyle\sum_{i,j}(\pi_{ij}\mathbf{V}_i)(\pi_{ij}\mathbf{V}_i)^\top\big) = g_1\big(\frac{\sigma^2\chi_n^2}{n}\mathbf{1}_{d\times d}\big) \,.$$

Given $\Sigma = x\mathbf{1}_{d\times d}$ for some $x > 0$, the explicit form of $g_1(\Sigma)$ and its derivative are given by Lemma D.36 as

$$g_1(\Sigma) = \frac{d\sigma^2\lambda^2}{(\lambda + dx)^2} \,, \qquad \partial g_1(\Sigma)\mathbf{1}_{d\times d} = -\frac{2d^2\sigma^2\lambda^2}{(\lambda+dx)^3} \,,$$

This implies

$$\mathrm{Var}[R^T] = \mathrm{Var}\Big[-\frac{2d^2\sigma^2\lambda^2}{(\lambda+d\sigma^2)^3}\frac{\sigma^2\chi_n^2}{n}\Big] = \frac{4d^4\sigma^8\lambda^4}{n^2(\lambda+d\sigma^2)^6}\mathrm{Var}[\chi_n^2] = \frac{8d^2\sigma^8\tilde{\lambda}^4}{n(\tilde{\lambda}+\sigma^2)^6} \,,$$

where we have used $\mathrm{Var}[\chi_n^2] = 2n$ and $\lambda = d\tilde{\lambda}$. Moreover

$$\begin{aligned}\mathrm{Var}[R^{\Phi\mathcal{X}}] &= \mathrm{Var}\Big[\frac{d\sigma^2\lambda^2}{(\lambda + d\sigma^2\chi_n^2/n)^2}\Big] = \mathrm{Var}\Big[\frac{d\sigma^2\tilde{\lambda}^2}{(\tilde{\lambda} + \sigma^2\chi_n^2/n)^2}\Big] \\ &= d^2\sigma^4\tilde{\lambda}^4\mathrm{Var}\Big[\frac{1}{(\tilde{\lambda} + \sigma^2\chi_n^2/n)^2}\Big] .\end{aligned}$$

Taking a difference and multiplying by n gives the desired result:

$$n|\mathrm{Var}[R^{\Phi\mathcal{X}}] - \mathrm{Var}[R^T]| = n\Big|d^2\sigma^4\tilde{\lambda}^4\mathrm{Var}\Big[\frac{1}{(\tilde{\lambda} + \sigma^2\chi_n^2/n)^2}\Big] - \frac{8d^2\sigma^8\tilde{\lambda}^4}{n(\tilde{\lambda} + \sigma^2)^6}\Big| .$$

□

Proof of Lemma D.35(ii). Note that a second-order Taylor expansion implies that almost surely there exists $\chi_\Delta^2 \in [n, \chi_n^2]$ such that

$$\begin{aligned}R^{\Phi\mathcal{X}} &= g_1\big(\frac{1}{nk}\textstyle\sum_{i,j}(\pi_{ij}\mathbf{V}_i)(\pi_{ij}\mathbf{V}_i)^\top\big) = g_1\big(\frac{\sigma^2\chi_n^2}{n}\mathbf{1}_{d\times d}\big) \\ &= g_1(\sigma^2\mathbf{1}_{d\times d}) + \partial g_1(\sigma^2\mathbf{1}_{d\times d})\frac{\sigma^2(\chi_n^2 - n)}{n}\mathbf{1}_{d\times d} \\ &\quad + \frac{1}{2}\partial^2 g_1\big(\frac{\sigma^2\chi_\Delta^2}{n}\mathbf{1}_{d\times d}\big)\big(\frac{\sigma^4(\chi_n^2 - n)^2}{n^2}\big)(\mathbf{1}_{d\times d})^{\otimes 2} .\end{aligned}$$

This implies

$$\begin{aligned}\mathrm{Var}[R^{\Phi\mathcal{X}}] &= \mathrm{Var}\Big[\partial g_1(\mu_S)\frac{\sigma^2\chi_n^2}{n}\mathbf{1}_{d\times d} + \partial^2 g_1\big(\frac{\sigma^2\chi_\Delta^2}{n}\mathbf{1}_{d\times d}\big)\big(\frac{\sigma^4(\chi_n^2 - n)^2}{2n^2}\big)(\mathbf{1}_{d\times d})^{\otimes 2}\Big] \\ &= \mathrm{Var}\Big[\partial g_1(\mu_S)\frac{\sigma^2\chi_n^2}{n}\mathbf{1}_{d\times d}\Big] + \mathrm{Var}\Big[\partial^2 g_1\big(\frac{\sigma^2\chi_\Delta^2}{n}\mathbf{1}_{d\times d}\big)\big(\frac{\sigma^4(\chi_n^2 - n)^2}{2n^2}\big)(\mathbf{1}_{d\times d})^{\otimes 2}\Big] \\ &\quad + 2\mathrm{Cov}\Big[\partial g_1(\mu_S)\frac{\sigma^2(\chi_n^2 - n)}{n}\mathbf{1}_{d\times d}, \partial^2 g_1\big(\frac{\sigma^2\chi_\Delta^2}{n}\mathbf{1}_{d\times d}\big)\big(\frac{\sigma^4(\chi_n^2 - n)^2}{2n^2}\big)(\mathbf{1}_{d\times d})^{\otimes 2}\Big]\end{aligned}$$

where the first term equals $\mathrm{Var}[R^T]$ by (D.70). Therefore by the triangle inequality, the difference in the variances of $R^{\Phi\mathcal{X}}$ and R^T can be written as

$$\begin{aligned}&n|\mathrm{Var}[R^{\Phi\mathcal{X}}] - \mathrm{Var}[R^T]| \\ &\geq 2n\Big|\mathrm{Cov}\Big[\partial g_1(\sigma^2\mathbf{1}_{d\times d})\frac{\sigma^2(\chi_n^2 - n)}{n}\mathbf{1}_{d\times d}, \partial^2 g_1\big(\frac{\sigma^2\chi_\Delta^2}{n}\mathbf{1}_{d\times d}\big)\big(\frac{\sigma^4(\chi_n^2 - n)^2}{2n^2}\big)(\mathbf{1}_{d\times d})^{\otimes 2}\Big]\Big| \qquad \text{(D.71)} \\ &\quad - n\Big|\mathrm{Var}\Big[\partial^2 g_1\big(\frac{\sigma^2\chi_\Delta^2}{n}\mathbf{1}_{d\times d}\big)\big(\frac{\sigma^4(\chi_n^2 - n)^2}{2n^2}\big)(\mathbf{1}_{d\times d})^{\otimes 2}\Big]\Big| . \qquad \text{(D.72)}\end{aligned}$$

Given $\Sigma = x\mathbf{1}_{d\times d}$ for some $x > 0$, the explicit form of derivatives of $g_1(\Sigma)$ are given by Lemma D.36 as

$$\partial g_1(\Sigma)\mathbf{1}_{d\times d} = -\frac{2d^2\sigma^2\lambda^2}{(\lambda + dx)^3} , \qquad \partial^2 g_1(\Sigma)(\mathbf{1}_{d\times d})^{\otimes 2} = \frac{6d^3\sigma^2\lambda^2}{(\lambda + dx)^4} .$$

Note that $\mathbb{E}[\chi_n^2 - n] = 0$ and $\lambda = d\tilde{\lambda}$. The covariance term can be computed as

$$\begin{aligned}\text{(D.71)} &= 2n\Big|\mathrm{Cov}\Big[-\frac{2d^2\sigma^2\lambda^2}{(\lambda + d\sigma^2)^3}\frac{\sigma^2(\chi_n^2 - n)}{n}, \frac{6d^3\sigma^2\lambda^2}{(\lambda + d\sigma^2\chi_\Delta^2/n)^4}\frac{\sigma^4(\chi_n^2 - n)^2}{2n^2}\Big]\Big| \\ &= \frac{12nd^5\sigma^{10}\lambda^4}{(\lambda + d\sigma^2)^3}\Big|\mathrm{Cov}\Big[\frac{(\chi_n^2 - n)}{n}, \frac{(\chi_n^2 - n)^2}{n^2(\lambda + d\sigma^2\chi_\Delta^2/n)^4}\Big]\Big|\end{aligned}$$

$$= \frac{12nd^5\sigma^{10}\lambda^4}{(\lambda+d\sigma^2)^3}\Big|\mathbb{E}\Big[\frac{(\chi_n^2-n)^3}{n^3(\lambda+d\sigma^2\chi_\Delta^2/n)^4}\Big]\Big| \;=\; \frac{12nd^2\sigma^{10}\tilde{\lambda}^4}{(\tilde{\lambda}+\sigma^2)^3}\Big|\mathbb{E}\Big[\frac{(\chi_n^2-n)^3}{n^3(\tilde{\lambda}+\sigma^2\chi_\Delta^2/n)^4}\Big]\Big|$$
$$= \frac{12nd^2\sigma^{10}\tilde{\lambda}^4}{(\tilde{\lambda}+\sigma^2)^3}E(n) \;=\; nd^2C_1E(n)\;,$$

where $C_1 := \frac{12\sigma^{10}\tilde{\lambda}^4}{(\tilde{\lambda}+\sigma^2)^3}$ is a constant not depending on n and d as required. The minus-variance term can be bounded as

$$\begin{aligned}
\text{(D.72)} \;&=\; -n\Big|\mathrm{Var}\Big[\frac{6d^3\sigma^2\lambda^2}{(\lambda+d\frac{\sigma^2\chi_\Delta^2}{n})^4}\Big(\frac{\sigma^4(\chi_n^2-n)^2}{2n^2}\Big)\Big]\Big| \;=\; -n\Big|\mathrm{Var}\Big[\frac{3d\sigma^2\tilde{\lambda}^2}{(\tilde{\lambda}+\frac{\sigma^2\chi_\Delta^2}{n})^4}\Big(\frac{\sigma^4(\chi_n^2-n)^2}{n^2}\Big)\Big]\Big| \\
&\overset{(a)}{\geq} -\mathbb{E}\Big[\frac{9nd^2\sigma^4\tilde{\lambda}^4}{(\tilde{\lambda}+\frac{\sigma^2\chi_\Delta^2}{n})^8}\Big(\frac{\sigma^8(\chi_n^2-n)^4}{n^4}\Big)\Big] \\
&\overset{(b)}{\geq} -\frac{9nd^2\sigma^{12}}{\tilde{\lambda}^4}\mathbb{E}\Big[\frac{(\chi_n^2-n)^4}{n^4}\Big] \\
&\overset{(c)}{\geq} -n^{-1}d^2\frac{9\sigma^{12}}{\tilde{\lambda}^4}\Big(K_4\max\big\{n^{-1/4}\big(\|\xi_1^2-1\|_{L_4}^4\big)^{1/4},\big(\|\xi_1^2-1\|_{L_2}^2\big)^{1/2}\big\}\Big)^4 \\
&=:\; -n^{-1}d^2\,C_2\;.
\end{aligned}$$

where in (a) we have upper bounded variance with a second moment, in (b) we have note that $\chi_\Delta^2 \geq 0$ and in (c) we have used Rosenthal's inequality from Lemma D.21 to show that there exists a universal constant K_4 such that

$$\begin{aligned}
\mathbb{E}\Big[\frac{(\chi_n^2-n)^4}{n^4}\Big] \;&=\; \frac{1}{n^4}\big\|\textstyle\sum_{i=1}^n(\xi_i^2-1)\big\|_{L_4}^4 \\
&\leq\; \frac{1}{n^4}\Big(K_4\max\big\{\big(\textstyle\sum_{i=1}^n\|\xi_i^2-1\|_{L_4}^4\big)^{1/4},\big(\textstyle\sum_{i=1}^n\|\xi_i^2-1\|_{L_2}^2\big)^{1/2}\big\}\Big)^4\;. \\
&=\; \frac{1}{n^2}\Big(K_4\max\big\{n^{-1/4}\big(\|\xi_1^2-1\|_{L_4}^4\big)^{1/4},\big(\|\xi_1^2-1\|_{L_2}^2\big)^{1/2}\big\}\Big)^4 \;=\; \Theta(n^{-2})\;.
\end{aligned}$$

Therefore C_2 is $\Theta(1)$ as required, and we obtain the statement in (i) from the bounds on (D.71) and (D.72):

$$n|\mathrm{Var}[R^{\Phi\mathcal{X}}]-\mathrm{Var}[R^T]| \;\geq\; nd^2C_1E(n)-n^{-1}d^2C_2\;.$$

□

Proof of Lemma D.35(iii). Write $\omega_n^2 := \sum_{i=1}^n Z_i$. Note that

$$\begin{aligned}
&|\mathrm{Var}[R^{\Phi\mathcal{X}}]-\mathrm{Var}[R^Z]| \;=\; \Big|\mathrm{Var}\big[g_1\big(\tfrac{\sigma^2\chi_n^2}{n}\mathbf{1}_{d\times d}\big)\big]-\mathrm{Var}\big[g_1\big(\tfrac{\sigma^2\omega_n^2}{n}\mathbf{1}_{d\times d}\big)\big]\Big| \\
&\leq\; \Big|\mathbb{E}\big[g_1\big(\tfrac{\sigma^2\chi_n^2}{n}\mathbf{1}_{d\times d}\big)\big]-\mathbb{E}\big[g_1\big(\tfrac{\sigma^2\omega_n^2}{n}\mathbf{1}_{d\times d}\big)\big]\Big|\Big|\mathbb{E}\big[g_1\big(\tfrac{\sigma^2\chi_n^2}{n}\mathbf{1}_{d\times d}\big)\big]+\mathbb{E}\big[g_1\big(\tfrac{\sigma^2\omega_n^2}{n}\mathbf{1}_{d\times d}\big)\big]\Big| \\
&\quad+\Big|\mathbb{E}\big[g_1\big(\tfrac{\sigma^2\chi_n^2}{n}\mathbf{1}_{d\times d}\big)^2-g_1\big(\tfrac{\sigma^2\omega_n^2}{n}\mathbf{1}_{d\times d}\big)^2\big]\Big|\;. \qquad \text{(D.73)}
\end{aligned}$$

We aim to bound (D.73) by mimicking the proof of Theorem 6.1 but use tighter control on dimensions since we know the specific form of the estimator. Write

$$\bar{W}_i(w) \;:=\; \frac{1}{n}\Big(\textstyle\sum_{i'=1}^{i-1}\xi_{i'}^2+w+\sum_{i'=i+1}^n Z_{i'}\Big)\;,$$

and denote $D_i^r g_{1;i}(w) := \partial^r g_1\big(\frac{\sigma^2}{n}\bar{W}_i(w)\mathbf{1}_{d\times d}\big)$ for $r = 0, 1, 2, 3$. Then analogous to the proof of Theorem 6.1, by a third-order Taylor expansion around 0 and noting that the first two moments of ξ_i^2 and $\mathbf{Z}_{ij1}$ match, we obtain that

$$\begin{aligned}
&\Big|\mathbb{E}\big[g_1\big(\frac{\sigma^2\chi_n^2}{n}\mathbf{1}_{d\times d}\big)\big] - \mathbb{E}\big[g_1\big(\frac{\sigma^2\omega_n^2}{n}\mathbf{1}_{d\times d}\big)\big]\Big| \\
&= \Big|\sum\nolimits_{i=1}^n \mathbb{E}\big[g_1\big(\frac{\sigma^2}{n}\bar{W}_i(\xi_i^2)\mathbf{1}_{d\times d}\big) - g_1\big(\frac{\sigma^2}{n}\bar{W}_i(Z_i)\mathbf{1}_{d\times d}\big)\big]\Big| \\
&\le \sum\nolimits_{i=1}^n \mathbb{E}\Big[\sup_{w\in[0,\xi_i^2]}\Big|D_i^3 g_{1;i}(w)\frac{\sigma^6(\xi_i^2)^3}{n^3}(\mathbf{1}_{d\times d})^{\otimes 3}\Big| + \sup_{w\in[0,Z_i]}\Big|D_i^3 g_{1;i}(w)\frac{\sigma^6(Z_i)^3}{n^3}(\mathbf{1}_{d\times d})^{\otimes 3}\Big|\Big] .
\end{aligned} \tag{D.74}$$

Similarly,

$$\begin{aligned}
&\Big|\mathbb{E}\big[g_1\big(\frac{\sigma^2\chi_n^2}{n}\mathbf{1}_{d\times d}\big)^2 - g_1\big(\frac{\sigma^2\omega_n^2}{n}\mathbf{1}_{d\times d}\big)^2\big]\Big| \\
&\le 2\sum\nolimits_{i=1}^n \mathbb{E}\Big[\sup_{w\in[0,\xi_i^2]}\Big|\big(g_{1;i}(w)D_i^3 g_{1;i}(w) + D_i g_{1;i}(w) D_i^2 g_{1;i}(w)\big)\frac{\sigma^6(\xi_i^2)^3}{n^3}(\mathbf{1}_{d\times d})^{\otimes 3}\Big| \\
&\qquad + \sup_{w\in[0,Z_i]}\Big|\big(g_{1;i}(w)D_i^3 g_{1;i}(w) + D_i g_{1;i}(w) D_i^2 g_{1;i}(w)\big)\frac{\sigma^6(Z_i)^3}{n^3}(\mathbf{1}_{d\times d})^{\otimes 3}\Big|\Big] .
\end{aligned} \tag{D.75}$$

Given $\Sigma = x\mathbf{1}_{d\times d}$ for some $x > 0$, the explicit forms of $g_1(\Sigma)$ and its derivatives from Lemma D.36 imply that

$$\begin{aligned}
g_{1;i}(w) &= \frac{d\sigma^2\lambda^2}{(\lambda + d\frac{\sigma^2\bar{W}_i(w)}{n})^2} , & D_i g_{1;i}(w)\mathbf{1}_{d\times d} &= -\frac{2d^2\sigma^2\lambda^2}{(\lambda + d\frac{\sigma^2\bar{W}_i(w)}{n})^3} , \\
D_i^2 g_{1;i}(w)(\mathbf{1}_{d\times d})^{\otimes 2} &= \frac{6d^3\sigma^2\lambda^2}{(\lambda + d\frac{\sigma^2\bar{W}_i(w)}{n})^4} , & D_i^3 g_{1;i}(w)(\mathbf{1}_{d\times d})^{\otimes 3} &= -\frac{24d^4\sigma^2\lambda^2}{(\lambda + d\frac{\sigma^2\bar{W}_i(w)}{n})^5} .
\end{aligned}$$

Therefore, by noting $\lambda = d\tilde{\lambda}$, we get

$$\begin{aligned}
\text{(D.74)} &= 24n^{-3}d^4\sigma^8\lambda^2\sum\nolimits_{i=1}^n \mathbb{E}\Big[\sup_{w\in[0,\xi_i^2]}\Big|\frac{(\xi_i^2)^3}{(\lambda + \frac{d\sigma^2\bar{W}_i(w)}{n})^5}\Big| + \sup_{w\in[0,Z_i]}\Big|\frac{(Z_i)^3}{(\lambda + \frac{d\sigma^2\bar{W}_i(w)}{n})^5}\Big|\Big] \\
&\overset{(a)}{\le} 24n^{-3}d^4\sigma^8\lambda^{-3}\sum\nolimits_{i=1}^n \mathbb{E}[(\xi_i^2)^3 + Z_i^3] \\
&= 24n^{-2}d\sigma^8\tilde{\lambda}^{-3}\mathbb{E}[(\xi_1^2)^3 + Z_1^3] \;=\; O(n^{-2}d) .
\end{aligned}$$

where in (a) we have used that $\bar{W}_i(w) \ge 0$ almost surely for $w \in [0, \xi_i^2]$ and for $w \in [0, Z_i]$. By the same argument,

$$\begin{aligned}
\text{(D.75)} &= 72n^{-3}d^5\sigma^{10}\lambda^4\sum\nolimits_{i=1}^n \mathbb{E}\Big[\sup_{w\in[0,\xi_i^2]}\Big|\frac{(\xi_i^2)^3}{(\lambda + d\frac{\sigma^2\bar{W}_i(w)}{n})^7}\Big| + \sup_{w\in[0,Z_i]}\Big|\frac{(Z_i^2)^3}{(\lambda + d\frac{\sigma^2\bar{W}_i(w)}{n})^7}\Big|\Big] \\
&\le 72n^{-2}d^2\sigma^{10}\tilde{\lambda}^{-3}\mathbb{E}[(\xi_i^2)^3 + Z_i^3] \;=\; O(n^{-2}d^2) .
\end{aligned}$$

Moreover,

$$\Big|\mathbb{E}\big[g_1\big(\frac{\sigma^2\chi_n^2}{n}\mathbf{1}_{d\times d}\big)\big] + \mathbb{E}\big[g_1\big(\frac{\sigma^2\omega_n^2}{n}\mathbf{1}_{d\times d}\big)\big]\Big| \;=\; |\mathbb{E}[g_{1;n}(\xi_n^2) + g_{1;1}(Z_1)]| \;=\; O(d) . \tag{D.76}$$

Finally the above three bounds imply that

$$n|\mathrm{Var}[R^{\Phi\mathcal{X}}] - \mathrm{Var}[R^{Z}]| \;\le\; n\,(\text{D.73}) \;\le\; n\,(\text{D.74}) \times (\text{D.76}) + n\,(\text{D.75}) \;=\; O(n^{-1}d^2)\;,$$

which is the desired bound. □

Lemma D.36. *Consider $\Sigma = x\mathbf{1}_{d\times d}$ for some $x > 0$ and $g_1(\Sigma) \coloneqq g_R(\Sigma,\Sigma)$ where g_R is defined as in (D.62) under the model (D.69). Then, the following derivative formulas hold:*

$$g_1(\Sigma) \;=\; \frac{d\sigma^2\lambda^2}{(\lambda+dx)^2}\;, \qquad \partial g_1(\Sigma)\mathbf{1}_{d\times d} \;=\; -\frac{2d^2\sigma^2\lambda^2}{(\lambda+dx)^3}\;,$$
$$\partial^2 g_1(\Sigma)(\mathbf{1}_{d\times d})^{\otimes 2} \;=\; \frac{6d^3\sigma^2\lambda^2}{(\lambda+dx)^4}\;, \qquad \partial^3 g_1(\Sigma)(\mathbf{1}_{d\times d})^{\otimes 3} \;=\; -\frac{24d^4\sigma^2\lambda^2}{(\lambda+dx)^5}\;.$$

Proof. First note that

$$\mathbb{E}[\|\mathbf{Y}_{new}\|_2^2] \;=\; \mathbb{E}[\|\mathbf{V}_{new}\|_2^2] \;=\; \sigma^2 d\;, \quad \mathbb{E}[\mathbf{V}_{new}\mathbf{Y}_{new}^\top] \;=\; \mathbb{E}[\mathbf{V}_{new}\mathbf{V}_{new}^\top] \;=\; \sigma^2\mathbf{1}_{d\times d}\;,$$

which allows us to write

$$g_1(\Sigma) = \sigma^2 d - 2\sigma^2\sum\nolimits_{r,s=1}^{d} g_{B;rs}(\Sigma,\Sigma) + \sigma^2\sum\nolimits_{r,s,t=1}^{d} g_{B;rt}(\Sigma,\Sigma)g_{B;ts}(\Sigma,\Sigma),$$

where we have recalled the expression

$$g_{B;rs}(\Sigma,\Sigma) \;=\; \mathbf{e}_r^\top(\Sigma+\lambda\mathbf{I}_{d\times d})^{-1}\Sigma\mathbf{e}_s\;.$$

Denoting $\tilde{\Sigma} = (\Sigma+\lambda\mathbf{I}_{d\times d})^{-1} = (\Sigma+\lambda\mathbf{I}_d)^{-1}$, the partial derivative of $g_{B;rs}$ has been computed in the proof of Lemma D.32(i) as

$$\begin{aligned}\frac{\partial g_{B;rs}(\Sigma,\Sigma)}{\partial\Sigma_{r_1s_1}} &= -\mathbf{e}_r^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_1}\mathbf{e}_{s_1}^\top\tilde{\Sigma}^{-1}\Sigma\mathbf{e}_s + \mathbf{e}_r^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_1}\mathbb{I}_{\{s=s_1\}}\\ &= \mathbf{e}_r^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_1}\mathbf{e}_{s_1}^\top\big(-\tilde{\Sigma}^{-1}\Sigma+\tilde{\Sigma}^{-1}\tilde{\Sigma}\big)\mathbf{e}_s \;=\; \psi(d)\tilde{\lambda}\mathbf{e}_r^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_1}\mathbf{e}_{s_1}^\top\tilde{\Sigma}^{-1}\mathbf{e}_s\;,\end{aligned}$$

Similarly

$$\begin{aligned}\frac{\partial^2 g_{B;rs}(\Sigma,\Sigma)}{\partial\Sigma_{r_1s_1}\partial\Sigma_{r_2s_2}} &= \sum\nolimits_{l_1,l_2\in\{1,2\};\,l_1\neq l_2}\Big(\mathbf{e}_r\tilde{\Sigma}^{-1}\mathbf{e}_{r_{l_1}}\mathbf{e}_{s_{l_1}}^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_{l_2}}\mathbf{e}_{s_{l_2}}^\top\tilde{\Sigma}^{-1}\Sigma\mathbf{e}_s\\ &\qquad - \mathbf{e}_r^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_{l_1}}\mathbf{e}_{s_{l_1}}^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_{l_2}}\mathbb{I}_{\{s=s_{l_2}\}}\Big)\\ &= -\psi(d)\tilde{\lambda}\sum\nolimits_{l_1,l_2\in\{1,2\};\,l_1\neq l_2}\mathbf{e}_r\tilde{\Sigma}^{-1}\mathbf{e}_{r_{l_1}}\mathbf{e}_{s_{l_1}}^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_{l_2}}\mathbf{e}_{s_{l_2}}^\top\tilde{\Sigma}^{-1}\mathbf{e}_s\;,\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^3 g_{B;rs}(\Sigma,\Sigma)}{\partial\Sigma_{r_1s_1}\partial\Sigma_{r_2s_2}\partial\Sigma_{r_3s_3}} &= -\sum\nolimits_{\substack{l_1,l_2,l_3\in\{1,2,3\}\\ l_1,l_2,l_3\text{ distinct}}}\Big(\mathbf{e}_r^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_{l_1}}\mathbf{e}_{s_{l_1}}^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_{l_2}}\mathbf{e}_{s_{l_2}}^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_{l_3}}\mathbf{e}_{s_{l_3}}^\top\tilde{\Sigma}^{-1}\Sigma\mathbf{e}_s\\ &\qquad - \mathbf{e}_r^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_{l_1}}\mathbf{e}_{s_{l_1}}^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_{l_2}}\mathbf{e}_{s_{l_2}}^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_{l_3}}\mathbb{I}_{\{s=s_{l_3}\}}\Big)\\ &= \psi(d)\tilde{\lambda}\sum\nolimits_{\substack{l_1,l_2,l_3\in\{1,2,3\}\\ l_1,l_2,l_3\text{ distinct}}}\mathbf{e}_r^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_{l_1}}\mathbf{e}_{s_{l_1}}^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_{l_2}}\mathbf{e}_{s_{l_2}}^\top\tilde{\Sigma}^{-1}\mathbf{e}_{r_{l_3}}\mathbf{e}_{s_{l_3}}^\top\tilde{\Sigma}^{-1}\mathbf{e}_s\;.\end{aligned}$$

On the other hand, since $\Sigma = x\mathbf{1}_{d\times d}$, a calculation gives

$$\tilde{\Sigma}^{-1} \;=\; (x\mathbf{1}_{d\times d} + \lambda \mathbf{I}_d)^{-1} \;=\; \frac{1}{\lambda(\lambda+dx)}\big((\lambda+dx)\mathbf{I}_d - x\mathbf{1}_{d\times d}\big)\;, \tag{D.77}$$

in which case, denoting $J_{r,s}(x) \coloneqq (\mathbb{I}_{\{r=s\}}(\lambda+(d-1)x) - \mathbb{I}_{\{r\neq s\}}x)$, we have

$$\begin{aligned}
g_{B;rs}(\Sigma,\Sigma) \;&=\; \frac{x}{\lambda\,(\lambda+dx)}\big((\lambda+dx)-dx\big) \;=\; \frac{x}{\lambda+dx}\;,\\
\frac{\partial g_{B;rs}(\Sigma,\Sigma)}{\partial \Sigma_{r_1s_1}} \;&=\; \frac{J_{r,r_1}(x)J_{s,s_1}(x)}{\lambda\,(\lambda+dx)^2}\;,\\
\frac{\partial^2 g_{B;rs}(\Sigma,\Sigma)}{\partial \Sigma_{r_1s_1}\partial\Sigma_{r_2s_2}} \;&=\; -\sum\nolimits_{l_1,l_2\in\{1,2\};\, l_1\neq l_2} \frac{J_{r,r_{l_1}}(x)J_{s_{l_1},r_{l_2}}(x)J_{s_{l_2},s}(x)}{\lambda^2\,(\lambda+dx)^3}\;,\\
\frac{\partial^3 g_{B;rs}(\Sigma,\Sigma)}{\partial \Sigma_{r_1s_1}\partial\Sigma_{r_2s_2}\partial\Sigma_{r_3s_3}} \;&=\; \sum\nolimits_{\substack{l_1,l_2,l_3\in\{1,2,3\}\\ l_1,l_2,l_3 \text{ distinct}}} \frac{J_{r,r_{l_1}}(x)J_{s_{l_1},r_{l_2}}(x)J_{s_{l_2},r_{l_3}}(x)J_{s_{l_3},s}(x)}{\lambda^3\,(\lambda+dx)^4}\;.
\end{aligned}$$

Note that $J_{r,s}(x) = J_{s,r}(x)$ and $\sum_{r=1}^d J_{r,s}(x) = \lambda$. These formulas and the above derivatives imply that

$$\begin{aligned}
g_1(\Sigma) \;&=\sigma^2 d - 2\sigma^2\sum\nolimits_{r,s=1}^d g_{B;rs}(\Sigma,\Sigma) + \sigma^2\sum\nolimits_{r,s,t=1}^d g_{B;rt}(\Sigma,\Sigma)g_{B;ts}(\Sigma,\Sigma)\\
&=\sigma^2 d - 2\frac{\sigma^2 x d^2}{\lambda+dx} + \frac{\sigma^2x^2d^3}{(\lambda+dx)^2} \;=\; \frac{d\sigma^2\lambda^2}{(\lambda+dx)^2}\;,\\
\partial g_1(\Sigma)\mathbf{1}_{d\times d} \;&=\; \sum\nolimits_{r_1,s_1=1}^d \frac{\partial g_1(\Sigma)}{\partial\Sigma_{r_1s_1}}\\
&=\; -2\sigma^2\sum\nolimits_{r,s,r_1,s_1}\frac{\partial g_{B;rs}(\Sigma,\Sigma)}{\partial\Sigma_{r_1s_1}} + 2\sigma^2\sum\nolimits_{r,s,t,r_1,s_1}\frac{\partial g_{B;rt}(\Sigma,\Sigma)}{\partial\Sigma_{r_1s_1}}g_{B;ts}(\Sigma,\Sigma)\\
&=\; -\frac{2d^2\sigma^2\lambda}{(\lambda+dx)^2} + \frac{2d^3\sigma^2x\lambda}{(\lambda+dx)^3} \;=\; -\frac{2d^2\sigma^2\lambda^2}{(\lambda+dx)^3}\;,
\end{aligned}$$

$$\begin{aligned}
\partial^2 g_1(\Sigma)(\mathbf{1}_{d\times d})^{\otimes 2} \;&=\; \sum\nolimits_{r_1,s_1,r_2,s_2=1}^d \frac{\partial^2 g_1(\Sigma)}{\partial\Sigma_{r_1s_1}\partial\Sigma_{r_2s_2}}\\
=\; -2\sigma^2\sum\nolimits_{r,s,r_1,s_1,r_2,s_2}&\frac{\partial^2 g_{B;rs}(\Sigma,\Sigma)}{\partial\Sigma_{r_1s_1}\partial\Sigma_{r_2s_2}} + 2\sigma^2\sum\nolimits_{r,s,t,r_1,s_1,r_2,s_2}\frac{\partial^2 g_{B;rt}(\Sigma,\Sigma)}{\partial\Sigma_{r_1s_1}\partial\Sigma_{r_2s_2}}g_{B;ts}(\Sigma,\Sigma)\\
+\,2\sigma^2\sum\nolimits_{r,s,t,r_1,s_1,r_2,s_2}&\frac{\partial g_{B;rt}(\Sigma,\Sigma)}{\partial\Sigma_{r_1s_1}}\frac{\partial g_{B;ts}(\Sigma,\Sigma)}{\partial\Sigma_{r_2s_2}}\\
= \frac{4d^3\sigma^2\lambda}{(\lambda+dx)^3} - \frac{4d^4\sigma^2\lambda x}{(\lambda+dx)^4} &+ \frac{2d^3\sigma^2\lambda^2}{(\lambda+dx)^4} \;=\; \frac{6d^3\sigma^2\lambda^2}{(\lambda+dx)^4}\;,
\end{aligned}$$

$$\begin{aligned}
\partial^3 g_1(\Sigma)(\mathbf{1}_{d\times d})^{\otimes 3} \;&=\; -2\sigma^2\sum\nolimits_{r,s,r_1,s_1,r_2,s_2,r_3,s_3}\frac{\partial^3 g_{B;rs}(\Sigma,\Sigma)}{\partial\Sigma_{r_1s_1}\partial\Sigma_{r_2s_2}\partial\Sigma_{r_3s_3}}\\
&\quad+2\sigma^2\sum\nolimits_{r,s,t,r_1,s_1,r_2,s_2,r_3,s_3}\frac{\partial^3 g_{B;rt}(\Sigma,\Sigma)}{\partial\Sigma_{r_1s_1}\partial\Sigma_{r_2s_2}\partial\Sigma_{r_3s_3}}g_{B;ts}(\Sigma,\Sigma)\\
&\quad+6\sigma^2\sum\nolimits_{r,s,t,r_1,s_1,r_2,s_2,r_3,s_3}\frac{\partial^2 g_{B;rt}(\Sigma,\Sigma)}{\partial\Sigma_{r_1s_1}\partial\Sigma_{r_2s_2}}\frac{\partial g_{B;ts}(\Sigma,\Sigma)}{\partial\Sigma_{r_3s_3}}\\
&=\; -\frac{12d^4\sigma^2\lambda}{(\lambda+dx)^4} + \frac{12d^5\sigma^2\lambda x}{(\lambda+dx)^5} - \frac{12d^4\sigma^2\lambda^2}{(\lambda+dx)^5} \;=\; -\frac{24d^4\sigma^2\lambda^2}{(\lambda+dx)^5}\;,
\end{aligned}$$

which completes the proof. □

D.6.5. Maximum of exponentially many correlated random variables

The challenge in Proposition 6.13 is that $f(\mathbf{x}_{11:nk}) := \max_{1\le l\le d_n} \frac{1}{nk}\sum_{i\le n}\sum_{j\le k} x_{ijl} \in \mathbb{R}^{d_n}$ is that the statistic is non-smooth and we also need to have careful control to deal with growing dimensions. We employ Corollary D.9, a result that adapts Theorem 4.1 to this setting by introducing smooth approximating functions $f^{(t)}$. In this section, we first propose an appropriate choice of $f^{(t)}$ to yield a suitable bound similar to Theorem 4.1. The use of Corollary D.9 introduces additional moment terms of f to be controlled as n, d_n grow, for which the bounds are obtained via a martingale difference argument in Lemma D.38. Finally, putting the results together allow us to compute a bound for $d_{\mathcal{H}}$ and for difference in variances in both the augmented and unaugmented cases.

The function of interest can be written as $f(\mathbf{x}_{11},\dots,\mathbf{x}_{nk}) = g\big(\frac{1}{nk}\sum_{i,j}\mathbf{x}_{ij}\big)$ where

$$g(\mathbf{x}) \;:=\; \max_{1\le s\le d_n} x_s \in \mathbb{R}\,. \tag{D.78}$$

We first propose the choice of approximating function g_t for g and present its approximating quality in terms of t and explicit forms for its derivatives.

Lemma D.37. *Consider $g:\mathbb{R}^{d_n}\to\mathbb{R}$ defined in (D.78). Define, for $t>0$,*

$$g^{(t)}(\mathbf{x}) \;:=\; \frac{\log\big(\sum_{s\le d_n} e^{t\log(d_n)x_s}\big)}{t\log(d_n)}\,.$$

Then for every t, $g^{(t)}$ is infinitely differentiable, and $|g^{(t)}(\mathbf{x}) - g(\mathbf{x})| \le \frac{1}{t}$. Moreover, defining

$$\omega_l(\mathbf{x}) \;:=\; \frac{\exp(t\log(d_n)x_l)}{\sum_{s=1}^{d_n}\exp(t\log(d_n)x_s)}\,,$$

the derivatives of g_t are given by

(i) $\frac{\partial}{\partial x_l} g^{(t)}(\mathbf{x}) = \omega_l(\mathbf{x})$,

(ii) $\frac{\partial^2}{\partial x_{l_2}\partial x_{l_1}} g^{(t)}(\mathbf{x}) = -(t\log(d_n))\,\omega_{l_2}(\mathbf{x})\,\omega_{l_1}(\mathbf{x}) + \mathbb{I}_{\{l_1=l_2\}}(t\log(d_n))\,\omega_{l_1}(\mathbf{x})$,

(iii)
$$\begin{aligned}&\frac{\partial^3}{\partial x_{l_3}\partial x_{l_2}\partial x_{l_1}} g^{(t)}(\mathbf{x})\\ =&(t\log(d_n))^2\,\omega_{l_3}(\mathbf{x})\,\omega_{l_2}(\mathbf{x})\,\omega_{l_1}(\mathbf{x}) - \mathbb{I}_{\{l_2=l_3\}}(t\log(d_n))^2\,\omega_{l_2}(\mathbf{x})\,\omega_{l_1}(\mathbf{x})\\ &- \mathbb{I}_{\{l_1=l_3\}}(t\log(d_n))^2\,\omega_{l_2}(\mathbf{x})\,\omega_{l_1}(\mathbf{x}) - \mathbb{I}_{\{l_1=l_2\}}(t\log(d_n))^2\,\omega_{l_2}(\mathbf{x})\,\omega_{l_1}(\mathbf{x})\\ &+ \mathbb{I}_{\{l_1=l_2=l_3\}}(t\log(d_n))^2\omega_{l_1}(\mathbf{x}_{11:nk})\,.\end{aligned}$$

In particular, this implies that $\tilde{\nu}_r^{(t)}$ defined in Corollary D.9 with respect to $g^{(t)}$ satisfy

$$\tilde{\nu}_1^{(t)} \le 1\,,\qquad \tilde{\nu}_2^{(t)} \le 2t\log(d_n)\,,\qquad \tilde{\nu}_3^{(t)} \le 5t^2\log(d_n)^2\,.$$

Proof. $g^{(t)}$ is infinitely differentiable as it is a composition of infinitely differentiable

functions. The approximation error is given by

$$|g(\mathbf{x}) - g^{(t)}(\mathbf{x})| = \left| \frac{\log\left(\sum_{l=1}^{d_n} \exp(t\log(d_n)x_l)\right) - \log\left(\exp(t\log(d_n)\max_{r\le d_n} x_r)\right)}{t\log(d_n)} \right|$$
$$= \left| \frac{\log\left(\sum_{l=1}^{d_n} \exp\left(t\log(d_n)(x_l - \max_{r\le d_n} M_r)\right)\right)}{t\log(d_n)} \right| \overset{(a)}{\le} \frac{\log(d_n)}{t\log(d_n)} = \frac{1}{t}.$$

The inequality at (a) is obtained by noting that $\exp\left(t\log(d_n)(x_l - \max_{r\le d_n} x_r)\right) \in (0,1]$ and that it attains 1 for some l. The sum inside the logarithm therefore lies in $[1, d_n]$.

The derivatives are obtained by repeated applications of chain rule.

For the final identities, note that $\sum_{l=1}^{d_n} \omega_l(\mathbf{x}) = 1$ and $\omega_l(\mathbf{x}) \in [0,1]$. Therefore we get

$$\left\| \frac{\partial}{\partial \mathbf{x}} g^{(t)}(\mathbf{x}) \right\|_1 = \left| \sum_{l=1}^{d_n} \omega_l(\mathbf{x}) \right| = 1 ,$$

for any $\mathbf{x} \in \mathbb{R}^{d_n}$. Similarly for the second and third derivatives,

$$\left\| \frac{\partial^2}{\partial \mathbf{x}^2} g^{(t)}(\mathbf{x}) \right\|_1 \le 2t\log(d_n) , \qquad \left\| \frac{\partial^3}{\partial \mathbf{x}^3} g^{(t)}(\mathbf{x}) \right\|_1 \le 5t^2 \log(d_n)^2 .$$

Recall that for $g^{(t)}$ with output dimension 1, the noise stability terms are given by

$$\tilde{\nu}_r^{(t)} = \max_{i\le n} \zeta_{i;12}\left(\left\| \frac{\partial}{\partial \mathbf{x}^r} g^{(t)}\left(\overline{\mathbf{W}}_i(\bullet)\right) \right\|_1 \right) ,$$

Since the bounds above apply to $\|\frac{\partial}{\partial \mathbf{x}^r} g^{(t)}(\overline{\mathbf{W}}_i(\bullet))\|_1$ almost surely, we obtain the desired bounds for $\tilde{\nu}_1^{(t)}$, $\tilde{\nu}_2^{(t)}$ and $\tilde{\nu}_3^{(t)}$. □

Recall that in Theorem D.8, the terms $\|f(\Phi\mathcal{X})\|_{L_2}$ and $\|f(\mathcal{Z})\|_{L_2}$ are introduced in the bound and needs to be controlled. Bounding the moment of a maximum of exponentially many correlated coordinates is made possible by the following lemma, which makes use of Rosenthal's inequality for a martingale difference sequence (Dharmadhikari et al., 1968a).

Lemma D.38. *Consider i.i.d. zero-mean random vectors $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ in $\mathbb{R}^{d_n}$. Denote $\mathbf{Y}_{i,l}$ as the l^{th} coordinate of $\mathbf{Y}_i$ for $l \le d_n$. For any $m \ge 3$, if $M_m := \|\max_{l\le d_n} |\mathbf{Y}_{1,l}|\|_{L_m} < \infty$, then there exists a constant C_m that does not depend on d_n or $(\mathbf{Y}_i)_{i\le n}$ such that*

$$\left\| \max_{l\le d_n} \frac{1}{n} \sum_{i\le n} \mathbf{Y}_{i,l} \right\|_{L_m} \le \inf_{\nu\in\mathbb{R}} \left[2n^{-(1-\nu)} + \log(d_n) n^{-\nu}\, M_2^2 + n^{-1/2} C_m M_m \right] .$$

In particular, this implies that for $f(\Phi\mathcal{X})$ and $f(\mathcal{Z})$ defined in Proposition 6.13, if $\log(d_n) = o(n^a)$ for some $a \ge 0$, then

$$\tilde{\nu}_0^{(t)} = o(n^{-(1-a)/2} + t^{-1}) , \qquad \|f(\Phi\mathcal{X})\|_{L_2}, \|f(\mathcal{Z})\|_{L_2} = o(n^{-(1-a)/2}) .$$

Proof. The general idea is to apply the triangle inequality to the following quantities:

$$\left| \mathbb{E}\left[\max_{l\le d_n} \frac{1}{n} \sum_{i\le n} \mathbf{Y}_{i,l} \right] \right| , \quad \left\| \max_{l\le d_n} \frac{1}{n} \sum_{i\le n} \mathbf{Y}_{i,l} - \left| \mathbb{E}\left[\max_{l\le d_n} \frac{1}{n} \sum_{i\le n} \mathbf{Y}_{i,l} \right] \right| \right\|_{L_m} ,$$

which are controlled separately. The first quantity is controlled by a second-order Taylor expansion on a smooth approximating function of the maximum, and the second quantity is controlled by martingale bounds.

The first step is to bound $\mathbb{E}\big[\max_{l\le d_n}\frac{1}{n}\sum_{i\le n}\mathbf{Y}_{i,l}\big]$. For $\mathbf{y}_1,\dots,\mathbf{y}_n\in\mathbb{R}^{d_n}$, consider the following function which can be expressed in terms of $g^{(t)}$ from Lemma D.37:

$$F_{t;\alpha}(\mathbf{y}_1,\dots,\mathbf{y}_n) \;:=\; \frac{\log\big(\sum_{l\le d_n}\exp(t\log(d_n)\frac{1}{n^\alpha}\sum_{i\le n}y_{il})\big)}{n^{1-\alpha}t\log(d_n)} \;=\; n^{-(1-\alpha)}g^{(t)}\big(\frac{1}{n^\alpha}\sum_{i\le n}\mathbf{y}_i\big).$$

For g defined in Lemma D.37, $F_{t;\alpha}$ then satisfies

$$\begin{aligned}\Big|F_{t;\alpha}(\mathbf{y}_1,\dots,\mathbf{y}_n)-\max_{l\le d_n}\frac{1}{n}\sum_{i\le n}y_{il}\Big| \;&=\; \frac{1}{n^{1-\alpha}}\Big|g^{(t)}\Big(\frac{1}{n^\alpha}\sum_{i\le n}\mathbf{y}_i\Big)-g\Big(\frac{1}{n^\alpha}\sum_{i\le n}\mathbf{y}_i\Big)\Big|\\ &\le\; \frac{1}{n^{1-\alpha}t}\,,\end{aligned}$$

and recall the intermediate bounds in the proof of Lemma D.37 that for any $\mathbf{y}\in\mathbb{R}^{d_n}$,

$$\|\partial g^{(t)}(\mathbf{y})\| \;\le\; 1\,,\qquad \|\partial^2 g^{(t)}(\mathbf{y})\| \;\le\; 2t\log(d_n)\,,\qquad \|\partial^3 g^{(t)}(\mathbf{y})\| \;\le\; 5t^2\log(d_n)^2\,.$$

Therefore

$$\big|\mathbb{E}[F_{t;\alpha}(\mathbf{Y}_1,\dots,\mathbf{Y}_n)]-\mathbb{E}\big[\max_{l\le d_n}\frac{1}{n}\sum_{i\le n}\mathbf{Y}_{i,l}\big]\big\| \;\le\; \frac{1}{n^{1-\alpha}t}\,,$$

and by expanding a telescoping sum followed by a Taylor expansion, we get

$$\begin{aligned}&|\mathbb{E}[F_{t;\alpha}(\mathbf{Y}_1,\dots,\mathbf{Y}_n)]| \;=\; \frac{1}{n^{1-\alpha}}\big|\mathbb{E}\big[g^{(t)}\big(\frac{1}{n^\alpha}\sum_{i\le n}\mathbf{Y}_i\big)\big]\big|\\
&\le \frac{1}{n^{1-\alpha}}|g^{(t)}(\mathbf{0})|+\frac{1}{n^{1-\alpha}}\sum_{i=1}^n\Big|\mathbb{E}\Big[g^{(t)}\big(\frac{1}{n^\alpha}\sum_{i'=1}^{i}\mathbf{Y}_{i'}\big)-g^{(t)}\big(\frac{1}{n^\alpha}\sum_{i'=1}^{i-1}\mathbf{Y}_{i'}\big)\Big]\Big|\\
&\le \frac{1}{n^{1-\alpha}t}+\frac{1}{n}\sum_{i\le n}\big|\mathbb{E}\big[\partial g^{(t)}\big(\frac{1}{n^\alpha}\sum_{i'=1}^{i-1}\mathbf{Y}_{i'}\big)\mathbf{Y}_i\big]\big|\\
&\quad+\frac{1}{2n^{1+\alpha}}\sum_{i\le n}\big|\mathbb{E}\big[\sup_{\mathbf{z}\in[\mathbf{0},\mathbf{Y}_i]}\partial^2 g^{(t)}\big(\frac{1}{n^\alpha}\sum_{i'=1}^{i-1}\mathbf{Y}_{i'}+\mathbf{z}\big)\mathbf{Y}_i\mathbf{Y}_i^\top\big]\big|\\
&\overset{(a)}{\le} \frac{1}{n^{1-\alpha}t}+\frac{1}{2n^{1+\alpha}}\sum_{i\le n}\mathbb{E}\big[\sup_{\mathbf{z}\in[\mathbf{0},\mathbf{Y}_i]}\|\partial^2 g^{(t)}\big(\frac{1}{n^\alpha}\sum_{i'=1}^{i-1}\mathbf{Y}_{i'}+\mathbf{z}\big)\|_1(\max_{l\le d_n}|\mathbf{Y}_{i,l}|)^2\big]\\
&\overset{(b)}{\le} \frac{1}{n^{1-\alpha}t}+\frac{t\log(d_n)}{n^\alpha}\,\|\max_{l\le d_n}|\mathbf{Y}_{1,l}|\|_{L_2}^2 \;=\; \frac{1}{n^{1-\alpha}t}+\frac{t\log(d_n)}{n^\alpha}\,M_2^2\,.\end{aligned}$$

To get (a), we have used the fact that $(\mathbf{Y}_i)_{i\le n}$ are i.i.d. with $\mathbb{E}[\mathbf{Y}_i]=0$ followed by applying Hölder's inequality. To get (b), we have used the bounds on the second derivative of g_t from Lemma D.37. Now by triangle inequality and taking $t=n^{\alpha-\nu}$, we obtain a bound on the mean of the maximum as

$$\begin{aligned}\Big|\mathbb{E}\Big[\max_{l\le d_n}\frac{1}{n}\sum_{i\le n}\mathbf{Y}_{i,l}\Big]\Big| \;&\le\; \inf_{\nu\in\mathbb{R}}\Big[\frac{2}{n^{1-\alpha}n^{\alpha-\nu}}+\frac{\log(d_n)}{n^\alpha n^{-\alpha+\nu}}\,M_2^2\Big]\\ &=\; \inf_{\nu\in\mathbb{R}}\big[2n^{-(1-\nu)}+\log(d_n)n^{-\nu}\,M_2^2\big]\,.\end{aligned}\tag{D.79}$$

The second step is to control

$$\big\|\max_{l\le d_n}\frac{1}{n}\sum_{i\le n}\mathbf{Y}_{i,l}-\mathbb{E}\big[\max_{l\le d_n}\frac{1}{n}\sum_{i\le n}\mathbf{Y}_{i,l}\big]\big\|_{L_m}\,.\tag{D.80}$$

Define the filtrations $(\mathcal{F}_i)_{i=0}^n$ by $\mathcal{F}_i \coloneqq \sigma(\mathbf{Y}_1, \ldots, \mathbf{Y}_i)$, and consider the martingale $(\mathbf{S}_i)_{i=0}^n$ with respect to $(\mathcal{F}_i)_{i=0}^n$ defined by $\mathbf{S}_0 \coloneqq 0$ and, for $1 \le i \le n$,

$$\mathbf{S}_i \;\coloneqq\; \mathbb{E}\big[\max_{l\le d_n} \frac{1}{n}\sum_{j\le n} \mathbf{Y}_{j,l}\big|\mathcal{F}_i\big] - \mathbb{E}\big[\max_{l\le d_n} \frac{1}{n}\sum_{j\le n} \mathbf{Y}_{j,l}\big] \;.$$

Note that the quantity to be controlled in (D.80) is exactly $\mathbf{S}_n$. We also define the martingale difference sequence $(\mathbf{D}_i)_{i=1}^n$ by

$$\mathbf{D}_i \coloneqq \mathbf{S}_i - \mathbf{S}_{i-1} = \mathbb{E}\big[\max_{l\le d_n} \frac{1}{n}\sum_{j\le n} \mathbf{Y}_{j,l}\big|\mathcal{F}_i\big] - \mathbb{E}\big[\max_{l\le d_n} \frac{1}{n}\sum_{j\le n} \mathbf{Y}_{j,l}\big|\mathcal{F}_{i-1}\big].$$

Then by a bound on moments of martingales (Dharmadhikari et al., 1968a), there exist constants C'_m that do not depend on n, d_n or $(\mathbf{Y}_i)_{i\le n}$ such that

$$\text{(D.80)} \;=\; \|\mathbf{S}_n\|_{L_m} \;\le\; C'_m n^{1/2}\big(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[|\mathbf{D}_i|^m\big]\big)^{1/m} \;. \tag{D.81}$$

Moreover, by independence between $\mathbf{Y}_i$ and $\{\mathbf{Y}_j\}_{j\neq i}$, we have that almost surely

$$\mathbb{E}\big[\max_{l\le d_n} \sum_{j\neq i} \mathbf{Y}_{j,l}\big|\mathcal{F}_i\big] \;=\; \mathbb{E}\big[\max_{l\le d_n} \sum_{j\neq i} \mathbf{Y}_{j,l}\big|\mathcal{F}_{i-1}\big] \;. \tag{D.82}$$

To control the martingale differences, we note that $n\,\mathbf{D}_i$ satisfies, almost surely,

$$\begin{aligned}
n\,\mathbf{D}_n \;&\le \mathbb{E}\Big[\max_{l\le d_n} \mathbf{Y}_{i,l} + \max_{l\le d_n}\sum_{j\neq i} \mathbf{Y}_{j,l}\Big|\mathcal{F}_i\Big] - \mathbb{E}\Big[\max_{l\le d_n}\sum_{j\le n} \mathbf{Y}_{j,l}\Big|\mathcal{F}_{i-1}\Big] \\
&\overset{\text{(D.82)}}{=} \max_{l\le d_n} \mathbf{Y}_{i,l} + \mathbb{E}\Big[\max_{l\le d_n}\sum_{j\neq i} \mathbf{Y}_{j,l}\Big|\mathcal{F}_{i-1}\Big] - \mathbb{E}\Big[\max_{l\le d_n}\sum_{j\le n} \mathbf{Y}_{j,l}\Big|\mathcal{F}_{i-1}\Big] \\
&\le \max_{l\le d_n} \mathbf{Y}_{i,l} + \mathbb{E}\Big[\max_{l\le d_n}\sum_{j\le n} \mathbf{Y}_{j,l} + \max_{l\le d_n}(-\mathbf{Y}_{i,l})\Big|\mathcal{F}_{i-1}\Big] \\
&\quad - \mathbb{E}\Big[\max_{l\le d_n}\sum_{j\le n} \mathbf{Y}_{j,l}\Big|\mathcal{F}_{i-1}\Big] \\
&= \max_{l\le d_n} \mathbf{Y}_{i,l} + \mathbb{E}\big[\max_{l\le d_n}(-\mathbf{Y}_{i,l})\big] \\
&\le \max_{l\le d_n} |\mathbf{Y}_{i,l}| + \mathbb{E}[\max_{l\le d_n} |\mathbf{Y}_{i,l}|\,] \;.
\end{aligned}$$

Similarly by expanding the $\mathbb{E}[\,\bullet\,|\mathcal{F}_{i-1}]$ term, almost surely,

$$\begin{aligned}
-n\,\mathbf{D}_n \;&\le \mathbb{E}\Big[\max_{l\le d_n} \mathbf{Y}_{i,l} + \max_{l\le d_n}\sum_{j\neq i} \mathbf{Y}_{j,l}\Big|\mathcal{F}_{i-1}\Big] - \mathbb{E}\Big[\max_{l\le d_n}\sum_{j\le n} \mathbf{Y}_{j,l}\Big|\mathcal{F}_i\Big] \\
&\overset{\text{(D.82)}}{=} \mathbb{E}\big[\max_{l\le d_n} \mathbf{Y}_{i,l}\big] + \mathbb{E}\Big[\max_{l\le d_n}\sum_{j\neq i} \mathbf{Y}_{j,l}\Big|\mathcal{F}_i\Big] - \mathbb{E}\Big[\max_{l\le d_n}\sum_{j\le n} \mathbf{Y}_{j,l}\Big|\mathcal{F}_i\Big] \\
&= \mathbb{E}\big[\max_{l\le d_n} \mathbf{Y}_{i,l}\big] + \mathbb{E}\Big[\max_{l\le d_n}\sum_{j\le n} \mathbf{Y}_{j,l} + \max_{l\le d_n}(-\mathbf{Y}_{i,l})\Big|\mathcal{F}_i\Big] \\
&\quad - \mathbb{E}\Big[\max_{l\le d_n}\sum_{j\le n} \mathbf{Y}_{j,l}\Big|\mathcal{F}_i\Big] \\
&= \mathbb{E}\big[\max_{l\le d_n} \mathbf{Y}_{i,l}\big] + \max_{l\le d_n}(-\mathbf{Y}_{i,l}) \\
&\le \mathbb{E}\big[\max_{l\le d_n} |\mathbf{Y}_{i,l}|\big] + \max_{l\le d_n} |\mathbf{Y}_{i,l}| \;.
\end{aligned}$$

This implies the m^{th} moment of $|n\,\mathbf{D}_i|$ can be bounded by

$$\begin{aligned}
\mathbb{E}[|n\,\mathbf{D}_i|^m] \;&\le \mathbb{E}\big[\big(\mathbb{E}\big[\max_{l\le d_n} |\mathbf{Y}_{i,l}|\big] + \max_{l\le d_n} |\mathbf{Y}_{i,l}|\big)^m\big] \\
&= \sum_{r=1}^m \binom{m}{r}\, \mathbb{E}\big[\max_{l\le d_n} |\mathbf{Y}_{i,l}|\big]^r \mathbb{E}\big[\big(\max_{l\le d_n} |\mathbf{Y}_{i,l}|\big)^{m-r}\big] \\
&\le 2^m \mathbb{E}\big[\big(\max_{l\le d_n} |\mathbf{Y}_{i,l}|\big)^m\big] \;=\; 2^m (M_m)^m \;.
\end{aligned}$$

Substituting into (D.81) yields the bound on (D.80):

$$\text{(D.80)} \;\le C'_m n^{-1/2}\big(\tfrac{1}{n}\textstyle\sum_{i=1}^n \mathbb{E}\big[|n\mathbf{D}_i|^m\big]\big)^{1/m} \;\le\; n^{-1/2}C_m M_m \;,$$

where we have defined $C_m \coloneqq 2C'_m$. Therefore by applying triangular inequality together with (D.79), the bound on the mean of a maximum, we get

$$\big\|\max_{l\le d_n}\tfrac{1}{n}\textstyle\sum_{i\le n}\mathbf{Y}_{i,l}\big\|_{L_m} \;\le\; \inf_{\nu\in\mathbb{R}}\big[2n^{-(1-\nu)} + \log(d_n)n^{-\nu}\,M_2^2 + n^{-1/2}C_m M_m\big] \;,$$

which is the required bound. For the final bound, note that if we choose $\mathbf{Y}_i$ above to be $\frac{1}{k}\sum_{j=1}^k \phi_{ij}\mathbf{X}_i$ from Proposition 6.13, $(\mathbf{Y}_i)_{i=1}^n$ are indeed i.i.d. zero-mean, and

$$M_6 \;=\; \Big\|\max_{l\le d_n}\frac{1}{k}\textstyle\sum_{j=1}^k(\phi_{1j}\mathbf{X}_1)_l\Big\|_{L_6} \;\le\; \Big\|\max_{l\le d_n}(\phi_{11}\mathbf{X}_1)_l\Big\|_{L_6} \overset{(d)}{<} \infty \;,$$

where (d) is by the assumption in Proposition 6.13. Then applying the bound above and plugging in $\log(d_n) = o(n^a)$ gives

$$\begin{aligned}\|f(\Phi\mathcal{X})\|_{L_2} \;\le \|f(\Phi\mathcal{X})\|_{L_6} \;&=\; \big\|\max_{l\le d_n}\tfrac{1}{n}\textstyle\sum_{i\le n}\mathbf{Y}_{i,l}\big\|_{L_6} \\ &= o\big(\inf_{\nu\in\mathbb{R}}\big[2n^{-(1-\nu)} + n^{a-\nu}M_6^2 + n^{-1/2}C_2M_6\big]\big) \\ &\overset{(e)}{=} o\big(2n^{-\frac{1-a}{2}} + n^{-\frac{1-a}{2}}M_2^2 + n^{-1/2}C_2M_2\big) \;=\; o(n^{-(1-a)/2}) \;,\end{aligned}$$

where we have set $\nu = \frac{1+\alpha}{2}$ in (e). This is the desired bound for $\|f(\Phi\mathcal{X})\|_{L_2}$. Applying the exact same argument to $f(\mathcal{Z})$ yields the same bound:

$$\|f(\mathcal{Z})\|_{L_2} \;\le \|f(\mathcal{Z})\|_{L_6} \;=\; o(n^{-(1-a)/2}) \;.$$

By definition of $\tilde{\nu}_0^{(t)}$,

$$\tilde{\nu}_0^{(t)} \;=\; O(\max\{\|f(\Phi\mathcal{X})\|_{L_6}, \|f(\mathcal{Z})\|_{L_6}\} + \varepsilon(t)) \;=\; o(n^{-(1-a)/2} + t^{-1}) \;.$$

□

We are now ready to prove Proposition 6.13.

Proof of Proposition 6.13. Consider the approximation function g_t defined in Lemma D.37, which satisfies the condition of Corollary D.9 with approximation quality $\varepsilon(t) = \frac{1}{t}$. By assumption we have $\log(d_n) = o(n^{1/10})$. Take $t = o(n^{11/20})$. Then by Lemma D.38, $\|f(\Phi\mathcal{X})\|_{L_2}, \|f(\mathcal{Z})\|_{L_2} = o(n^{-9/20})$ and

$$\tilde{\nu}_0^{(t)} \;=\; o(n^{-9/20} + t^{-1}) = o(n^{-9/20}) \;,$$

and by Lemma D.37,

$$\tilde{\nu}_1^{(t)} \le 1 \;, \qquad \tilde{\nu}_2^{(t)} \le 2t\log(d_n) \;=\; o(n^{13/20}) \;, \qquad \tilde{\nu}_3^{(t)} \le 5t^2\log(d_n)^2 \;=\; o(n^{24/20}) \;.$$

Moreover, by the assumption $\|\max_{l\le d_n}|(\phi_{11}\mathbf{X}_1)_l|\|_{L_6}, \|\max_{l\le d_n}|(\mathbf{Z}_{11})_l|\|_{L_6} < \infty$, the

moment terms in Corollary D.9 satisfy

$$\begin{aligned}\tilde{c}_X &= \frac{1}{6}\sqrt{\mathbb{E}\big[\max_{l\le d_n}|(\phi_{ij}\mathbf{X}_i)_l|^6\big]} \;=\; O(1)\;,\\ \tilde{c}_Z &= \frac{1}{6}\sqrt{\mathbb{E}\Big[\frac{1}{k}\sum_{j=1}^k \max_{l\le d_n}|(\mathbf{Z}_{1j})_l|^6\Big]} \overset{(a)}{=} \sqrt{\mathbb{E}[\max_{l\le d}|(\mathbf{Z}_{1j})_l|^6]}\\ &\overset{(b)}{=} O\big(\|\max_{l\le d_n}|(\phi_{ij}\mathbf{X}_i)_l|\|_{L_2}^3\,(\log d_n)^{3/2}\big) \;=\; o(n^{3/20})\;,\end{aligned}$$

In (a) we have used that the moment conditions (6.4) and the fact that $\mathbf{Z}_1$ is Gaussian implies that $\mathbf{Z}_{11},\dots,\mathbf{Z}_{1k}$ are identically distributed. In (b) we have used Lemma D.23 to bound the moment of maximum of a Gaussian. Therefore by Corollary D.9, we get that

$$\begin{aligned}d_{\mathcal{H}}(\sqrt{n}f(\Phi\mathcal{X}),\sqrt{n}f(\mathcal{Z})) &= o\big((n^{-1/2}(\tilde{\nu}_1^{(t)})^3 + 3n^{-1}\tilde{\nu}_1^{(t)}\tilde{\nu}_2^{(t)} + n^{-3/2}\tilde{\nu}_3^{(t)})(\tilde{c}_X+\tilde{c}_Z)\\ &\qquad + \sqrt{n}\varepsilon(t)\big)\\ &= o\big(n^{-7/20}+n^{-4/20}+n^{-3/20}+n^{-1/20}\big) \xrightarrow{n\to\infty} 0\;,\end{aligned}$$

and

$$\begin{aligned}n|\mathrm{Var}[f(\Phi\mathcal{X})]-\mathrm{Var}[f(\mathcal{Z})]| &= o\big(n^{-1}(\tilde{\nu}_0^{(t)}\tilde{\nu}_3^{(t)}+\tilde{\nu}_1^{(t)}\tilde{\nu}_2^{(t)})(\tilde{c}_X+\tilde{c}_Z)\\ &\qquad + n(\|f(\Phi\mathcal{X})\|_{L_2}+\|f(\mathcal{Z})\|_{L_2})\varepsilon(t)+n\varepsilon(t)^2\big)\\ &= o\big(n^{-2/20}+n^{-4/20}+1+n^{-2/20}\big) \xrightarrow{n\to\infty} 0\;,\end{aligned}$$

which are the desired convergences. □

D.6.6. Derivation of examples: softmax ensemble

In this section, we examine the effect of augmentation on the softmax ensemble estimator

$$f(\mathbf{x}_{11},\dots,\mathbf{x}_{nk}) \;:=\; \sum_{r=1}^{m_n}\beta_r\frac{\exp\big(-t\frac{\log(m_n)}{nk}\sum_{i=1}^n\sum_{j=1}^k L(\beta_r,\mathbf{x}_{ij})\big)}{\sum_{s=1}^{m_n}\exp\big(-t\frac{\log(m_n)}{nk}\sum_{i=1}^n\sum_{j=1}^k L(\beta_s,\mathbf{x}_{ij})\big)} \quad\text{for}\quad \mathbf{x}_{ij}\in\mathbb{R}^d\;.$$

The $\log(m_n)$ scaling is justified in Lemma D.39. To prove Proposition 6.14 about the effect of augmentations on testing data size, we consider the modified augmentations $\pi_{ij},\tau_{ij}:\mathbb{R}^d\to\mathbb{R}^{m_n}$ defined by

$$\begin{aligned}\pi_{ij}(\mathbf{X}_i) &:= \big(L(\beta_1,\phi_{ij}\mathbf{X}_i),\dots,L(\beta_{m_n},\phi_{ij}\mathbf{X}_i)\big)\;,\\ \tau_{ij}(\mathbf{X}_i) &:= \Big(\frac{1}{c_\phi}(L(\beta_1,\mathbf{X}_i)-\mu_1)+\mu_1,\dots,\frac{1}{c_\phi}(L(\beta_{m_n},\mathbf{X}_i)-\mu_{m_n})+\mu_{m_n}\Big)\;.\end{aligned}\tag{D.83}$$

Then defining the function $g:\mathbb{R}^{m_n}\to\mathbb{R}^p$ by

$$g(\mathbf{x}) \;:=\; \sum_{r=1}^{m_n}\beta_r\frac{\exp(-t\log(m_n)\,x_r)}{\sum_{s=1}^{m_n}\exp(-t\log(m_n)\,x_s)}\;,\tag{D.84}$$

we can write the two quantities of interest as

$$f(\Phi\mathcal{X}) \;=\; g\big(\frac{1}{nk}\sum_{i=1}^n\sum_{j=1}^k\pi_{ij}(\mathbf{X}_i)\big)\;,\quad f^*(\mathcal{X}) \;=\; g\big(\frac{1}{nk}\sum_{i=1}^n\sum_{j=1}^k\tau_{ij}(\mathbf{X}_i)\big)\;.\tag{D.85}$$

This allows us to invoke Corollary D.10 to study convergence of each estimator to their first-order Taylor expansions and provide an explicit formula of c_ϕ:

$$c_\phi \;:=\; \frac{(\partial g(\mu))^\top \big(\frac{1}{n}\Sigma_{11}\big)\partial g(\mu)}{(\partial g(\mu))^\top \big(\frac{1}{nk}\Sigma_{11} + \frac{k-1}{nk}\Sigma_{12}\big)\partial g(\mu)} \;\geq\; 1\;. \tag{D.86}$$

The fact that Corollary D.10 is adapted for high dimensions is what allows m_n to grow exponentially in n.

In the proof of Proposition 6.14, we first compare $f(\Phi\mathcal{X})$ to a first order Taylor expansion $f^T(\mathcal{Z})$ that is equal in distribution to $f^T(\tilde{\mathcal{Z}}^\phi)$ for some surrogate variables $\tilde{\mathcal{Z}}^\phi$ to be specified, which is in turn compared to $f^*(\mathcal{X})$ defined in Proposition 6.14.

Proof of Proposition 6.14. Recall from (D.84) that

$$g(\mathbf{x}) \;:=\; \textstyle\sum_{r=1}^{m_n} \beta_r \frac{\exp(-t\log(m_n)\,x_r)}{\sum_{s=1}^{m_n}\exp(-t\log(m_n)\,x_s)}\;.$$

Denote $w_r := \frac{\exp(-t\log(m_n)\,x_r)}{\sum_{s=1}^{m_n}\exp(-t\log(m_n)\,x_s)}$, and use $(\beta)_l$ to denote the lth coordinate of a $\mathbb{R}^p$ vector. To apply Corollary D.10 to g, we first compute the partial derivatives of each lth coordinate of g as

$$\begin{aligned}
\frac{\partial g_l(\mathbf{x})}{\partial x_{r_1}} \;&=\; -\,t\log(m_n)(\beta_{r_1})_l \frac{w_{r_1}}{\sum_{s=1}^{m_n} w_s} + t\log(m_n)\textstyle\sum_{s'=1}^{m_n}(\beta_{s'})_l \frac{w_{s'}w_{r_1}}{(\sum_{s=1}^{m_n} w_s)^2}\;,\\
\frac{\partial^2 g_l(\mathbf{x})}{\partial x_{r_2}\partial x_{r_1}} \;&= t^2\log(m_n)^2(\beta_{r_1})_l \frac{w_{r_1}}{\sum_{s=1}^{m_n} w_s}\mathbb{I}_{\{r_1=r_2\}} - t^2\log(m_n)^2(\beta_{r_1})_l \frac{w_{r_1}w_{r_2}}{(\sum_{s=1}^{m_n} w_s)^2}\\
&\quad - t^2\log(m_n)^2(\beta_{r_2})_l \frac{w_{r_2}w_{r_1}}{(\sum_{s=1}^{m_n} w_s)^2} - t^2\log(m_n)^2\textstyle\sum_{r'=1}^{m_n}(\beta_{r'})_l \frac{w_{r'}w_{r_1}}{(\sum_{s=1}^{m_n} w_s)^2}\mathbb{I}_{\{r_1=r_2\}}\\
&\quad + 2t^2\log(m_n)^2\textstyle\sum_{s'=1}^{m_n}(\beta_{s'})_l \frac{w_{s'}w_{r_1}w_{r_2}}{(\sum_{s=1}^{m_n} w_s)^3}\;.
\end{aligned}$$

Since $w_r \geq 0$ for all $r \leq m_n$, we get by triangle inequality and Cauchy-Schwarz that

$$\begin{aligned}
\Big\|\frac{\partial g_l(\mathbf{x})}{\partial \mathbf{x}}\Big\|_1 \;&\leq t\log(m_n)\textstyle\sum_{r_1=1}^{m_n}\Big(\big|(\beta_{r_1})_l\big|\Big|\frac{w_{r_1}}{\sum_{s=1}^{m_n} w_s}\Big| + \sum_{s'=1}^{m_n}\big|(\beta_{s'})_l\big|\Big|\frac{w_{s'}w_{r_1}}{(\sum_{s=1}^{m_n} w_s)^2}\Big|\Big)\\
&\leq t\log(m_n)(\sup_{r\leq m_n}|(\beta_r)_l|)\Big(\frac{\sum_{r_1=1}^{m_n} w_{r_1}}{\sum_{s=1}^{m_n} w_s} + \frac{\sum_{r_1,s'=1}^{m_n} w_{s'}w_{r_1}}{(\sum_{s=1}^{m_n} w_s)^2}\Big)\\
&= 2t\log(m_n)(\sup_{r\leq m_n}|(\beta_r)_l|)\;,
\end{aligned}$$

and similarly

$$\Big\|\frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x}^2}\Big\|_1 \;\leq 6t^2\log(m_n)^2(\sup_{r\leq m_n}|(\beta_r)_l|)\;.$$

Recall that $\tilde{\kappa}_r := \sum_{l\leq p}\big\|\sup_{\mathbf{w}\in[\mathbf{0},\bar{\mathbf{X}}]}\big\|\partial^r g_l(\mu+\mathbf{w})\big\|_1\big\|_{L_6}$. Since $\log m_n = o(n^{1/9})$, t is fixed and $\sum_{l=1}^p(\sup_{r\leq m_n}|(\beta_r)_l|)$ is assumed to be $O(1)$, the above bounds imply

$$\begin{aligned}
\tilde{\kappa}_1 \;&\leq 2t\log(m_n)\textstyle\sum_{l=1}^p(\sup_{r\leq m_n}|(\beta_r)_l|) \;=\; o(n^{1/9})\;,\\
\tilde{\kappa}_2 \;&\leq 6t^2\log(m_n)^2\textstyle\sum_{l=1}^p(\sup_{r\leq m_n}|(\beta_r)_l|) \;=\; o(n^{2/9})\;,
\end{aligned}$$

where we have used the assumption $\sum_{l=1}^{p}(\max_{r\le m_n}|(\beta_r)_l|) = O(1)$. Now recall from (D.85) that

$$f(\Phi\mathcal{X}) \;=\; g\big(\tfrac{1}{nk}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{k}\pi_{ij}(\mathbf{X}_i)\big)\;,$$

where π_{ij} is as defined in (D.83). Take the surrogate variables $\mathcal{Z}$ to be $\{\mathbf{Z}_i\}_{i\le n}$, where $\mathbf{Z}_2,\ldots,\mathbf{Z}_n$ are i.i.d. copies of the Gaussian vector $\mathbf{Z}_1$ specified in the statement in Proposition 6.14. The moment terms in Corollary D.10 can be bounded as

$$\begin{aligned}
\tilde{c}_X \;&=\frac{1}{6}\sqrt{\mathbb{E}\big[\max_{r\le m_n}|(\pi_{11}(\mathbf{X}_i))_l|^6\big]} \;=\; \frac{1}{6}\big\|\max_{r\le m_n}|L(\beta_r,\phi_{11}\mathbf{X}_1)|\big\|_{L_6}^3 \overset{(a)}{=} O(1)\;,\\
\tilde{c}_Z \;&=\frac{1}{6}\sqrt{\mathbb{E}\Big[\frac{1}{k}\textstyle\sum_{j=1}^{k}\max_{r\le m_n}|(\mathbf{Z}_{1j})_r|^6\Big]} \overset{(b)}{=} \frac{1}{6}\sqrt{\mathbb{E}\big[\max_{r\le m_n}|(\mathbf{Z}_{11})_r|^6\big]}\\
&\overset{(c)}{=} O\big(\big\|\max_{r\le m_n}|L(\beta_r,\phi_{11}\mathbf{X}_1)|\big\|_{L_2}^3(\log m_n)^{3/2}\big) \;=\; o(n^{1/6})\;,
\end{aligned}$$

where (a) is by assumption of Proposition 6.14 and (b) is by noting that the Gaussianity of $\mathbf{Z}_1$ and its specified moments imply that $\mathbf{Z}_{11},\ldots,\mathbf{Z}_{1k}$ are identically distributed. In (c), we have used Lemma D.23. By Corollary D.10 and that $\log m_n = o(n^{1/9})$, we get that

$$d_{\mathcal{H}}(\sqrt{n}f(\Phi\mathcal{X}),\sqrt{n}f^T(\mathcal{Z})) \;=o\big(n^{-1/2+1/9}\tilde{\kappa}_2\big)+O\big(n^{-1/2}\tilde{\kappa}_1^3\tilde{c}_Z\big) \;=\; o(1)\;, \tag{D.87}$$

$$n\big\|\mathrm{Var}[f(\Phi\mathcal{X})]-\mathrm{Var}\big[f^T(\mathcal{Z})\big]\big\| \;=o\big(n^{-1/2+1/6}\tilde{\kappa}_1\tilde{\kappa}_2+n^{-1+2/9}\tilde{\kappa}_2^2\big) \;=\; o(1)\;. \tag{D.88}$$

Next we compare $f^T(\mathcal{Z})$ to $f^T(\tilde{\mathcal{Z}}^\phi)$ for some $\tilde{\mathcal{Z}}^\phi$. Writing $\mu := \{\mu_r\}_{r\le m_n} = \mathbb{E}[\pi_{11}(\mathbf{X}_1)] = \mathbb{E}[\{L(\beta_r,\phi_{11}\mathbf{X}_1)\}_{r\le m_n}] = \mathbb{E}[\mathbf{Z}_{11}]$, we can express

$$f^T(\mathcal{Z}) \;=\; g(\mu)+(\partial g(\mu))^\top\Big(\frac{1}{nk}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{k}\mathbf{Z}_{ij}-\mu\Big)\;,$$

where we have viewed $\partial g(\mu)$ as a vector in $\mathbb{R}^{m_n}$ instead of a map $\mathbb{R}^{m_n}\to\mathbb{R}$ as before and hence included a transpose. On the other hand, write $\tilde{\mathcal{Z}} := \{\tilde{\mathbf{Z}}_i\}_{i\le n}$, where $\tilde{\mathbf{Z}}_2,\ldots,\tilde{\mathbf{Z}}_n$ are i.i.d. copies of the Gaussian vector $\tilde{\mathbf{Z}}_1$ specified in the statement in Proposition 6.14. Then by distributional invariance assumption $\phi_{11}\mathbf{X}_1 \overset{d}{=} \mathbf{X}_1$ and the assumption on $\tilde{\mathbf{Z}}_1$, we get that for $1\le j\le k$,

$$\mathbb{E}[\tilde{\mathbf{Z}}_{1j}] = \mathbb{E}[\tilde{\mathbf{Z}}_{11}] = \mathbb{E}[\tau_{11}(\mathbf{X}_1)] = \mathbb{E}[\{L(\beta_r,\mathbf{X}_1)\}_{r\le m_n}] = \mathbb{E}[\{L(\beta_r,\phi_{11}\mathbf{X}_1)\}_{r\le m_n}] = \mu\;.$$

Therefore, writing

$$\Sigma_{11} \;:=\; \mathrm{Var}[\pi_{11}(\mathbf{X}_1)]\;, \qquad\qquad \Sigma_{12} \;:=\; \mathrm{Cov}[\pi_{11}(\mathbf{X}_1),\pi_{12}(\mathbf{X}_1)]\;,$$

we get that $\big(\frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k}\tilde{\mathbf{Z}}_{ij}-\mu\big)$ and $\big(\frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k}\mathbf{Z}_{ij}-\mu\big)$ are both zero-mean Gaussian vectors in $\mathbb{R}^{m_n}$ with variance given by

$$\mathrm{Var}\Big[\frac{1}{nk}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{k}\mathbf{Z}_{ij}-\mu\Big] \;=\; \frac{1}{nk}\Sigma_{11}+\frac{k-1}{nk}\Sigma_{12}\;,$$

$$\mathrm{Var}\Big[\frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k}\tilde{\mathbf{Z}}_{ij}-\mu\Big] \;=\; \frac{1}{nk}\mathrm{Var}[\tau_{11}(\mathbf{X}_1)]+\frac{k-1}{nk}\mathrm{Cov}[\tau_{11}(\mathbf{X}_1),\tau_{12}(\mathbf{X}_1)] \overset{(a)}{=} \frac{1}{n}\Sigma_{11}\;,$$

where (a) is because

$$\begin{aligned}\mathrm{Var}[\tau_{11}(\mathbf{X}_1)] \;&=\; \mathrm{Cov}[\tau_{11}(\mathbf{X}_1),\tau_{12}(\mathbf{X}_1)] \;=\; \mathrm{Var}[\{L(\beta_r,\mathbf{X}_1)\}_{r\le m_n}]\\ &=\; \mathrm{Var}[\{L(\beta_r,\phi_{11}\mathbf{X}_1)\}_{r\le m_n}] \;=\; \mathrm{Var}[\pi_{11}(\mathbf{X}_1)] \;=\; \Sigma_{11}\;.\end{aligned}$$

In particular, this implies that $f^T(\mathcal{Z})$ is a Gaussian vector with mean $g(\mu)$ and variance satisfying

$$\mathrm{Var}[f^T(\mathcal{Z})] \;=(\partial g(\mu))^\top\big(\frac{1}{nk}\Sigma_{11}+\frac{k-1}{nk}\Sigma_{12}\big)\partial g(\mu) \overset{(a)}{\le} (\partial g(\mu))^\top\big(\frac{1}{n}\Sigma_{11}\big)\partial g(\mu)\;,$$

where we have noted in (a) that $\Sigma_{11}=\mathrm{Var}[\pi_{11}(\mathbf{X}_1)]\succeq\mathrm{Cov}[\pi_{11}(\mathbf{X}_1),\pi_{12}(\mathbf{X}_1)]=\Sigma_{12}$ by Lemma D.19. This suggests that if we define c_ϕ as in (D.86),

$$c_\phi \;:=\; \frac{(\partial g(\mu))^\top\big(\frac{1}{n}\Sigma_{11}\big)\partial g(\mu)}{(\partial g(\mu))^\top\big(\frac{1}{nk}\Sigma_{11}+\frac{k-1}{nk}\Sigma_{12}\big)\partial g(\mu)} \;\ge\; 1\;,$$

then $f^T(\mathcal{Z})$ is equal in distribution to the Gaussian vector

$$g(\mu)+\frac{1}{c_\phi}(\partial g(\mu))^\top\Big(\frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k}\tilde{\mathbf{Z}}_{ij}-\mu\Big) \;=:\; f^T\big(\tilde{\mathcal{Z}}^\phi\big)\;,$$

where we have defined $\tilde{\mathcal{Z}}^\phi:=\{\tilde{\mathbf{Z}}_i^\phi\}_{i\le n}$, where $\tilde{\mathbf{Z}}_i^\phi:=\frac{1}{c_\phi}(\tilde{\mathbf{Z}}_i-\mu)+\mu$. This together with (D.87) and the triangle inequality implies

$$\begin{aligned}&d_{\mathcal{H}}(\sqrt{n}f(\Phi\mathcal{X}),\sqrt{n}f^T(\tilde{\mathcal{Z}}^\phi))\\ &\le\; d_{\mathcal{H}}(\sqrt{n}f(\Phi\mathcal{X}),\sqrt{n}f^T(\mathcal{Z}))+\sup_{h\in\mathcal{H}}\big|\mathbb{E}[h(\sqrt{n}f(\Phi\mathcal{X}))-h(\sqrt{n}f^T(\mathcal{Z}^\phi))]\big|\\ &=\; o(1)+0 \;=\; o(1)\;,\end{aligned}\tag{D.89}$$

and similarly by (D.88),

$$\begin{aligned}&n\|\mathrm{Var}[f(\Phi\mathcal{X})]-\mathrm{Var}[f^T(\tilde{\mathcal{Z}}^\phi)\|\\ &\le\; n\|\mathrm{Var}[f(\Phi\mathcal{X})]-\mathrm{Var}[f^T(\mathcal{Z})]\|+n\|\mathrm{Var}[f^T(\mathcal{Z})]-\mathrm{Var}[f^T(\tilde{\mathcal{Z}}^\phi)]\|\\ &=\; o(1)+0 \;=\; o(1)\;.\end{aligned}\tag{D.90}$$

Finally, note that $\tilde{\mathbf{Z}}_i^\phi$ match the first two moments of

$$\frac{1}{c_\phi}(\{L(\beta_r,\mathbf{X}_i)\}_{r\le m_n}-\mu)+\mu \;=\; \Big\{\frac{1}{c_\phi}(L(\beta_r,\mathbf{X}_i)-\mu_r)+\mu_r\Big\}_{r\le m_n} \;=\; \tau_{ij}(\mathbf{X}_i)\;,$$

where we have invoked the definition of τ_{ij} from (D.83). We check that the bounds on $\tilde{c}_X^\phi$ and $\tilde{c}_Z^\phi$ (i.e. $\tilde{c}_X$ and $\tilde{c}_Z$ defined as in Corollary D.10 but for $\tau_{11}(\mathbf{X}_1)$ and $\tilde{\mathbf{Z}}_1^\phi$) holds:

$$\begin{aligned}\tilde{c}_X^\phi \;&=\frac{1}{6}\sqrt{\mathbb{E}\big[\max_{r\le m_n}|(\tau_{11}(\mathbf{X}_i))_l|^6\big]}\\ &=\frac{1}{6}\big\|\max_{r\le m_n}\big|\frac{1}{c_\phi}(L(\beta_r,\mathbf{X}_1)-\mu_r)+\mu_r\big|\,\big\|_{L_6}^3\\ &\le\frac{1}{6}\Big(\frac{1}{c_\phi}\big\|\max_{r\le m_n}\big|L(\beta_r,\mathbf{X}_1)\big|\,\big\|_{L_6}+\frac{c_\phi-1}{c_\phi}\max_{r\le m_n}|\mu_r|\Big)^3\end{aligned}$$

$$\leq \frac{1}{6}\Big(\frac{1}{c_\phi}\big\|\max_{r\leq m_n}\big|L(\beta_r,\mathbf{X}_1)\big|\,\big\|_{L_6}+\frac{c_\phi-1}{c_\phi}\|\max_{r\leq m_n}L(\beta_r,\mathbf{X}_1)\|_{L_1}\Big)^3$$
$$\leq \frac{1}{6}\big\|\max_{r\leq m_n}\big|L(\beta_r,\mathbf{X}_1)\big|\,\big\|_{L_6}^3 \;=\; O(1)\;,$$

where the last bound is by assumption of Proposition 6.14. Similarly for the moment term of the surrogate variable, by noting that $\tilde{\mathbf{Z}}_i^\phi := \frac{1}{c_\phi}(\tilde{\mathbf{Z}}_i-\mu)+\mu$, we have

$$\begin{aligned}\tilde{c}_Z^\phi &= \frac{1}{6}\sqrt{\mathbb{E}\Big[\frac{1}{k}\sum_{j=1}^k\max_{r\leq m_n}|(\tilde{\mathbf{Z}}_{1j}^\phi)_r|^6\Big]} \;=\; \frac{1}{6}\sqrt{\mathbb{E}\big[\max_{r\leq m_n}|(\tilde{\mathbf{Z}}_{11}^\phi)_r|^6\big]}\\
&\leq \frac{1}{6}\Big(\frac{1}{c_\phi}\big\|\max_{r\leq m_n}\big|\tilde{\mathbf{Z}}_{11}\big|\,\big\|_{L_6}+\frac{c_\phi-1}{c_\phi}\|\max_{r\leq m_n}L(\beta_r,\mathbf{X}_1)\|_{L_1}\Big)^3\\
&\leq \frac{1}{6}\max\Big\{\big\|\max_{r\leq m_n}\big|\tilde{\mathbf{Z}}_{11}\big|\,\big\|_{L_6},\,\|\max_{r\leq m_n}L(\beta_r,\mathbf{X}_1)\|_{L_6}\Big\}^3\\
&= O\big(\big\|\max_{r\leq m_n}\big|\tilde{\mathbf{Z}}_{11}\big|\,\big\|_{L_6}^3\big) \;=\; O((\log m_n)^{3/2}) \;=\; o(n^{1/6}),\end{aligned}$$

where in the last display line, we have used Lemma D.23 similar to how we have bounded $\tilde{c}_Z$. The bounds on derivative terms of g are the same as before, and therefore we get an analogous result to (D.87) and (D.88) for $\tilde{\mathcal{X}} := \{\tau_{ij}(\mathbf{X}_i)\}_{i\leq n,j\leq k}$:

$$d_{\mathcal{H}}(\sqrt{n}f(\tilde{\mathcal{X}}),\sqrt{n}f^T(\tilde{\mathcal{Z}}^\phi)) \;=\; o(1)\;,\quad n\big\|\mathrm{Var}[f(\tilde{\mathcal{X}})]-\mathrm{Var}\big[f^T(\tilde{\mathcal{Z}}^\phi)\big]\big\| \;=\; o(1)\;.$$

Observe that by construction, $f(\tilde{\mathcal{X}}) = f^*(\mathcal{X})$ as noted in (D.85). Combining the above bounds with (D.89) and (D.90) by the triangle inequality, we get

$$d_{\mathcal{H}}(\sqrt{n}f(\Phi\mathcal{X}),\sqrt{n}f^*(\mathcal{X})) \;=\; o(1)\;,\quad n\big\|\mathrm{Var}[f(\Phi\mathcal{X})]-\mathrm{Var}\big[f^*(\mathcal{X})\big]\big\| \;=\; o(1)\;.$$

This completes the proof of Proposition 6.14. □

The following lemma justifies the $\log(m_n)$ scaling in the definition of softmax ensemble in Proposition 6.14:

$$\beta_t \;:=\; \sum_{r=1}^{m_n}\beta_r\frac{\exp\big(-t\frac{\log(m_n)}{nk}\sum_{i=1}^n\sum_{j=1}^k L(\beta_r,\mathbf{x}_{ij})\big)}{\sum_{s=1}^m\exp\big(-t\frac{\log(m_n)}{nk}\sum_{i=1}^n\sum_{j=1}^k L(\beta_s,\mathbf{x}_{ij})\big)}\;.$$

Lemma D.39. *Fix n, k, m_n, the training set on which β_r's are trained and the testing set $\{\mathbf{x}_{11},\ldots,\mathbf{x}_{nk}\}$. Denote $\bar{L}(\beta_r) := \frac{1}{nk}\sum_{i=1}^n\sum_{j=1}^k L(\beta_r,\mathbf{x}_{ij})$ and $\bar{L}_{min} := \min_{s\leq m_n}\bar{L}(\beta_s)$. Define the minimizing set $S := \{r\in\{1,\ldots,m_n\}\mid \bar{L}(\beta_r)=\bar{L}_{min}\}$, i.e. S is the indexing set of β_r's that minimize $\bar{L}$. Consider $\beta^S(\mathbf{x}_{11:nk}) := \frac{1}{|S|}\sum_{r\in S}\beta_r(\mathbf{x}_{1:n_1})$, an average of β_r's within the minimizing set S, and define $\bar{L}_2 := \min_{s\notin S}\bar{L}(\beta_s)$. Then, if $|S|=m_n$, $\beta^t(\mathbf{x}_{11:nk}) = \beta^S(\mathbf{x}_{11:nk})$ and otherwise*

$$\begin{aligned}&\big\|\beta^t(\mathbf{x}_{11:nk})-\beta^S(\mathbf{x}_{11:nk})\big\|\\
&\leq\; 2\max_{r\leq m_n}\|\beta_r(\mathbf{x}_{11:nk})\|\,\Big(1-\frac{1}{1+\exp(\log(m_n)-t\log(m_n)(\bar{L}_2-\bar{L}_{min}))}\Big).\end{aligned}$$

Notably if $\max_{r\leq m_n}\|\beta_r(\mathbf{x}_{11:nk})\|$ is bounded, then as $t\to\infty$, $\beta^t(\mathbf{x}_{11:nk})\to\beta^S(\mathbf{x}_{11:nk})$.

Proof of Lemma D.39. Denote $w_r := \frac{\exp(-t\log(m_n)\bar{L}(\beta_r))}{\sum_{s\le m_n}\exp(-t\log(m_n)\bar{L}(\beta_s))}$. If $|S| = m_n$, the equality holds since $w_r = \frac{1}{m_n}$ for all r. If $|S| \neq m_n$, $w_r < \frac{1}{|S|}$ for $r \in S$. By triangle inequality,

$$\begin{aligned}
\left\|\beta^t(\mathbf{x}_{11:nk}) - \beta^S(\mathbf{x}_{11:nk})\right\| &= \left\|\sum_{r=1}^{m_n}\beta_r(\mathbf{x}_{1:n_1})w_r - \sum_{r\in S}\beta_r(\mathbf{x}_{1:n_1})\frac{1}{|S|}\right\| \\
&= \left\|\sum_{r\notin S}\beta_r(\mathbf{x}_{1:n_1})w_r - \sum_{r\in S}\beta_r(\mathbf{x}_{1:n_1})\Big(\frac{1}{|S|} - w_r\Big)\right\| \\
&\le \sum_{r\notin S}\left\|\beta_r(\mathbf{x}_{1:n_1})w_r\right\| + \sum_{r\in S}\left\|\beta_r(\mathbf{x}_{1:n_1})\Big(\frac{1}{|S|} - w_r\Big)\right\| \\
&\le \max_{1\le r\le m_n}\left\|\beta_r(\mathbf{x}_{1:n_1})\right\|\Big(\sum_{r\notin S} w_r + \sum_{r\in S}\Big(\frac{1}{|S|} - w_r\Big)\Big) \\
&= \max_{1\le r\le m_n}\left\|\beta_r(\mathbf{x}_{1:n_1})\right\|\Big(\sum_{r\notin S} w_r + 1 - \sum_{r\in S} w_r\Big) \\
&= 2\max_{1\le r\le m_n}\left\|\beta_r(\mathbf{x}_{1:n_1})\right\|\Big(1 - \sum_{r\in S} w_r\Big)\,. \qquad \text{(D.91)}
\end{aligned}$$

The final equality is by $\sum_{r\le m_n} w_r = 1$. For $r \in S$, by construction, $\bar{L}(\beta_r) = \bar{L}_{min}$, so

$$\begin{aligned}
\sum_{r\in S} w_r &= \sum_{r\in S}\frac{\exp(-t\log(m_n)\bar{L}_{min})}{|S|\exp(-t\log(m_n)\bar{L}_{min}) + \sum_{s\notin S}\exp(-t\log(m_n)\bar{L}(\beta_s))} \\
&= \frac{1}{1 + \frac{1}{|S|}\sum_{s\notin S}\exp(-t\log(m_n)(\bar{L}(\beta_s) - \bar{L}_{min}))} \\
&\le \frac{1}{1 + \frac{m_n - |S|}{|S|}\exp(-t\log(m_n)(\bar{L}_2 - \bar{L}_{min}))} \\
&\le \frac{1}{1 + \exp(\log(m_n) - t\log(m_n)(\bar{L}_2 - \bar{L}_{min}))}\,.
\end{aligned}$$

Substituting the bound into (D.91) recovers the desired result. The last statement is obtained by noting that if $|S| = m$, $\beta^t(\mathbf{x}_{11:nk}) = \beta^S(\mathbf{x}_{11:nk})$ by definition, and if not, $\bar{L}_2 - \bar{L}_{\min} > 0$, so the bound proved above goes to zero as $t \to \infty$. □

D.7 Proof for Section 6.5 and Appendix D.2.2

We follow the notation in Section 6.5 and Appendix D.2.2. We first prove a list of results on $f_\lambda^{(1)}$ and $f_\lambda^{(2)}$, collected in Lemma D.40, that are useful for subsequent derivations. Section D.7.1 presents the proofs for results in Appendix D.2.2 and Appendix D.7.2 presents the proofs for Section 6.5.

Throughout, for a real symmetric matrix $A \in \mathbb{R}^{n\times n}$, we denote $\lambda_1(A) \le \ldots \le \lambda_d(A)$ as its eigenvalues and denote the associated eigenvectors as $v_1(A), \ldots, v_d(A)$.

Lemma D.40. *Let A, A' and B be $\mathbb{R}^{d\times d}$ symmetric matrices and fix $\lambda \ge 0$.*

(i) The following bounds control the sizes of $f_\lambda^{(1)}$ and $f_\lambda^{(2)}$:

$$|f_\lambda^{(1)}(A)| \le \max_{l\le d;\ \lambda_l(A)\neq -\lambda}\frac{\lambda^2}{(\lambda_l(A)+\lambda)^2}\|\beta\|^2 \qquad \textit{for } \lambda > 0\,,$$

$$
\begin{aligned}
|f_0^{(1)}(A)| &\le \textstyle\sum_{l=1}^{d} \mathbb{I}_{\{\lambda_l(A)=0\}} \|\beta\|^2 \,, \\
|f_\lambda^{(2)}(A,B)| &\le \max_{l\le d;\, \lambda_l(A)\neq -\lambda} \frac{d\,\sigma_\epsilon^2\,\|B\|_{op}}{n(\lambda_l(A)+\lambda)^2} \,.
\end{aligned}
$$

(ii) The following bounds hold for the approximations of $f_0^{(1)}$ by $f_\lambda^{(1)}$ and $f_0^{(2)}$ by $f_\lambda^{(2)}$, where $\lambda > 0$:

$$
\begin{aligned}
\big|f_\lambda^{(1)}(A) - f_0^{(1)}(A)\big| &\le \max_{l\le d;\, \lambda_l(A)\notin\{0,-\lambda\}} \frac{\lambda^2\|\beta\|^2}{(\lambda_l(A)+\lambda)^2} \,, \\
\big|f_\lambda^{(2)}(A,B) - f_0^{(2)}(A,B)\big| &\le \frac{\sigma_\epsilon^2}{n\lambda^2} \textstyle\sum_{l=1}^{d} \mathbb{I}_{\{\lambda_l(A)\in\{0,-\lambda\}\}} |v_l(A)^\top B\, v_l(A)| \\
&\quad + \frac{\lambda d\,\sigma_\epsilon^2}{n} \max_{l\le d;\, \lambda_l(A)\notin\{0,-\lambda\}} \frac{|\lambda + 2\lambda_l(A)|\,\|B\|_{op}}{\lambda_l(A)^2(\lambda_l(A)+\lambda)^2} \,.
\end{aligned}
$$

Now suppose additionally that $\lambda > 0$, $\lambda_1(A) \ge -\lambda/2$ and $\lambda_1(A') \ge -\lambda/2$. Then we have

(iii) the following bounds hold on the effect of perturbing the argument of $f_\lambda^{(1)}$ and $f_\lambda^{(2)}$:

$$
\begin{aligned}
\big|f_\lambda^{(1)}(A) - f_\lambda^{(1)}(A')\big| &\le 4\,\|\beta\|^2 \|A - A'\|_{op} \\
\big|f_\lambda^{(2)}(A,B) - f_\lambda^{(2)}(A',B)\big| &\le \frac{16\,\sigma_\epsilon^2 d}{n\lambda^3} \|A - A'\|_{op} \|B\|_{op} \,.
\end{aligned}
$$

Proof of Lemma D.40. To prove (i), we first note that for $\lambda > 0$,

$$
\begin{aligned}
\big|\, f_\lambda^{(1)}(A)\,\big| &= \lambda^2 \big|\beta^\top \big(A + \lambda \mathbf{I}_d\big)^{-2} \beta\big| \\
&= \textstyle\sum_{l=1}^{d} \frac{\lambda^2\,\|\beta\|^2}{(\lambda_l(A)+\lambda)^2} \mathbb{I}_{\{\lambda_l(A)\neq -\lambda\}} \;\le\; \max_{l\le d;\, \lambda_l(A)\neq -\lambda} \frac{\lambda^2\|\beta\|^2}{(\lambda_l(A)+\lambda)^2} \,,
\end{aligned}
$$

whereas for $\lambda = 0$, we have

$$
\begin{aligned}
\big|\, f_0^{(1)}(A)\,\big| &= \big\|\big(A^\dagger A - \mathbf{I}_d\big)\beta\big\|^2 \\
&= \textstyle\sum_{l=1}^{d} \Big((0-1)^2 \mathbb{I}_{\{\lambda_l(A)=0\}} + (1-1)^2 \mathbb{I}_{\{\lambda_l(A)\neq 0\}}\Big) \|\beta\|^2 \\
&= \textstyle\sum_{l=1}^{d} \mathbb{I}_{\{\lambda_l(A)=0\}} \|\beta\|^2 \,.
\end{aligned}
$$

Meanwhile for $\lambda \ge 0$, we have

$$
\begin{aligned}
\big|\, f_\lambda^{(2)}(A,B)\,\big| &= \frac{\sigma_\epsilon^2}{n} \big|\mathrm{Tr}\big(\big(A + \lambda \mathbf{I}_d\big)^{-2} B\big)\big| \\
&\le \frac{\sigma_\epsilon^2\,\|B\|_{op}}{n} \Big| \textstyle\sum_{l=1}^{d} \frac{\mathbb{I}\{\lambda_l(A)\neq -\lambda\}}{(\lambda_l(A)+\lambda)^2} \Big| \;=\; \max_{l\le d;\, \lambda_l(A)\neq -\lambda} \frac{d\,\sigma_\epsilon^2\,\|B\|_{op}}{n(\lambda_l(A)+\lambda)^2} \,.
\end{aligned}
$$

To prove (ii), note that by assumption $\lambda > 0$. The first difference can be bounded as

$$
\begin{aligned}
\big|f_\lambda^{(1)}(A) - f_0^{(1)}(A)\big| &= \big|\beta^\top \big(\big(A^\dagger A - \mathbf{I}_d\big)^2 - \lambda^2 \big(A + \lambda \mathbf{I}_d\big)^{-2} \big)\beta\big| \\
&\le \|\beta\|^2 \,\big\|\big(A^\dagger A - \mathbf{I}_d\big)^2 - \lambda^2 \big(A + \lambda \mathbf{I}_d\big)^{-2}\big\|_{op} \\
&\overset{(a)}{=} \|\beta\|^2 \,\max\Big\{ \Big|(-1)^2 - \frac{\lambda^2}{\lambda^2}\Big|,\, \max_{l\le d;\, \lambda_l(A)\neq 0} \Big|0^2 - \frac{\lambda^2 \mathbb{I}\{\lambda_l(A)\neq -\lambda\}}{(\lambda_l(A)+\lambda)^2}\Big| \Big\}
\end{aligned}
$$

$$\leq \max_{l\leq d;\, \lambda_l(A)\notin\{0,-\lambda\}} \frac{\lambda^2\|\beta\|^2}{(\lambda_l(A)+\lambda)^2} \,.$$

In (a), we have noted that all matrices involved share the same set of eigenvectors. The second difference can be controlled as

$$\begin{aligned}
\big|f_\lambda^{(2)}(A,B) - f_0^{(2)}(A,B)\big| \;&=\; \frac{\sigma_\epsilon^2}{n}\Big|\mathrm{Tr}\Big(\Big(\big(A+\lambda\mathbf{I}_d\big)^{-2} - A^{-2}\Big)B\Big)\Big| \\
&\leq \frac{\sigma_\epsilon^2}{n}\sum_{l=1}^d \Big|\Big(\frac{\mathbb{I}\{\lambda_l(A)\neq -\lambda\}}{(\lambda_l(A)+\lambda)^2} - \frac{\mathbb{I}\{\lambda_l(A)\neq 0\}}{\lambda_l(A)^2}\Big)(v_l(A)^\top B\, v_l(A))\Big| \\
&\leq \frac{\sigma_\epsilon^2}{n}\sum_{l=1}^d \Big(\frac{\mathbb{I}\{\lambda_l(A)\in\{0,-\lambda\}\}}{\lambda^2} + \mathbb{I}_{\{\lambda_l(A)\notin\{0,-\lambda\}\}}\frac{|\lambda^2+2\lambda\lambda_l(A)|}{\lambda_l(A)^2(\lambda_l(A)+\lambda)^2}\Big)|v_l(A)^\top B\, v_l(A)| \\
&\leq \frac{\sigma_\epsilon^2}{n\lambda^2}\sum_{l=1}^d \mathbb{I}_{\{\lambda_l(A)\in\{0,-\lambda\}\}}|v_l(A)^\top B\, v_l(A)| \\
&\quad + \frac{\lambda d\,\sigma_\epsilon^2\|B\|_{op}}{n}\max_{l\leq d;\,\lambda_l(A)\notin\{0,-\lambda\}}\frac{|\lambda+2\lambda_l(A)|}{\lambda_l(A)^2(\lambda_l(A)+\lambda)^2} \,.
\end{aligned}$$

To prove (iii), we first note that by assumption, $\lambda_l(A)\geq -\lambda/2 > -\lambda$ for all $l\leq d$, so the map $\tilde{A}\mapsto(\tilde{A}+\lambda\mathbf{I}_d)^{-1}$ is smooth in the local neighbourhood of the line segment $[0,A]$; the same holds for A'. We can now apply the mean value theorem to $f_\lambda^{(1)}$ and $f_\lambda^{(2)}$ by computing their first derivatives: Writing $\tilde{A}_t = t(A-A')+A'$, we have

$$\begin{aligned}
\big|f_\lambda^{(1)}(A) - f_\lambda^{(1)}(A')\big| \;&\leq\; \sup_{t\in[0,1]}\Big|\lambda^2\beta^\top(\tilde{A}_t+\lambda\mathbf{I}_d)^{-1}(A-A')\big(\tilde{A}_t+\lambda\mathbf{I}_d\big)^{-1}\beta\Big| \\
&\leq \frac{\lambda^2\|\beta\|^2\|A-A'\|_{op}}{(\lambda/2)^2} \;=\; 4\,\|\beta\|^2\|A-A'\|_{op} \,.
\end{aligned}$$

In the last line, we have noted that all eigenvalues of $t(A-A')+A'$ are bounded from below by $-\lambda/2$. Similarly we have

$$\begin{aligned}
&\big|f_\lambda^{(2)}(A,B) - f_\lambda^{(2)}(A',B)\big| \\
&\leq \frac{\sigma_\epsilon^2}{n}\sum_{\substack{q_1,q_2\in\mathbb{N}\\ q_1+q_2=3}}\sup_{t\in[0,1]}\big|\mathrm{Tr}\big((\tilde{A}_t+\lambda\mathbf{I}_d)^{-q_1}(A-A')(\tilde{A}_t+\lambda\mathbf{I}_d)^{-q_2}B\big)\big| \\
&\leq \frac{2\sigma_\epsilon^2 d}{n}\|A-A'\|_{op}\big\|(\tilde{A}_t+\lambda\mathbf{I}_d)^{-1}\big\|_{op}^3\|B\|_{op} \\
&\leq \frac{16\,\sigma_\epsilon^2 d}{n\lambda^3}\|A-A'\|_{op}\|B\|_{op} \,.
\end{aligned}$$

□

D.7.1. Proofs for Appendix D.2.2

The proof exploits the assumption below on the distribution of the extreme eigenvalues of $\bar{\mathbf{X}}_1$, $\bar{\mathbf{X}}_2$, $\bar{\mathbf{Z}}_1$ and $\bar{\mathbf{Z}}_2$, as well as the alignment of their zero eigenspace.

Proof of Lemma D.15. First note that by the triangle inequality, almost surely

$$\begin{aligned}
&\big|f_\lambda(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2) - f_0(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2)\big| \\
&\leq \big|f_\lambda^{(1)}(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2) - f_0^{(1)}(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2)\big| + \big|f_\lambda^{(2)}(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2) - f_0^{(2)}(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2)\big| \,.
\end{aligned}$$

Applying Lemma D.40(ii), we get that almost surely

$$\left|f_\lambda^{(1)}(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2)-f_0^{(1)}(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2)\right| \;\le\; \lambda^2\|\beta\|^2 \max_{l\le d;\,\lambda_l(\bar{\mathbf{X}}_1)\notin\{0,-\lambda\}} \frac{1}{(\lambda_l(\bar{\mathbf{X}}_1)+\lambda)^2}\,,$$

and

$$\begin{aligned}\left|f_\lambda^{(2)}(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2)-f_0^{(2)}(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2)\right| \;\le\; &\frac{\sigma_\epsilon^2}{n\lambda^2}\sum_{l=1}^d \mathbb{I}_{\{\lambda_l(\bar{\mathbf{X}}_1)\in\{0,-\lambda\}\}}\big(v_l(\bar{\mathbf{X}}_1)^\top\bar{\mathbf{X}}_2\,v_l(\bar{\mathbf{X}}_1)\big)\\ &+\frac{\lambda d\,\sigma_\epsilon^2\,\|\bar{\mathbf{X}}_2\|_{op}}{n}\max_{l\le d;\,\lambda_l(\bar{\mathbf{X}}_1)\notin\{0,-\lambda\}}\frac{|\lambda+2\lambda_l(\bar{\mathbf{X}}_1)|}{\lambda_l(\bar{\mathbf{X}}_1)^2(\lambda_l(\bar{\mathbf{X}}_1)+\lambda)^2}\,.\end{aligned}$$

The above bound can be simplified by noting that all eigenvalues of $\bar{\mathbf{X}}_1$ are non-negative, which implies that almost surely for all $1\le l\le d$,

$$\begin{aligned}&\frac{\mathbb{I}\{\lambda_l(\bar{\mathbf{X}}_1)\notin\{0,-\lambda\}\}}{(\lambda_l(\bar{\mathbf{X}}_1)+\lambda)^2} \le \frac{\mathbb{I}\{\lambda_l(\bar{\mathbf{X}}_1)\neq 0\}}{\lambda_l(\bar{\mathbf{X}}_1)^2} \le \big\|\bar{\mathbf{X}}_1^\dagger\big\|_{op}^2\,, \quad \mathbb{I}_{\{\lambda_l(\bar{\mathbf{X}}_1)\in\{0,-\lambda\}\}} = \mathbb{I}_{\{\lambda_l(\bar{\mathbf{X}}_1)=0\}}\\ &\frac{\mathbb{I}\{\lambda_l(\bar{\mathbf{X}}_1)\notin\{0,-\lambda\}\}\times|\lambda+2\lambda_l(\bar{\mathbf{X}}_1)|}{\lambda_l(\bar{\mathbf{X}}_1)^2(\lambda_l(\bar{\mathbf{X}}_1)+\lambda)^2} \le \frac{2\,\mathbb{I}\{\lambda_l(\bar{\mathbf{X}}_1)\neq 0\}}{\lambda_l(\bar{\mathbf{X}}_1)^3} \le 2\big\|\bar{\mathbf{X}}_1^\dagger\big\|_{op}^3\,.\end{aligned}$$

Combining the bounds above and applying Assumption 6.2 gives that

$$\begin{aligned}\left|f_\lambda^{(1)}(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2)-f_0^{(1)}(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2)\right| &= O_{\gamma'}(\lambda^2)\,,\\ \left|f_\lambda^{(2)}(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2)-f_0^{(2)}(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2)\right| &= O_{\gamma'}\Big(\lambda+\frac{1}{n\lambda^2}\Big)\,,\\ \left|f_\lambda(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2)-f_0(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2)\right| &= O_{\gamma'}\Big(\lambda+\lambda^2+\frac{1}{n\lambda^2}\Big)\end{aligned}$$

with probability $1-o_{\gamma'}(1)$. By the definition of the Lévy–Prokhorov metric d_P (D.8), we obtain

$$\begin{aligned}d_P\big(f_\lambda^{(1)}(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2)\,,\,f_0^{(1)}(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2)\big) &= O_{\gamma'}(\lambda^2)\,,\\ d_P\big(f_\lambda^{(2)}(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2)\,,\,f_0^{(2)}(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2)\big) &= O_{\gamma'}\Big(\lambda+\frac{1}{n\lambda^2}\Big)\,,\\ d_P\big(f_\lambda(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2)\,,\,f_0(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2)\big) &= O_{\gamma'}\Big(\lambda+\lambda^2+\frac{1}{n\lambda^2}\Big)\,,\end{aligned}$$

which proves the first bound. The second bound follows from applying the same argument with $\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2$ replaced by $\bar{\mathbf{Z}}_1,\bar{\mathbf{Z}}_1+E_{12}$. □

The next proof exploits orthogonal invariance of isotropic Gaussians.

Proof of Lemma D.16. Consider the $\mathbb{R}^{d\times nk}$-valued random matrix

$$\mathbf{U} \;\coloneqq\; \big(\mathbf{V}_1+\xi_{11},\mathbf{V}_1+\xi_{12},\ldots,\mathbf{V}_n+\xi_{nk}\big)\,,$$

We can then express

$$\bar{\mathbf{Z}}_1 \;=\; \frac{1}{nk}\mathbf{U}\mathbf{U}^\top\,.$$

Notice that under (6.22), $\mathbf{U}$ have i.i.d. rows, each of which has a covariance matrix

$$\mathbf{I}_n\otimes\big(\mathbf{1}_{k\times k}+\sigma_A^2\mathbf{I}_k\big) \;=\; \mathbf{I}_n\otimes k\,Q_k^\top D_k Q_k\,.$$

This implies that we can express, for some choice of $\eta'_1, \ldots, \eta'_d \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_{nk})$, almost surely

$$\mathbf{U} \;=\; \sqrt{k} \begin{pmatrix} \leftarrow (\eta'_1)^\top \rightarrow \\ \vdots \\ \leftarrow (\eta'_d)^\top \rightarrow \end{pmatrix} (\mathbf{I}_n \otimes D_k^{1/2} Q_k) \;=:\; \sqrt{k}\, \mathbf{H} (\mathbf{I}_n \otimes D_k^{1/2} Q_k) \;,$$

and therefore almost surely we have

$$\bar{\mathbf{Z}}_1 \;= \frac{k}{nk} \mathbf{H} \left(\mathbf{I}_n \otimes D_k^{1/2} Q_k Q_k^\top D_k^{1/2} \right) \mathbf{H}^\top \;=\; \frac{1}{n} \mathbf{H} \left(\mathbf{I}_n \otimes D_k \right) \mathbf{H}^\top \;,$$

where $\mathbf{H}$ is an $\mathbb{R}^{d \times nk}$ matrix with i.i.d. standard Gaussian entries. Meanwhile, observing that

$$\bar{\mathbf{Z}}_2 \;= \frac{1}{nk} \mathbf{U} K K^\top \mathbf{U}^\top$$

proves the second statement. The final statement follows by identifying $\eta_{11}, \ldots, \eta_{nk}$ as the column vectors of $\mathbf{H}$, which yields

$$\bar{\mathbf{Z}}_1 \;=\; \frac{1}{n} \sum\nolimits_{i=1}^n \left(\frac{k + \sigma_A^2}{k} \eta_{i1} \eta_{i1}^\top + \frac{\sigma_A^2}{k} \sum\nolimits_{j=2}^k \eta_{ij} \eta_{ij}^\top \right) .$$

By recalling that

$$Q_k \;\coloneqq \begin{pmatrix} k^{-1/2} & \cdots & k^{-1/2} \\ \leftarrow & \mathbf{v}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{v}_{k-1}^\top & \rightarrow \end{pmatrix} ,$$

and observing that

$$\left(\mathbf{V}_1 + \xi_{11}, \mathbf{V}_1 + \xi_{12}, \ldots, \mathbf{V}_n + \xi_{nk} \right) \;=\; \mathbf{U} \;=\; \sqrt{k} \begin{pmatrix} \uparrow & & \uparrow \\ \eta_{11} & \cdots & \eta_{nk} \\ \downarrow & & \downarrow \end{pmatrix} (\mathbf{I}_n \otimes D_k^{1/2} Q_k) \;,$$

we obtain that

$$\eta_{i1} \;=\; \frac{1}{k} \sum\nolimits_{j=1}^k (\mathbf{V}_i + \xi_{ij}) \times \frac{\sqrt{k}}{\sqrt{k + \sigma_A^2}}$$

and therefore we can express

$$\begin{aligned} \bar{\mathbf{Z}}_1 \;&= \frac{1}{n} \sum_{i=1}^n \left(\left(\frac{1}{k} \sum_{j=1}^k (\mathbf{V}_i + \xi_{ij}) \right) \left(\frac{1}{k} \sum_{j=1}^k (\mathbf{V}_i + \xi_{ij}) \right)^\top + \frac{\sigma_A^2}{k} \sum_{j=2}^k \eta_{ij} \eta_{ij}^\top \right) \\ &= \bar{\mathbf{Z}}_2 + \frac{\sigma_A^2}{nk} \sum\nolimits_{i=1}^n \sum\nolimits_{j=2}^k \eta_{ij} \eta_{ij}^\top \;. \end{aligned}$$

□

Proof of Lemma D.17. We first verify Assumption 6.2. Under (6.22), we can apply Lemma D.16 to express

$$\bar{\mathbf{Z}}_1 \;=\; \frac{1}{n} \mathbf{H} \left(\mathbf{I}_n \otimes D_k \right) \mathbf{H}^\top \;,$$

where $D_k \in \mathbb{R}^{k \times k}$ is a positive diagonal matrix with minimum eigenvalue $\sigma_A^2 / k > 0$

and $\mathbf{H}$ is an $\mathbb{R}^{d\times nk}$ matrix with i.i.d. standard Gaussian entries. Given a real symmetric matrix A, let $\sigma_{\min}(A)$ denote its minimum non-zero eigenvalue and $\sigma_{\min;>0}(A)$ denote its minimum non-zero eigenvalue. Then almost surely

$$\begin{aligned}\|\bar{\mathbf{X}}_1^\dagger\|_{op} \overset{d}{=} \|\bar{\mathbf{Z}}_1^\dagger\|_{op} &= \big(\sigma_{\min;>0}(\bar{\mathbf{Z}}_1)\big)^{-1} \\ &= \Big(\sigma_{\min;>0}\Big(\frac{1}{n}\big(\mathbf{I}_n\otimes D_k^{1/2}\big)\mathbf{H}\mathbf{H}^\top\big(\mathbf{I}_n\otimes D_k^{1/2}\big)\Big)\Big)^{-1} \\ &\le \frac{1}{\sigma_A^2}\Big(\sigma_{\min;>0}\Big(\frac{1}{nk}\mathbf{H}\mathbf{H}^\top\Big)\Big)^{-1} \\ &= \frac{1}{\sigma_A^2}\Big(\sigma_{\min;>0}\Big(\frac{1}{nk}\textstyle\sum_{l=1}^d \eta_l\eta_l^\top\Big)\Big)^{-1}\,,\end{aligned}$$

where $\eta_1,\dots,\eta_d$ are some i.i.d. standard Gaussian vectors in $\mathbb{R}^{nk}$. Meanwhile, by the minimum singular value bound from Theorem 6.1 of Wainwright (2019), for any fixed $\epsilon>0$ and $nk\le d$,

$$\mathbb{P}\Big(\sigma_{\min}\Big(\frac{1}{d}\textstyle\sum_{l=1}^d \eta_l\eta_l^\top\Big) > \Big((1-\epsilon)-\frac{(nk)^{1/2}}{d^{1/2}}\Big)^2\Big) \;\ge\; 1-e^{-d\epsilon^2/2}\,,$$

so if $nk\le d$ with $\gamma'=\lim d/(kn)\in(1,\infty)$, we get that $\sigma_{\min}\big(\frac{1}{nk}\sum_{l=1}^d\eta_l\eta_l^\top\big)$ is bounded from below by some constant $c'_{\gamma'}\in(0,\infty)$ that only depends on γ'. This is still true if $nk\ge d$ with $\gamma'\in[0,1)$, since in this case

$$\begin{aligned}\sigma_{\min;>0}\Big(\frac{1}{nk}\textstyle\sum_{l=1}^d\eta_l\eta_l^\top\Big) &= \sigma_{\min;>0}\Big(\frac{1}{nk}\begin{pmatrix}\uparrow & & \uparrow\\ \eta_1 & \cdots & \eta_d\\ \downarrow & & \downarrow\end{pmatrix}\begin{pmatrix}\leftarrow\ \eta_1^\top\ \rightarrow\\ \vdots\\ \leftarrow\ \eta_d^\top\ \rightarrow\end{pmatrix}\Big) \\ &= \sigma_{\min}\Big(\frac{1}{nk}\begin{pmatrix}\leftarrow\ \eta_1^\top\ \rightarrow\\ \vdots\\ \leftarrow\ \eta_d^\top\ \rightarrow\end{pmatrix}\begin{pmatrix}\uparrow & & \uparrow\\ \eta_1 & \cdots & \eta_d\\ \downarrow & & \downarrow\end{pmatrix}\Big) \;=:\; \sigma_{\min}(\mathbf{W}_{nk})\,,\end{aligned}$$

and the same argument applies to the $\mathbb{R}^{d\times d}$ Wishart matrix $\mathbf{W}_{nk}$. This implies that $\|\bar{\mathbf{X}}_1^\dagger\|_{op}$ and $\|\bar{\mathbf{Z}}_1^\dagger\|_{op}$ are both $O_{\gamma'}(1)$ with probability $1-o_{\gamma'}(1)$ under the stated assumptions.

Meanwhile, by Lemma D.16 again,

$$\bar{\mathbf{X}}_2 \overset{d}{=} \bar{\mathbf{Z}}_2 \overset{a.s.}{=} \frac{1}{n}\mathbf{H}\big(\mathbf{I}_n\otimes D_k^{1/2}Q_k\big)KK^\top\big(\mathbf{I}_n\otimes Q_k^\top D_k^{1/2}\big)\mathbf{H}^\top$$

where $Q_k\in\mathbb{R}^{k\times k}$ is an orthogonal matrix. Therefore almost surely

$$\big\|\bar{\mathbf{Z}}_2\big\|_{op} \;\le\; \sigma_{\max}\big(KK^\top\big)\sigma_{\max}\big(\bar{\mathbf{Z}}_1\big) \;\le\; \frac{k+\sigma_A^2}{k}\times\sigma_{\max}\Big(\frac{1}{n}\textstyle\sum_{l=1}^d\eta_l\eta_l^\top\Big)\,, \tag{D.92}$$

where we have recalled from the definitions in Lemma D.16 that

$$\sigma_{\max}(KK^\top) = \sigma_{\max}\Big(\frac{1}{k}\mathbf{I}_n\otimes\mathbf{1}_{k\times k}\Big) = 1 \quad\text{and}\quad \big\|\mathbf{I}_n\otimes D_k^{1/2}Q_k\big\| \le \sqrt{\frac{k+\sigma_A^2}{k}}\,.$$

Applying the maximum singular value bound from Theorem 6.1 of Wainwright (2019) to $\frac{1}{nk}\sum_{l=1}^d\eta_l\eta_l^\top$ implies that $\|\bar{\mathbf{X}}_2\|_{op}$ is $O_{\gamma'}(1)$ with probability $1-o_{\gamma'}(1)$ provided that $nk\le d$ with $\gamma'=\lim d/nk>1$, and by noting again that

$$\sigma_{\max}\Big(\frac{1}{nk}\textstyle\sum_{l=1}^d\eta_l\eta_l^\top\Big) = \sigma_{\max}(\mathbf{W}_{nk})$$

for the $\mathbb{R}^{d\times d}$ Wishart matrix $\mathbf{W}_{nk}$, we get that the same holds when $nk \geq d$ with $\gamma' = \lim d/nk < 1$. This implies that $\|\bar{\mathbf{X}}_2\|_{op}$ and $\|\bar{\mathbf{Z}}_2\|_{op}$ are both $O_{\gamma'}(1)$ with probability $1 - o_{\gamma'}(1)$ under the stated assumptions.

The final quantity in Assumption 6.2 can be expressed as

$$\begin{aligned}\sum_{l=1}^{d} \mathbb{I}_{\{\lambda_l(\bar{\mathbf{X}}_1)=0\}} \big(v_l(\bar{\mathbf{X}}_1)^\top \bar{\mathbf{X}}_2 v_l(\bar{\mathbf{X}}_1)\big) &\overset{d}{=} \sum_{l=1}^{d} \mathbb{I}_{\{\lambda_l(\bar{\mathbf{Z}}_1)=0\}} \big(v_l(\bar{\mathbf{Z}}_1)^\top \bar{\mathbf{Z}}_2 v_l(\bar{\mathbf{Z}}_1)\big) \\ &= \sum_{l=1}^{d} \mathbb{I}\Big\{\lambda_l\Big(\frac{1}{nk}\sum_{l=1}^{d} \eta_l \eta_l^\top\Big) = 0\Big\} .\end{aligned}$$

Since $\bar{\mathbf{Z}}_1 = \frac{1}{n}\mathbf{H}(\mathbf{I}_n \otimes D_k)\mathbf{H}^\top$, where $\mathbf{I}_n \otimes D_k$ is positive-definite, if $v_l(\bar{\mathbf{Z}}_1)$ is a zero eigenvector of $\bar{\mathbf{Z}}_1$, then we must have $\mathbf{H}^\top v_l(\bar{\mathbf{Z}}_1) = \mathbf{0}$ almost surely. This implies

$$v_l(\bar{\mathbf{Z}}_1)^\top \bar{\mathbf{Z}}_2\, v_l(\bar{\mathbf{Z}}_1) = \frac{1}{n} v_l(\bar{\mathbf{Z}}_1)^\top \mathbf{H}\left(\mathbf{I}_n \otimes D_k^{1/2} Q_k\right) K K^\top \left(\mathbf{I}_n \otimes Q_k^\top D_k^{1/2}\right)\mathbf{H}^\top v_l(\bar{\mathbf{Z}}_1) = 0$$

almost surely, and therefore with probability $1 - o(1)$,

$$\sum_{l=1}^{d} \mathbb{I}_{\{\lambda_l(\bar{\mathbf{X}}_1)=0\}} \big(v_l(\bar{\mathbf{X}}_1)^\top \bar{\mathbf{X}}_2 v_l(\bar{\mathbf{X}}_1)\big) = \sum_{l=1}^{d} \mathbb{I}_{\{\lambda_l(\bar{\mathbf{Z}}_1)=0\}} \big(v_l(\bar{\mathbf{Z}}_1)^\top \bar{\mathbf{Z}}_2 v_l(\bar{\mathbf{Z}}_1)\big) = 0 .$$

This verifies Assumption 6.2.

To verify Assumption 6.1, we first note that since the entries of the matrices are all Gaussian, we automatically have $\max_{i\leq n, j\leq k, l\leq d} \|X_{ijl}\|_{L_{10}} = O(1)$. Meanwhile by (D.92),

$$\big\|\|\bar{\mathbf{X}}_2\|_{op}\big\|_{L_{60}} = \big\|\|\bar{\mathbf{Z}}_2\|_{op}\big\|_{L_{60}} \leq \frac{k+\sigma_A^2}{k}\Big\|\Big\|\frac{1}{nk}\sum_{l=1}^{d}\eta_l\eta_l^\top\Big\|_{op}\Big\|_{L_{60}} = \frac{k+\sigma_A^2}{k}\big\|\|\mathbf{W}_{nk}\|_{op}\big\|_{L_{60}}$$

where $\mathbf{W}_{nk}$ is the $\mathbb{R}^{d\times d}$ Wishart matrix defined above. By Theorem 4.6.1 of Vershynin (2018), there exists some constant $C_1 > 0$ such that, for all $t > 0$,

$$\mathbb{P}\Big(\big\|\mathbf{W}_{nk} - \mathbf{I}_d\big\|_{op} > 2C_1\frac{\sqrt{d}+t}{\sqrt{nk}} + C_1^2\frac{(\sqrt{d}+t)^2}{nk}\Big) \leq 2\exp(-t^2) .$$

Using that $d/(kn) = O(1)$, we get that for every fixed $m \in \mathbb{N}$, there exists some constant $C_m > 0$ depending on m such that

$$\mathbb{E}\big[\|\bar{\mathbf{Z}}_2 - \mathbf{I}_d\|_{op}^m\big] \leq \int_0^\infty \mathbb{P}\big(\|\mathbf{W}_{nk} - \mathbf{I}_d\|_{op} > s^{1/m}\big)ds \leq C_m .$$

This implies

$$\big\|\|\bar{\mathbf{X}}_2\|_{op}\big\|_{L_{60}} = \big\|\|\bar{\mathbf{Z}}_2\|_{op}\big\|_{L_{60}} \leq \big\|\|\mathbf{W}_{nk} - \mathbf{I}_d\|_{op}\big\|_{L_{60}} + \|\mathbf{I}_d\|_{op} = O(1) ,$$

which verifies Assumption 6.1. □

D.7.2. Proofs for Section 6.5

Proof of Proposition 6.10: Universality for oracle augmentation The proof adapts the two-moment matching argument from Theorem 6.1 to utilize the matching of four

moments. For $1 \le i \le n$ and $1 \le l \le d$, define the $\mathbb{R}^k$ vectors

$$\tilde{\mathbf{X}}_{il} \coloneqq (X_{i1l}, \dots, X_{ikl}) \qquad \text{and} \qquad \tilde{\mathbf{Z}}_{il} \coloneqq (Z_{i1l}, \dots, Z_{ikl}) .$$

We also rewrite

$$\begin{aligned}
\bar{\mathbf{X}}_1 &= \frac{1}{nk} \sum_{i=1}^n \sum_{l_1,l_2=1}^d \tilde{\mathbf{X}}_{il_1}^\top \tilde{\mathbf{X}}_{il_2} \mathbf{e}_{l_1} \mathbf{e}_{l_2}^\top =: S_1(\tilde{\mathbf{X}}_{11}, \dots, \tilde{\mathbf{X}}_{nd}) , \\
\bar{\mathbf{X}}_2 &= \frac{1}{nk^2} \sum_{i=1}^n \sum_{l_1,l_2=1}^d \sum_{j_1,j_2=1}^k X_{ij_1l_1} X_{ij_2l_2} \mathbf{e}_{l_1} \mathbf{e}_{l_2}^\top =: S_2(\tilde{\mathbf{X}}_{11}, \dots, \tilde{\mathbf{X}}_{nd}) , \\
\bar{\mathbf{Z}}_1 &= S_1(\tilde{\mathbf{Z}}_{11}, \dots, \tilde{\mathbf{Z}}_{nd}) , \qquad \bar{\mathbf{Z}}_2 = S_2(\tilde{\mathbf{Z}}_{11}, \dots, \tilde{\mathbf{Z}}_{nd}) .
\end{aligned}$$

As mentioned in Remark D.4, Theorem 6.1 can be directly extended to the independent but non-i.i.d. case, and we shall use it to replace the sequence of independent vectors $(\tilde{\mathbf{X}}_{11}, \dots, \tilde{\mathbf{X}}_{nd})$ by $(\tilde{\mathbf{Z}}_{11}, \dots, \tilde{\mathbf{Z}}_{nd})$ (note that in this case, k in Theorem 6.1 is set to 1). We also seek to exploit the fact that $\tilde{\mathbf{X}}_{ij}$ and $\tilde{\mathbf{Z}}_{ij}$ matches in the first four moments by assumption. By replacing the third-order Taylor expansion in Theorem 6.1 by a fifth-order Taylor expansion and a fifth-order Faà di Bruno's formula, we obtain that

$$\begin{aligned}
&d_{\mathcal{H}}\big(f_\lambda(\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2) , f_\lambda(\bar{\mathbf{Z}}_1, \bar{\mathbf{Z}}_2)\big) \\
&\le \sum_{i=1}^n \sum_{l=1}^d \frac{\sqrt{\mathbb{E}\big(\sum_{j=1}^k X_{ijl}^2\big)^5} + \sqrt{\mathbb{E}\big(\sum_{j=1}^k Z_{ijl}^2\big)^5}}{120} \\
&\quad \times \big(\theta_{1;10;X}^5 + 10\theta_{1;8;X}^3\theta_{2;8;X} + 10\theta_{1;6;X}^2\theta_{3;6;X} + 15\theta_{1;6;X}\theta_{2;6;X}^2 + 10\theta_{2;4;X}\theta_{3;4;X} \\
&\qquad + 5\theta_{1;4;X}\theta_{4;4;X} + \theta_{5;2;X} \\
&\qquad + \theta_{1;10;Z}^5 + 10\theta_{1;8;Z}^3\theta_{2;8;Z} + 10\theta_{1;6;Z}^2\theta_{3;6;Z} + 15\theta_{1;6;Z}\theta_{2;6;Z}^2 + 10\theta_{2;4;Z}\theta_{3;4;Z} \\
&\qquad + 5\theta_{1;4;Z}\theta_{4;4;Z} + \theta_{5;2;Z}\big) ,
\end{aligned}$$

where, for $m \ge 2$, $q \in \mathbb{N}$ and $r \in \{1, 2\}$, we define

$$\begin{aligned}
\theta_{q;m;X} &\coloneqq \max_{i\le n, l\le d} \Big\| \Big\| \partial_{il}^q f_\lambda\Big(\bar{\mathbf{W}}_{il}^{(1)}(\Theta\tilde{\mathbf{X}}_{il}), \bar{\mathbf{W}}_{il}^{(2)}(\Theta\tilde{\mathbf{X}}_{il})\Big) \Big\| \Big\|_{L_m} , \\
\theta_{q;m;Z} &\coloneqq \max_{i\le n, l\le d} \Big\| \Big\| \partial_{il}^q f_\lambda\Big(\bar{\mathbf{W}}_{il}^{(1)}(\Theta\tilde{\mathbf{Z}}_{il}), \bar{\mathbf{W}}_{il}^{(2)}(\Theta\tilde{\mathbf{Z}}_{il})\Big) \Big\| \Big\|_{L_m} , \\
\bar{\mathbf{W}}_{il}^{(r)}(\mathbf{x}) &\coloneqq S_r\big(\tilde{\mathbf{X}}_{\le il}, \mathbf{x}, \tilde{\mathbf{Z}}_{\ge il}\big) .
\end{aligned}$$

$\Theta \sim \mathrm{Uniform}[0, 1]$ is independent of all other random variables, $\tilde{\mathbf{X}}_{\le il}$ is the sequence formed by $\tilde{\mathbf{X}}_{i'l'}$'s such that (i', l') is before (i, l) in the lexicographical order, and $\tilde{\mathbf{Z}}_{\ge il}$ corresponds to $\tilde{\mathbf{Z}}_{i'l'}$'s such that (i', l') comes after (i, l). Now note that by the Jensen's inequality, we have

$$\sqrt{\mathbb{E}\big(\sum_{j=1}^k X_{ijl}^2\big)^5} = k^{5/2}\sqrt{\mathbb{E}\big(\frac{1}{k}\sum_{j=1}^k X_{ijl}^2\big)^5} \le k^{5/2} \max_{j\le k} \|X_{ijl}\|_{L_{10}}^5 \le k^{5/2} c_0^5 ,$$

where we have used Assumption 6.1 for the last inequality. Similarly

$$\sqrt{\mathbb{E}\big(\sum_{j=1}^k X_{ijl}^2\big)^5} \le k^{5/2} \max_{j\le k} \|Z_{ijl}\|_{L_{10}}^5 \le C' k^{5/2} c_0^5$$

for some absolute constant $C' > 0$; in the bound above, we have used that Z_{ijl} matches X_{ijl} in the first two moments, the moment formula of a Gaussian and that $\|X_{ijl}\|_{L_1} \le \|X_{ijl}\|_{L_2} \le \|X_{ijl}\|_{L_{10}}$. This implies that for some absolute constant $C'' > 0$, we have

$$\begin{aligned} & d_{\mathcal{H}}\big(f_\lambda(\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2)\,,\, f_\lambda(\bar{\mathbf{Z}}_1, \bar{\mathbf{Z}}_2)\big) \\ \le\ & C'' n d k^{5/2}\big(\theta^5_{1;10;X} + 10\theta^3_{1;8;X}\theta_{2;8;X} + 10\theta^2_{1;6;X}\theta_{3;6;X} + 15\theta_{1;6;X}\theta^2_{2;6;X} + 10\theta_{2;4;X}\theta_{3;4;X} \\ & \quad + 5\theta_{1;4;X}\theta_{4;4;X} + \theta_{5;2;X} \\ & \quad + \theta^5_{1;10;Z} + 10\theta^3_{1;8;Z}\theta_{2;8;Z} + 10\theta^2_{1;6;Z}\theta_{3;6;Z} + 15\theta_{1;6;Z}\theta^2_{2;6;Z} + 10\theta_{2;4;Z}\theta_{3;4;Z} \\ & \quad + 5\theta_{1;4;Z}\theta_{4;4;Z} + \theta_{5;2;Z}\big)\,. \end{aligned} \tag{D.93}$$

The remaining proof controls the derivatives. We will perform a detailed calculation of the first derivative, comment on the shared pattern and state the remaining derivatives. We first write x_{ijl} as the l-th coordinate of $\mathbf{x}_{ij}$ and note that

$$\begin{aligned} \frac{\partial S_1(\mathbf{x}_{11}, \ldots, \mathbf{x}_{nd})}{\partial x_{ijl}} &= \frac{1}{nk}\sum\nolimits_{l'=1}^{d} x_{ijl'}\big(\mathbf{e}_l\mathbf{e}_{l'}^\top + \mathbf{e}_{l'}\mathbf{e}_l^\top\big) = \frac{1}{nk}\Big(\mathbf{e}_l\begin{pmatrix} x_{ij1} \\ \vdots \\ x_{ijd}\end{pmatrix}^\top + \begin{pmatrix} x_{ij1} \\ \vdots \\ x_{ijd}\end{pmatrix}\mathbf{e}_l^\top\Big)\,, \\ \frac{\partial^2 S_1(\mathbf{x}_{11}, \ldots, \mathbf{x}_{nd})}{\partial x_{ijl}^2} &= \frac{1}{nk}\big(\mathbf{e}_l\mathbf{e}_{l'}^\top + \mathbf{e}_{l'}\mathbf{e}_l^\top\big)\,, \qquad \frac{\partial^3 S_1(\mathbf{x}_{11}, \ldots, \mathbf{x}_{nd})}{\partial x_{ijl}^3} = \mathbf{0}\,, \\ \frac{\partial S_2(\mathbf{x}_{11}, \ldots, \mathbf{x}_{nd})}{\partial x_{ijl}} &= \frac{1}{nk^2}\sum\nolimits_{l'=1}^{d}\sum\nolimits_{j'=1}^{k} x_{ij'l'}\big(\mathbf{e}_l\mathbf{e}_{l'}^\top + \mathbf{e}_{l'}\mathbf{e}_l^\top\big) \\ &= \frac{1}{nk^2}\sum\nolimits_{j'=1}^{k}\Big(\mathbf{e}_l\begin{pmatrix} x_{ij'1} \\ \vdots \\ x_{ij'd}\end{pmatrix}^\top + \begin{pmatrix} x_{ij'1} \\ \vdots \\ x_{ij'd}\end{pmatrix}\mathbf{e}_l^\top\Big)\,, \\ \frac{\partial^2 S_2(\mathbf{x}_{11}, \ldots, \mathbf{x}_{nd})}{\partial x_{ijl}^2} &= \frac{1}{nk^2}\big(\mathbf{e}_l\mathbf{e}_{l'}^\top + \mathbf{e}_{l'}\mathbf{e}_l^\top\big)\,, \qquad \frac{\partial^3 S_2(\mathbf{x}_{11}, \ldots, \mathbf{x}_{nd})}{\partial x_{ijl}^3} = \mathbf{0}\,. \end{aligned}$$

Meanwhile, since $\bar{\mathbf{W}}^{(1)}_{il}(t\tilde{\mathbf{X}}_{il})$ is positive semi-definite almost surely for all $t \in [0,1]$, the map $A \mapsto (A + \lambda\mathbf{I})^{-1}$ is differentiable in the local neighborhood of the line segment $[0, \bar{\mathbf{W}}^{(1)}_{il}(t\tilde{\mathbf{X}}_{il})]$ with respect to the Euclidean norm. For positive semi-definite matrix $A \in \mathbb{R}^{d\times d}$ and another matrix $B \in \mathbb{R}^{d\times d}$, denoting $A_\lambda \coloneqq A + \lambda\mathbf{I}_d$, we can compute

$$\begin{aligned} \frac{\partial f^{(1)}_\lambda(A)}{\partial A_{ij}} &= -\sum\nolimits_{\substack{q_1,q_2\in\mathbb{N}\\ q_1+q_2=3}} \lambda^2\beta^\top A_\lambda^{-q_1} E_{ij} A_\lambda^{-q_2}\beta\,, \\ \frac{\partial f^{(2)}_\lambda(A,B)}{\partial A_{ij}} &= \frac{\sigma_\epsilon^2}{n}\sum\nolimits_{\substack{q_1,q_2\in\mathbb{N}\\ q_1+q_2=3}} \mathrm{Tr}\big(A_\lambda^{-q_1}E_{ij}A_\lambda^{-q_2}B\big)\,, \qquad \frac{\partial f^{(2)}_\lambda(A,B)}{\partial B_{ij}} = \frac{\sigma_\epsilon^2}{n}\,\mathrm{Tr}\big(A_\lambda^{-2}E_{ij}\big)\,. \end{aligned}$$

Fix $m \in [2, 10]$. Using a chain rule with the derivatives computed above, we can calculate

$$\begin{aligned} \theta_{1;m;X} &= \Big\|\Big\|\partial_{il}\, f_\lambda\Big(\bar{\mathbf{W}}^{(1)}_{il}(\Theta\tilde{\mathbf{X}}_{il}), \bar{\mathbf{W}}^{(2)}_{il}(\Theta\tilde{\mathbf{X}}_{il})\Big)\Big\|\Big\|_{L_m} \\ &\le \Big\|\Big\|\partial_{il}\, f^{(1)}_\lambda\Big(\bar{\mathbf{W}}^{(1)}_{il}(\Theta\tilde{\mathbf{X}}_{il})\Big)\Big\|\Big\|_{L_m} + \Big\|\Big\|\partial_{il}\, f^{(2)}_\lambda\Big(\bar{\mathbf{W}}^{(1)}_{il}(\Theta\tilde{\mathbf{X}}_{il}), \bar{\mathbf{W}}^{(2)}_{il}(\Theta\tilde{\mathbf{X}}_{il})\Big)\Big\|\Big\|_{L_m} \\ &= \Big\|\Big(\sum_{j=1}^{k}\Big(-\lambda^2\sum\nolimits_{\substack{q_1,q_2\in\mathbb{N}\\ q_1+q_2=3}}\beta^\top\big(\bar{\mathbf{W}}^{(1)}_{il}(\Theta\tilde{\mathbf{X}}_{il}) + \lambda\mathbf{I}_d\big)^{-q_1}\frac{1}{nk}\Theta\big(\mathbf{e}_l(\pi_{ij}\mathbf{V}_i)^\top + (\pi_{ij}\mathbf{V}_i)\mathbf{e}_l^\top\big)\end{aligned}$$

$$
\begin{aligned}
&\Big(\bar{\mathbf{W}}_{il}^{(1)}(\Theta\tilde{\mathbf{X}}_{il}) + \lambda\mathbf{I}_d\big)^{-q_2}\beta\Big)^2\Big)^{1/2}\Big\|_{L_m} \\
&+\Big\|\Big(\sum_{j=1}^{k}\Big(\frac{\sigma_\epsilon^2}{n}\sum\nolimits_{\substack{q_1,q_2\in\mathbb{N}\\ q_1+q_2=3}}\mathrm{Tr}\big(\big(\bar{\mathbf{W}}_{il}^{(1)}(\Theta\tilde{\mathbf{X}}_{il}) + \lambda\mathbf{I}_d\big)^{-q_1} \\
&\qquad\times\frac{1}{nk}\Theta\big(\mathbf{e}_l(\pi_{ij}\mathbf{X}_i)^\top + (\pi_{ij}\mathbf{X}_i)\mathbf{e}_l^\top\big)\times\big(\bar{\mathbf{W}}_{il}^{(1)}(\Theta\tilde{\mathbf{X}}_{il}) + \lambda\mathbf{I}_d\big)^{-q_2}\bar{\mathbf{W}}_{il}^{(2)}(\Theta\tilde{\mathbf{X}}_{il})\big) \\
&\quad+\frac{\sigma_\epsilon^2}{n}\mathrm{Tr}\Big(\big(\bar{\mathbf{W}}_{il}^{(1)}(\Theta\tilde{\mathbf{X}}_{il}) + \lambda\mathbf{I}_d\big)^{-2}\frac{1}{nk^2}\sum_{j'=1}^{k}\Theta\big(\mathbf{e}_l(\pi_{ij'}\mathbf{X}_i)^\top + (\pi_{ij'}\mathbf{X}_i)\mathbf{e}_l^\top\big)\Big)\Big)^2\Big)^{1/2}\Big\|_{L_m} \\
&\le\ \frac{2\lambda^2\|\beta\|^2}{nk}\Big\|\Big\|\big(\bar{\mathbf{W}}_{il}^{(1)}(\Theta\tilde{\mathbf{X}}_{il}) + \lambda\mathbf{I}_d\big)^{-1}\Big\|_{op}^3\times\Big(\sum\nolimits_{j=1}^{k}\|\pi_{ij}\mathbf{V}_i\|^2\Big)^{1/2}\Big\|_{L_m} \\
&+\frac{4\sigma_\epsilon^2}{n^2k}\Big\|\Big\|\big(\bar{\mathbf{W}}_{il}^{(1)}(\Theta\tilde{\mathbf{X}}_{il}) + \lambda\mathbf{I}_d\big)^{-1}\Big\|_{op}^3\times\Big\|\bar{\mathbf{W}}_{il}^{(2)}(\Theta\tilde{\mathbf{X}}_{il})\Big\|_{op} \\
&\qquad\times\Big(\sum\nolimits_{j=1}^{k}\|\pi_{ij}\mathbf{X}_i\|^2\Big)^{1/2}\Big\|_{L_m} \\
&+\frac{2\sigma_\epsilon^2}{n^2k^{3/2}}\Big\|\Big\|\big(\bar{\mathbf{W}}_{il}^{(1)}(\Theta\tilde{\mathbf{X}}_{il}) + \lambda\mathbf{I}_d\big)^{-1}\Big\|_{op}^2\times\Big(\sum\nolimits_{j'=1}^{k}\|\pi_{ij'}\mathbf{X}_i\|\Big)\Big\|_{L_m}.
\end{aligned}
$$

To simplify this bound, notice that since $\bar{\mathbf{W}}_{il}^{(1)}(\Theta\tilde{\mathbf{X}}_{il})$ is positive semi-definite, almost surely

$$
\Big\|\big(\bar{\mathbf{W}}_{il}^{(1)}(\Theta\tilde{\mathbf{X}}_{il}) + \lambda\mathbf{I}_d\big)^{-1}\Big\|_{op}\ \le\ \frac{1}{\lambda}\,.
$$

Meanwhile since $m \ge 2$, by the Jensen's inequality,

$$
\begin{aligned}
\Big\|\Big(\sum\nolimits_{j=1}^{k}\|\pi_{ij}\mathbf{V}_i\|^2\Big)^{1/2}\Big\|_{L_m}\ &=\ k^{1/2}\Big(\mathbb{E}\Big[\Big(\frac{1}{k}\sum\nolimits_{j=1}^{k}\|\pi_{ij}\mathbf{V}_i\|^2\Big)^{m/2}\Big]\Big)^{1/m} \\
&\le k^{1/2}\,\max\nolimits_{j\le k}\Big(\mathbb{E}\Big[\|\pi_{ij}\mathbf{V}_i\|^m\Big]\Big)^{1/m} \\
&= d^{1/2}k^{1/2}\,\max\nolimits_{j\le k}\Big(\mathbb{E}\Big[\Big(\frac{1}{d}\sum\nolimits_{l=1}^{d}X_{ijl}^2\Big)^{m/2}\Big]\Big)^{1/m} \\
&\le d^{1/2}k^{1/2}\,\max\nolimits_{i\le n,j\le k,l\le d}\|X_{ijl}\|_{L_m} = O(d^{1/2}k^{1/2})\,,
\end{aligned}
$$

where we have applied Assumption 6.1 by noting that $m \le 12$. Similarly

$$
\Big\|\sum\nolimits_{j'=1}^{k}\|\pi_{ij'}\mathbf{X}_i\|\Big\|_{L_m} = O(d^{1/2})\,.
$$

Applying Assumption 6.1 again and noting that $|\Theta| \le 1$ almost surely, we have

$$
\begin{aligned}
\Big\|\Big\|\bar{\mathbf{W}}_{il}^{(2)}(\Theta\tilde{\mathbf{X}}_{il})\Big\|_{op}\Big\|_{L_m}\ &\le\ \Big\|\Big\|\frac{1}{n}\sum\nolimits_{i'=1}^{i-1}\Big(\frac{1}{k}\sum\nolimits_{j=1}^{k}(\pi_{i'j}\mathbf{V}_i)\Big)\Big(\frac{1}{k}\sum\nolimits_{j=1}^{k}(\pi_{i'j}\mathbf{V}_i)\Big)^\top\Big\|_{op}\Big\|_{L_m} \\
&\quad+\Big\|\Big\|\frac{\Theta^2}{n}\Big(\frac{1}{k}\sum\nolimits_{j=1}^{k}(\pi_{ij}\mathbf{V}_i)\Big)\Big(\frac{1}{k}\sum\nolimits_{j=1}^{k}(\pi_{ij}\mathbf{V}_i)\Big)^\top\Big\|_{op}\Big\|_{L_m} \\
&\quad+\Big\|\Big\|\frac{1}{n}\sum\nolimits_{i'=i+1}^{n}\Big(\frac{1}{k}\sum\nolimits_{j=1}^{k}\mathbf{Z}_i\Big)\Big(\frac{1}{k}\sum\nolimits_{j=1}^{k}\mathbf{Z}_i\Big)^\top\Big\|_{op}\Big\|_{L_m} \\
&\le\ \frac{i-1}{n}c_0\ +\ \frac{1}{n}c_0\ +\ \frac{n-i}{n}c_0\ =\ O(1)\,.
\end{aligned}
$$

Combining the above calculations and noting additionally that $\|\beta\| = O(1)$, $\sigma_\epsilon = O(1)$

and $d = O(n)$, we get that the first derivative term can be bounded as

$$\theta_{1;m;X} \;=\; O\Big(\frac{d^{1/2}\lambda^{-1}}{nk^{1/2}} + \frac{d^{1/2}(\lambda^{-3}+\lambda^{-2})}{n^2k^{1/2}}\Big) \;=\; O\Big(\frac{\max\{1,\lambda^{-3}\}}{n^{1/2}k^{1/2}}\Big)\,.$$

By using the same argument and additionally bounding $\|Z_{ijl}\|_{L_m}$ by $C''\|X_{ijl}\|_{L_m}$ for some absolute constant C'', we also have

$$\theta_{1;m;Z} \;=\; O\Big(\frac{\max\{1,\lambda^{-3}\}}{n^{1/2}k^{1/2}}\Big)\,.$$

To handle the higher-order derivative terms up to the fifth order, notice that in the above calculation, differentiating $f_\lambda^{(1)}$ and $f_\lambda^{(2)}$ with respect to $\bar{\mathbf{W}}_{il}^{(1)}(\Theta\tilde{\mathbf{X}}_{il})$ results in

- an additional $(\bar{\mathbf{W}}_{il}^{(1)}(\Theta\tilde{\mathbf{X}}_{il}) + \lambda\mathbf{I}_d)^{-1}$ term, which contributes an $1/\lambda$ factor, and
- an additional $\dfrac{\partial\bar{\mathbf{W}}_{il}^{(1)}(\Theta\tilde{\mathbf{X}}_{il})}{\partial X_{ijl}} = \dfrac{\Theta}{nk}\big(\mathbf{e}_l(\pi_{ij}\mathbf{X}_i)^\top + (\pi_{ij}\mathbf{X}_i)\mathbf{e}_l^\top\big)$ term, which contributes an $d^{1/2}/nk$ factor,

whereas differentiating $f_\lambda^{(2)}$ with respect to $\bar{\mathbf{W}}_{il}^{(2)}(\Theta\tilde{\mathbf{X}}_{il})$ results in

- an additional $\Big\|\bar{\mathbf{W}}_{il}^{(2)}(\Theta\tilde{\mathbf{X}}_{il})\Big\|_{op}$ term, which is $O(1)$, and
- an additional $\dfrac{\partial\bar{\mathbf{W}}_{il}^{(2)}(\Theta\tilde{\mathbf{X}}_{il})}{\partial X_{ijl}} = \dfrac{\Theta}{nk^2}\sum_{j'=1}^{k}\big(\mathbf{e}_l(\pi_{ij'}\mathbf{X}_i)^\top + (\pi_{ij'}\mathbf{X}_i)\mathbf{e}_l^\top\big)$ term, which contributes an $d^{1/2}/nk$ factor.

We also note a few additional points:

- The initial sizes of $f_\lambda^{(1)}$ and $f_\lambda^{(2)}$ before differentiation are $O(1)$ and $O(n^{-1}\lambda^{-2})$ respectively, and that the norm we compute in $\theta_{q;m;X}$ has a persisting $k^{1/2}$ factor;
- The higher derivatives will also involve higher derivatives of $\bar{\mathbf{W}}_{il}^{(1)}(\Theta\tilde{\mathbf{X}}_{il})$ and $\bar{\mathbf{W}}_{il}^{(2)}(\Theta\tilde{\mathbf{X}}_{il})$ with respect to X_{ijl}. But since the third derivatives vanish, the only additional terms are their second derivatives, which brings the sizes of the first derivatives down from $O(d^{1/2}/nk)$ down to $O(1/nk)$;
- The q-th derivative involves at most one copy of $\bar{\mathbf{W}}_{il}^{(2)}(\Theta\tilde{\mathbf{X}}_{il})$ and q copies of $\pi_{ij}\mathbf{V}_i$, so the bounding constant involves at most $(q+1)m$-th moments of $\bar{\mathbf{X}}_2$, $\bar{\mathbf{Z}}_2$ and $\pi_{ij}\mathbf{V}_i$. As Assumption 6.1 controls moments up to the order $60 \geq (q+1)m$ for $q \leq 5$, it yields the necessary moment controls for computing up to the fifth derivative.

One can therefore perform a tedious calculation to verify that each further differentiation brings a multiplicative factor of at most $\max\{1, \lambda^{-1}\}n^{-1/2}$ to the overall upper bound, i.e. for $1 \leq q \leq 5$,

$$\max\{\theta_{q;m;X}\,,\,\theta_{q;m;Z}\} \;=\; O\Big(\frac{\max\{1,\lambda^{-2-q}\}}{n^{q/2}k^{1/2}}\Big)\,.$$

Plugging the bounds into (D.93) implies

$$\begin{aligned}
&d_{\mathcal{H}}\big(f_\lambda(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2)\,,\,f_\lambda(\bar{\mathbf{Z}}_1,\bar{\mathbf{Z}}_2)\big)\\
&\qquad\leq C''ndk^{5/2}\\
&\qquad\quad\times\big(\theta_{1;10;X}^5+10\theta_{1;8;X}^3\theta_{2;8;X}+10\theta_{1;6;X}^2\theta_{3;6;X}+15\theta_{1;6;X}\theta_{2;6;X}^2+10\theta_{2;4;X}\theta_{3;4;X}\\
&\qquad\qquad+5\theta_{1;4;X}\theta_{4;4;X}+\theta_{5;2;X}\\
&\qquad\qquad+\theta_{1;10;Z}^5+10\theta_{1;8;Z}^3\theta_{2;8;Z}+10\theta_{1;6;Z}^2\theta_{3;6;Z}+15\theta_{1;6;Z}\theta_{2;6;Z}^2+10\theta_{2;4;Z}\theta_{3;4;Z}\\
&\qquad\qquad+5\theta_{1;4;Z}\theta_{4;4;Z}+\theta_{5;2;Z}\big)\\
&\qquad=O\Big(ndk^{5/2}\times\frac{\max\{1,\lambda^{-2-5}\}}{n^{5/2}k^{1/2}}\Big)\;=\;O\Big(\frac{k^2\max\{1,\lambda^{-7}\}}{n^{1/2}}\Big)\,,
\end{aligned}$$

where we have again used $d = O(n)$. This proves the universality statement for $\lambda > 0$ fixed.

For the ridgeless case, recall from Lemma D.18 that $d_P(\bullet,\star) \leq 8^{4/5}d_{\mathcal{H}}(\bullet,\star)^{1/5}$. By the triangle inequality and Lemma D.15, we have that for every $\lambda \in (0,1]$,

$$\begin{aligned}
&d_P\big(f_0(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2)\,,\,f_0(\bar{\mathbf{Z}}_1,\bar{\mathbf{Z}}_2)\big)\\
&\leq\; d_P\big(f_0(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2)\,,\,f_\lambda(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2)\big)+8^{\frac{4}{5}}d_{\mathcal{H}}\big(f_\lambda(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2)\,,\,f_\lambda(\bar{\mathbf{Z}}_1,\bar{\mathbf{Z}}_2)\big)^{\frac{1}{5}}\\
&\qquad+d_P\big(f_\lambda(\bar{\mathbf{Z}}_1,\bar{\mathbf{Z}}_2)\,,\,f_0(\bar{\mathbf{Z}}_1,\bar{\mathbf{Z}}_2)\big)\\
&=\;O\Big(\lambda+\lambda^2+\frac{1}{n\lambda}+\Big(\frac{k^2\max\{1,\lambda^{-7}\}}{n^{1/2}}\Big)^{1/5}\Big)\,.
\end{aligned}$$

Since $d = O(n)$ and $1 \leq k^2 = o(n^{1/2})$, setting $\lambda = k^{1/7}n^{-1/28}$ implies that the above bound is $o(1)$, which finishes the proof.

□

Proof of Proposition 6.11: Oracle augmentation via unaugmented risk The proof consists of three steps: We first quantify the error of approximating $\bar{\mathbf{Z}}_1$ by

$$\bar{\mathbf{Z}}_2+\frac{(k-1)\sigma_A^2}{k}\mathbf{I}_d$$

in the risk in the case $\lambda > 0$. This is followed by a similar approximation for the case $\lambda = 0$. Then we compute the limiting risk by reducing the risk to that of an unaugmented ridge regressor.

Step 1: Replace $\bar{\mathbf{Z}}_1$ in $f_\lambda(\bar{\mathbf{Z}}_1, \bar{\mathbf{Z}}_2)$ for $\lambda > 0$. Recall from Lemma D.16 that

$$\bar{\mathbf{Z}}_1 \;=\; \bar{\mathbf{Z}}_2 + \Delta \;,$$

where we denote the following rescaled Wishart matrix

$$\Delta \;:=\; \frac{\sigma_A^2}{nk} \sum_{i=1}^n \sum_{j=2}^k \eta_{ij}\eta_{ij}^\top \;,$$

and η_{ij}'s are i.i.d. standard Gaussians in $\mathbb{R}^d$. Also note that

$$\frac{(k-1)\sigma_A^2}{k}\,\mathbf{I}_d \;=\; \mathbb{E}[\Delta] \;.$$

This allows us to control

$$\begin{aligned}
&\Big|f_\lambda^{(1)}(\bar{\mathbf{Z}}_1) - f_\lambda^{(1)}\Big(\bar{\mathbf{Z}}_2 + \frac{(k-1)\sigma_A^2}{k}\mathbf{I}_d\Big)\Big| \\
&= \lambda^2\Big|\beta^\top\Big((\bar{\mathbf{Z}}_1 + \lambda\mathbf{I}_d)^{-2} - (\bar{\mathbf{Z}}_2 + \mathbb{E}[\Delta] + \lambda\mathbf{I}_d)^{-2}\Big)\beta\Big| \\
&\le \lambda^2\|\beta\|^2\Big\|(\bar{\mathbf{Z}}_1 + \lambda\mathbf{I}_d)^{-2}\left((\bar{\mathbf{Z}}_2 + \mathbb{E}[\Delta] + \lambda\mathbf{I}_d)^2 - (\bar{\mathbf{Z}}_1 + \lambda\mathbf{I}_d)^2\right)(\bar{\mathbf{Z}}_2 + \mathbb{E}[\Delta] + \lambda\mathbf{I}_d)^{-2}\Big\|_{op} \\
&\le \frac{\lambda^2\|\beta\|^2}{\lambda^2\big(\frac{k-1}{k}\sigma_A^2 + \lambda\big)^2}\Big\|\bar{\mathbf{Z}}_2 + \frac{(k-1)\sigma_A^2}{k}\mathbf{I}_d + \lambda\mathbf{I}_d + \bar{\mathbf{Z}}_1 + \lambda\mathbf{I}_d\Big\|_{op}\|\mathbb{E}[\Delta] - \Delta\|_{op} \\
&\le \frac{\|\beta\|^2}{\big(\frac{k-1}{k}\sigma_A^2 + \lambda\big)^2}\Big(\|\bar{\mathbf{Z}}_2\|_{op} + \|\bar{\mathbf{Z}}_1\|_{op} + \frac{(k-1)\sigma_A^2}{k} + 2\lambda\Big)\|\mathbb{E}[\Delta] - \Delta\|_{op} \;.
\end{aligned}$$

By adapting the proof of Lemma D.17 and using the maximum singular value bound from Theorem 6.1 of Wainwright (2019), we see that for any $\epsilon > 0$, with probability $1-\epsilon$ we have

$$\big\|\bar{\mathbf{Z}}_l\big\|_{op} \;\le\; \frac{k+\sigma_A^2}{k}\Big(1 + \sqrt{\frac{2\log(1/\epsilon)}{n}} + \sqrt{\frac{d}{n}}\,\Big)$$

for both $l = 1, 2$. Meanwhile, by noting that Δ is a rescaled sample covariance matrix of $n(k-1)$ i.i.d. isotropic Gaussians, by Theorem 4.6.1 of Vershynin (2018), there is some absolute constant $C' > 0$ such that for any $\epsilon > 0$, with probability $1-\epsilon$ we have

$$\|\Delta - \mathbb{E}[\Delta]\|_{op} \;\le\; C'\,\frac{(k-1)\sigma_A^2}{k}\Big(\frac{\sqrt{d} + \sqrt{\log(2/\epsilon)}}{\sqrt{n(k-1)}} + \frac{(\sqrt{d} + \sqrt{\log(2/\epsilon)}\,)^2}{n(k-1)}\Big) \;.$$

Also note that since $k \ge 2$, $\frac{k-1}{k} \in [\frac{1}{2}, 1]$. This implies that for some absolute constants $C'', C''' > 0$ such that with probability $1 - 3\epsilon$,

$$\begin{aligned}
&\Big|f_\lambda^{(1)}(\bar{\mathbf{Z}}_1) - f_\lambda^{(1)}\Big(\bar{\mathbf{Z}}_2 + \frac{(k-1)\sigma_A^2}{k}\mathbf{I}_d\Big)\Big| \\
&\quad\le C''\|\beta\|^2\,\frac{1}{(\sigma_A^2+\lambda)^2}\left(\frac{k+\sigma_A^2}{k}\Big(1 + \sqrt{\frac{2\log(1/\epsilon)}{n}} + \sqrt{\frac{d}{n}}\,\Big) + \frac{(k-1)\sigma_A^2}{k} + \lambda\right) \\
&\qquad\qquad \|\mathbb{E}[\Delta] - \Delta\|_{op} \\
&\quad\le \frac{C'''\|\beta\|^2}{(\sigma_A^2+\lambda)^2}\left(\frac{k+\sigma_A^2}{k}\Big(\sqrt{\frac{2\log(1/\epsilon)}{n}} + \sqrt{\frac{d}{n}}\,\Big) + 1 + \sigma_A^2 + \lambda\right)
\end{aligned}$$

$$\times\, \sigma_A^2\Big(\frac{\sqrt{d}+\sqrt{\log(2/\epsilon)}}{\sqrt{n(k-1)}}+\frac{(\sqrt{d}+\sqrt{\log(2/\epsilon)}\,)^2}{n(k-1)}\Big)\,.$$

Notice that by recycling the bound above, we have

$$\begin{aligned}
&\Big|f_\lambda^{(2)}(\bar{\mathbf{Z}}_1,\bar{\mathbf{Z}}_2)-f_\lambda^{(2)}\Big(\bar{\mathbf{Z}}_2+\frac{(k-1)\sigma_A^2}{k}\mathbf{I}_d,\bar{\mathbf{Z}}_2\Big)\Big|\\
&=\ \frac{\sigma_\epsilon^2}{n}\Big|\mathrm{Tr}\Big(\big(\bar{\mathbf{Z}}_1+\lambda\mathbf{I}_d\big)^{-2}\bar{\mathbf{Z}}_2-\Big(\bar{\mathbf{Z}}_2+\frac{(k-1)\sigma_A^2}{k}\mathbf{I}_d+\lambda\mathbf{I}_d\Big)^{-2}\bar{\mathbf{Z}}_2\Big)\Big|\\
&\le\ \frac{\sigma_\epsilon^2 d}{n}\,\|\bar{\mathbf{Z}}_2\|_{op}\,\Big\|\big(\bar{\mathbf{Z}}_1+\lambda\mathbf{I}_d\big)^{-2}-\Big(\bar{\mathbf{Z}}_2+\frac{(k-1)\sigma_A^2}{k}\mathbf{I}_d+\lambda\mathbf{I}_d\Big)^{-2}\Big\|_{op}\\
&\le\ \frac{C'''}{\lambda^2(\sigma_A^2+\lambda)^2}\Big(\frac{k+\sigma_A^2}{k}\Big(\sqrt{\frac{2\log(1/\epsilon)}{n}}+\sqrt{\frac{d}{n}}\,\Big)+1+\sigma_A^2+\lambda\Big)^2\\
&\quad\times\,\sigma_A^2\Big(\frac{\sqrt{d}+\sqrt{\log(2/\epsilon)}}{\sqrt{n(k-1)}}+\frac{(\sqrt{d}+\sqrt{\log(2/\epsilon)}\,)^2}{n(k-1)}\Big)
\end{aligned}$$

for some absolute constant $C'''>0$ with probability $1-3\epsilon$ for any $\epsilon>0$. By a union bound, we obtain that there exists some absolute constant $C>0$ such that for any $\epsilon>0$, with probability $1-6\epsilon$, we have

$$\begin{aligned}
&\Big|\,f_\lambda(\bar{\mathbf{Z}}_1,\bar{\mathbf{Z}}_2)-f_\lambda\Big(\bar{\mathbf{Z}}_2+\frac{(k-1)\sigma_A^2}{k}\mathbf{I}_d,\bar{\mathbf{Z}}_2\Big)\Big|\\
&\le\ C\,\frac{1}{(\sigma_A^2+\lambda)^2}\big(\|\beta\|^2+\lambda^{-2}\big)\Big(\frac{k+\sigma_A^2}{k}\Big(\sqrt{\frac{2\log(1/\epsilon)}{n}}+\sqrt{\frac{d}{n}}\,\Big)+1+\sigma_A^2+\lambda\Big)^2\\
&\quad\times\,\sigma_A^2\Big(\frac{\sqrt{d}+\sqrt{\log(2/\epsilon)}}{\sqrt{n(k-1)}}+\frac{(\sqrt{d}+\sqrt{\log(2/\epsilon)}\,)^2}{n(k-1)}\Big)\,.
\end{aligned}$$

In particular this implies that for $\lambda>0$ fixed, $d=O(n)$, $k\ge 2$ and $\sigma_A^2\le 1$,

$$\begin{aligned}
\Big|f_\lambda(\bar{\mathbf{Z}}_1,\bar{\mathbf{Z}}_2)-f_\lambda\Big(\bar{\mathbf{Z}}_2+\frac{(k-1)\sigma_A^2}{k}\mathbf{I}_d,\bar{\mathbf{Z}}_2\Big)\Big|\ &=O\Big(\max\Big\{1\,,\,\frac{d}{n}\Big\}^{3/2}\frac{\sqrt{d}}{\sqrt{n}}\,\frac{\sigma_A^2(1+\lambda^{-2})}{\sqrt{k}}\Big)\\
&=O\Big(\frac{\sigma_A^2}{\sqrt{k}}\,\frac{\sqrt{d}}{\sqrt{n}}\,\max\Big\{1\,,\,\frac{d}{n}\Big\}^{3/2}\Big)
\end{aligned}$$

with probability $1-O(e^{-\min\{d,n\}})$. By the definition of the Lévy-Prokhorov metric (D.8), we have

$$d_P\Big(f_\lambda(\bar{\mathbf{Z}}_1,\bar{\mathbf{Z}}_2)\,,\,f_\lambda\Big(\bar{\mathbf{Z}}_2+\frac{(k-1)\sigma_A^2}{k}\mathbf{I}_d,\bar{\mathbf{Z}}_2\Big)\Big)\ =\ O\Big(\frac{\sigma_A^2}{\sqrt{k}}\,\frac{\sqrt{d}}{\sqrt{n}}\,\max\Big\{1\,,\,\frac{d}{n}\Big\}^{3/2}\Big)\,. \tag{D.94}$$

Step 2: Approximate $f_0(\bar{\mathbf{Z}}_1,\bar{\mathbf{Z}}_2)$ by $f_\lambda\big(\bar{\mathbf{Z}}_2+\frac{(k-1)\sigma_A^2}{k}\mathbf{I}_d,\bar{\mathbf{Z}}_2\big)$. By Lemma D.17, we get that the assumptions of Lemma D.15 are fulfilled, and in particular in the proof of Lemma D.17 we have shown that the $1/(n\lambda^2)$ term in fact vanishes. This implies for λ small,

$$d_P\big(f_\lambda(\bar{\mathbf{Z}}_1,\bar{\mathbf{Z}}_2)\,,\,f_0(\bar{\mathbf{Z}}_1,\bar{\mathbf{Z}}_2)\big)\ =\ O(\lambda)\,.$$

Setting $\lambda=\sigma_A^{2/3}k^{-1/6}(d/n)^{1/2}$ and combining this bound with the d_P bound from above,

we obtain

$$d_P\Big(f_0(\bar{\mathbf{Z}}_1, \bar{\mathbf{Z}}_2)\,,\, f_{\frac{\sigma_A^{2/3}}{k^{1/6}}\frac{d^{1/2}}{n^{1/2}}}\Big(\bar{\mathbf{Z}}_2 + \frac{(k-1)\sigma_A^2}{k}\mathbf{I}_d, \bar{\mathbf{Z}}_2\Big)\Big) \;=\; O\bigg(\frac{\sigma_A^{2/3}}{k^{1/6}}\frac{d^{1/6}}{n^{1/6}}\max\Big\{1\,,\,\frac{d}{n}\Big\}^{1/2}\bigg)\,. \tag{D.95}$$

Step 3: Compute the limiting risk of $f_\lambda\big(\bar{\mathbf{Z}}_2 + \frac{(k-1)\sigma_A^2}{k}\mathbf{I}_d, \bar{\mathbf{Z}}_2\big)$. Define

$$\lambda_k \;:=\; \frac{(k-1)\sigma_A^2}{k} + \lambda\,, \qquad \sigma_k^2 \;:=\; \frac{k+\sigma_A^2}{k}\,, \qquad \tilde{\mathbf{Z}} \;:=\; \frac{1}{n}\sum\nolimits_{i=1}^n \eta_{i1}\eta_{i1}^\top\,,$$

where η_{i1}'s are the i.i.d. standard Gaussians defined in Lemma D.16. Recall also that

$$\bar{\mathbf{Z}}_2 \;=\; \frac{k+\sigma_A^2}{k}\,\frac{1}{n}\sum\nolimits_{i=1}^n \eta_{i1}\eta_{i1}^\top \;=\; \sigma_k^2\,\tilde{\mathbf{Z}}\,,$$

where η_{i1}'s are i.i.d. standard Gaussians. Observe that

$$\begin{aligned} f_\lambda\Big(\bar{\mathbf{Z}}_2 + \frac{(k-1)\sigma_A^2}{k}\,\mathbf{I}_d, \bar{\mathbf{Z}}_2\Big) \;&= \lambda^2\beta^\top\big(\bar{\mathbf{Z}}_2 + \lambda_k\,\mathbf{I}_d\big)^{-2}\beta + \frac{\sigma_\epsilon^2}{n}\mathrm{Tr}\big(\big(\bar{\mathbf{Z}}_2 + \lambda_k\,\mathbf{I}_d\big)^{-2}\bar{\mathbf{Z}}_2\big) \\ &= \frac{\lambda^2}{\lambda_k^2}\,f^{(1)}_{\lambda_k/\sigma_k^2}(\tilde{\mathbf{Z}}) + \frac{1}{\sigma_k^2}\,f^{(2)}_{\lambda_k/\sigma_k^2}(\tilde{\mathbf{Z}}, \tilde{\mathbf{Z}})\,. \end{aligned}$$

Denote the bias and variance parts of the risk defined in Hastie et al. (2022) as

$$R^{(1)}(\beta, \lambda, \gamma) := \|\beta\|^2\lambda^2\,\partial m_\gamma(-\lambda) \text{ and } R^{(2)}(\sigma, \lambda, \gamma) := \sigma^2\gamma\,\big(m_\gamma(-\lambda) - \lambda\partial m_\gamma(-\lambda)\big)\,,$$

where we recall $m_\gamma(z) = \frac{1-\gamma-z-\sqrt{(1-\gamma-z)^2-4\gamma z}}{2\gamma z}$. Now suppose k is fixed and $\lambda > 0$. By Corollary 5 of Hastie et al. (2022), we get that almost surely as $d, n \to \infty$ with $d/n \to \gamma$,

$$\begin{aligned} f_\lambda\Big(\bar{\mathbf{Z}}_2 + \frac{(k-1)\sigma_A^2}{k}\,\mathbf{I}_d, \bar{\mathbf{Z}}_2\Big) \;&\xrightarrow{a.s.}\; \frac{\lambda^2}{\lambda_k^2}\,R^{(1)}\Big(\beta, \frac{\lambda_k}{\sigma_k^2}, \gamma\Big) + \frac{1}{\lambda_k^2}\,R^{(2)}\Big(\sigma_\epsilon, \frac{\lambda_k}{\sigma_k^2}, \gamma\Big) \\ &= R^{(1)}\Big(\frac{\lambda}{\lambda_k}\beta, \frac{\lambda_k}{\sigma_k^2}, \gamma\Big) + R^{(2)}\Big(\frac{\sigma_\epsilon}{\sigma_k}, \frac{\lambda_k}{\sigma_k^2}, \gamma\Big) \\ &= R\Big(\frac{\lambda}{\lambda_k}\beta, \frac{\sigma_\epsilon}{\sigma_k}, \frac{\lambda_k}{\sigma_k^2}, \gamma\Big) \end{aligned} \tag{D.96}$$

for every $k \geq 2$ and $\sigma_A^2 \leq 1$. Note that Lemma D.17 shows that Assumptions 6.1 and 6.2 both hold under the isotropic setup, so the universality bounds in Proposition 6.10 hold. In the case $\lambda > 0$, combining the above first with (D.94) under the assumption that $\frac{\sigma_A^2}{\sqrt{k}}\frac{\sqrt{d}}{\sqrt{n}} = o(1)$ and then with Proposition 6.10, we have

$$f_\lambda(\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2) \;\xrightarrow{\mathbb{P}}\; \lim\, R\Big(\frac{\lambda}{\lambda_k}\beta, \frac{\sigma_\epsilon}{\sigma_k}, \frac{\lambda_k}{\sigma_k^2}, \gamma\Big)\,,$$

where $\lim$ denotes the limit under (6.20) with $\frac{\sigma_A^2}{\sqrt{k}}\frac{\sqrt{d}}{\sqrt{n}} = o(1)$. For the ridgeless case $\lambda = 0$, the same argument applies: Proposition 6.10 shows that $f_0(\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2)$ and $f_0(\bar{\mathbf{Z}}_1, \bar{\mathbf{Z}}_2)$ have the same distributional limit under (6.20), whereas (D.95) shows that $f_0(\bar{\mathbf{Z}}_1, \bar{\mathbf{Z}}_2)$ and $f_{\frac{\sigma_A^{2/3}}{k^{1/6}}\frac{d^{1/2}}{n^{1/2}}}\Big(\bar{\mathbf{Z}}_2 + \frac{(k-1)\sigma_A^2}{k}\mathbf{I}_d, \bar{\mathbf{Z}}_2\Big)$ have the same distributional limit under $\frac{\sigma_A^2}{\sqrt{k}}\frac{\sqrt{d}}{\sqrt{n}} = o(1)$. The distributional limit of $f_{\frac{\sigma_A^{2/3}}{k^{1/6}}\frac{d^{1/2}}{n^{1/2}}}\Big(\bar{\mathbf{Z}}_2 + \frac{(k-1)\sigma_A^2}{k}\mathbf{I}_d, \bar{\mathbf{Z}}_2\Big)$ under (6.20) is given by (D.96),

and we note that

$$R(\mathbf{0}, \sigma_\epsilon, \sigma_A^2, \gamma) \;=\; \lim_{\lambda \to 0^+} \; R\Big(\frac{\lambda}{\lambda + \sigma_A^2}\beta, \sigma_\epsilon, \lambda + \sigma_A^2, \gamma\Big)$$

exists by continuity as shown in Hastie et al. (2022).

□

Proof for Proposition 6.12: Two-stage augmentation The proof expresses the difference $R(\hat{\beta}_0^{(m)}) - \hat{L}_0^{(\mathrm{ora})} - \big\|\bar{\mathbf{X}}_1^\dagger \bar{\mathbf{X}}_\Delta(\tilde{\beta}_0^{(m)} - \beta)\big\|^2 - \sigma_\epsilon^2$ as two quantities involving averages and uses a concentration argument to show that they both converge to zero in probability.

We first recall from Appendix D.2.2 that

$$\hat{L}_0^{(\mathrm{ora})} \;=\; \beta^\top \big(\bar{\mathbf{X}}_1^\dagger \bar{\mathbf{X}}_1 - \mathbf{I}_d\big)^2 \beta + \frac{\sigma_\epsilon^2}{n} \mathrm{Tr}\big(\bar{\mathbf{X}}_1^\dagger \bar{\mathbf{X}}_2 \bar{\mathbf{X}}_1^\dagger\big) \;.$$

Meanwhile, recall that we have defined

$$\bar{\mathbf{X}}_\Delta \;=\; \frac{1}{n}\sum\nolimits_{i=1}^n \Big(\frac{1}{k}\sum\nolimits_{j=1}^k (\mathbf{V}_i + \xi_{ij})\Big)\Big(\frac{1}{k}\sum\nolimits_{j=1}^k \xi_{ij}\Big)^\top \;,$$

and denote $\bar{\mathbf{x}}_\epsilon = \frac{1}{n}\sum_{i=1}^n \big(\frac{1}{k}\sum_{j=1}^k (\mathbf{V}_i + \xi_{ij})\big)\epsilon_i$. Then we can express

$$\begin{aligned}
\hat{\beta}_0^{(m)} \;&= \bar{\mathbf{X}}_1^\dagger \Big(\frac{1}{n}\sum\nolimits_{i=1}^n \sum\nolimits_{j=1}^k (\mathbf{V}_i + \xi_{ij})(\mathbf{V}_i^\top \beta + \epsilon_i + \xi_{ij}^\top \tilde{\beta}_0^{(m)})\Big) \\
&= \bar{\mathbf{X}}_1^\dagger \bar{\mathbf{X}}_1 \beta + \bar{\mathbf{X}}_1^\dagger \bar{\mathbf{X}}_\Delta (\tilde{\beta}_\lambda^{(m)} - \beta) + \bar{\mathbf{X}}_1^\dagger \bar{\mathbf{x}}_\epsilon \;,
\end{aligned}$$

and therefore the risk of interest can be expressed as

$$\begin{aligned}
R(\hat{\beta}_0^{(m)}) \;&= \sigma_\epsilon^2 + \|\hat{\beta}_0^{(m)} - \beta\|^2 \\
&= \sigma_\epsilon^2 + \Big\|\big(\bar{\mathbf{X}}_1^\dagger \bar{\mathbf{X}}_1 - \mathbf{I}_d\big)\beta + \bar{\mathbf{X}}_1^\dagger \bar{\mathbf{X}}_\Delta (\tilde{\beta}_\lambda^{(m)} - \beta) + \bar{\mathbf{X}}_1^\dagger \bar{\mathbf{x}}_\epsilon\Big\|^2 \\
&\overset{(a)}{=} \sigma_\epsilon^2 + \beta^\top \big(\bar{\mathbf{X}}_1^\dagger \bar{\mathbf{X}}_1 - \mathbf{I}_d\big)^2 \beta + \big\|\bar{\mathbf{X}}_1^{-1} \bar{\mathbf{X}}_\Delta (\tilde{\beta}_\lambda^{(m)} - \beta)\big\|^2 \\
&\quad + 2(\tilde{\beta}_\lambda^{(m)} - \beta)^\top \bar{\mathbf{X}}_\Delta \bar{\mathbf{X}}_1^{-2} \bar{\mathbf{x}}_\epsilon + \bar{\mathbf{x}}_\epsilon^\top \bar{\mathbf{X}}_1^{-2} \bar{\mathbf{x}}_\epsilon \\
&= \sigma_\epsilon^2 + \hat{L}_0^{(\mathrm{ora})} + \big\|\bar{\mathbf{X}}_1^{-1} \bar{\mathbf{X}}_\Delta (\tilde{\beta}_\lambda^{(m)} - \beta)\big\|^2 \\
&\quad - 2\underbrace{(\tilde{\beta}_\lambda^{(m)} - \beta)^\top \bar{\mathbf{X}}_\Delta \bar{\mathbf{X}}_1^{-2} \bar{\mathbf{x}}_\epsilon}_{=:Q_1} - \underbrace{\Big(\bar{\mathbf{x}}_\epsilon^\top \bar{\mathbf{X}}_1^{-2} \bar{\mathbf{x}}_\epsilon - \frac{\sigma_\epsilon^2}{n}\mathrm{Tr}\big(\bar{\mathbf{X}}_1^\dagger \bar{\mathbf{X}}_2 \bar{\mathbf{X}}_1^\dagger\big)\Big)}_{=:Q_2} \;.
\end{aligned}$$

In (a), we have noted that $(\bar{\mathbf{X}}_1 \bar{\mathbf{X}}_1^\dagger - \mathbf{I}_d)\bar{\mathbf{X}}_1^\dagger = \mathbf{0}$ by the property of pseudo-inverse, which allows some cross-terms to vanish.

We now prove that Q_1 and Q_2 converge in probability to zero. By assumption, $\|\bar{\mathbf{X}}_1^\dagger\|_{op} + \|\bar{\mathbf{X}}_2\|_{op} + \|\bar{\mathbf{X}}_\Delta\|_{op} + \|\tilde{\beta}_\lambda^{(m)} - \beta\| \le C$ for some constant $C < \infty$ with probability $1 - o(1)$. Define the event

$$E \;:=\; \big\{\|\bar{\mathbf{X}}_1^\dagger\|_{op} + \|\bar{\mathbf{X}}_2\|_{op} + \|\bar{\mathbf{X}}_\Delta\|_{op} + \|\tilde{\beta}_\lambda^{(m)} - \beta\| \le C\big\} \;.$$

By the expression of $\bar{\mathbf{x}}_\epsilon$, we can write

$$Q_1 \;=\; (\tilde{\beta}_\lambda^{(m)} - \beta)^\top \bar{\mathbf{X}}_\Delta \bar{\mathbf{X}}_1^{-2} \bar{\mathbf{x}}_\epsilon \;=\; \frac{1}{n} \sum\nolimits_{i=1}^n (\tilde{\beta}_\lambda^{(m)} - \beta)^\top \bar{\mathbf{X}}_\Delta \bar{\mathbf{X}}_1^{-2} \Big(\frac{1}{k} \sum\nolimits_{j \le k} (\mathbf{V}_i + \xi_{ij}) \Big) \epsilon_i \; .$$

Conditioning on $\tilde{\mathcal{X}} = (V_i, \xi_{ij})_{i \le n, j \le k}$, we get that almost surely

$$\begin{aligned}
\mathbb{E}[\, Q_1 \,|\, \tilde{\mathcal{X}} \,] \; &= 0 \; , \\
\mathrm{Var}[\, Q_1 \,|\, \tilde{\mathcal{X}} \,] \; &= \frac{\sigma_\epsilon^2}{n^2} \sum_{i=1}^n \Bigg(\Big(\frac{1}{k} \sum_{j \le k} (\mathbf{V}_i + \xi_{ij}) \Big)^\top \bar{\mathbf{X}}_1^{-2} \bar{\mathbf{X}}_\Delta (\tilde{\beta}_\lambda^{(m)} - \beta) \\
&\qquad\qquad (\tilde{\beta}_\lambda^{(m)} - \beta)^\top \bar{\mathbf{X}}_\Delta \bar{\mathbf{X}}_1^{-2} \Big(\frac{1}{k} \sum_{j \le k} (\mathbf{V}_i + \xi_{ij}) \Big) \Bigg) \\
&= \frac{\sigma_\epsilon^2}{n} (\tilde{\beta}_\lambda^{(m)} - \beta)^\top \bar{\mathbf{X}}_\Delta \bar{\mathbf{X}}_1^{-2} \bar{\mathbf{X}}_2 \bar{\mathbf{X}}_1^{-2} \bar{\mathbf{X}}_\Delta (\tilde{\beta}_\lambda^{(m)} - \beta) \\
&\le \frac{\sigma_\epsilon^2}{n} \|\bar{\mathbf{X}}_2\|_{op} \|\bar{\mathbf{X}}_1^\dagger\|_{op}^4 \|\bar{\mathbf{X}}_\Delta\|_{op}^2 \|\tilde{\beta}_\lambda^{(m)} - \beta\|^2 \; ,
\end{aligned}$$

which is $O(n^{-1})$ on the event E. Therefore by splitting the probability according to E and applying the Markov's inequality, we obtain that for any $t > 0$,

$$\begin{aligned}
\mathbb{P}(|Q_1| > t) \; &\le \mathbb{P}(|Q_1| > t, E) + \mathbb{P}(E^c) \\
&= \mathbb{E}\big[\, \mathbb{P}(|Q_1| > t \,|\, \tilde{\mathcal{X}}\,) \, \mathbb{I}_E \big] + o(1) \\
&\le t^{-2} \, \mathbb{E}\big[\, \mathrm{Var}[Q_1 \,|\, \tilde{\mathcal{X}}] \, \mathbb{I}_E \big] + o(1) \;=\; o(1) \; ,
\end{aligned}$$

i.e. Q_1 converges to zero in probability. Q_2 can be handled by a similar argument: First note that $\mathbb{E}[Q_2] = 0$ since

$$\begin{aligned}
\mathbb{E}\big[\bar{\mathbf{x}}_\epsilon^\top \bar{\mathbf{X}}_1^{-2} \bar{\mathbf{x}}_\epsilon \,\big|\, \tilde{\mathcal{X}} \big] \; &= \frac{1}{n^2} \sum\nolimits_{i=1}^n \mathbb{E}\Big[\epsilon_i \Big(\frac{1}{k} \sum_{j \le k} (\mathbf{V}_i + \xi_{ij}) \Big)^\top \bar{\mathbf{X}}_1^{-2} \Big(\frac{1}{k} \sum_{j \le k} (\mathbf{V}_i + \xi_{ij}) \Big) \epsilon_i \,\Big|\, \tilde{\mathcal{X}} \Big] \\
&= \frac{\sigma_\epsilon^2}{n} \mathrm{Tr}\big(\bar{\mathbf{X}}_1^\dagger \bar{\mathbf{X}}_2 \bar{\mathbf{X}}_1^\dagger \big) \; .
\end{aligned}$$

While the expression of $\mathrm{Var}\big[Q_2 \,\big|\, \tilde{\mathcal{X}} \big]$ involves a complicated expansion of four sums, we note that since ϵ_i is zero-mean and independent, the only non-vanishing terms are of the form $\epsilon_i^2 \epsilon_{i'}^2$ with $i \neq i'$, with a multiplicity of $O(n^2)$, and ϵ_i^4, with a multiplicity of $O(n)$. Therefore, conditioning on the event E, we have that

$$\mathrm{Var}\big[Q_2 \,\big|\, \tilde{\mathcal{X}} \big] \;=\; O(n^{-2}) \;=\; o(1) \; ,$$

and applying the same argument of splitting the probability according to E followed by Markov's inequality gives that Q_2 converges to zero in probability. In summary, we have proved the desired statement that

$$R(\hat{\beta}_0^{(m)}) - \big(\sigma_\epsilon^2 + \hat{L}_0^{(\mathrm{ora})} + \big\| \bar{\mathbf{X}}_1^{-1} \bar{\mathbf{X}}_\Delta (\tilde{\beta}_\lambda^{(m)} - \beta) \big\|^2 \big) \;\xrightarrow{\mathbb{P}}\; 0 \; .$$

□

Appendix E

Proofs for Section 7.1.2

In this appendix, we prove Theorem 7.1 and Corollary 7.2. The proof recipe is similar to that of a standard CGMT: We start by proving a Gaussian min-max theorem (GMT) on discrete sets in Lemma E.1, proceed to extend it to compact sets in Lemma E.2, and then prove the results in Theorem 7.1. Corollary 7.2 then follows directly from Theorem 7.1(ii).

As with the standard CGMT, the Gaussian min-max theorem (GMT) on discrete sets is proved for a surrogate optimisation problem. Let $(\xi_l)_{l\leq M}$ be a collection of univariate standard Gaussians independent of $\mathbf{H}$, and define

$$\begin{aligned}\Psi^{\xi}_{\mathcal{S}_d,\mathcal{S}_n} &:= \min_{w\in\mathcal{S}_d}\max_{u\in\mathcal{S}_n} L^{\xi}_{\Psi}(w,u)\ ,\\ &\text{where}\quad L^{\xi}_{\Psi}(w,u) := w^\top \mathbf{H}u + \sum\nolimits_{l=1}^{M}\xi_l\,\|w\|_{\Sigma^{(l)}}\|u\|_{\tilde{\Sigma}^{(l)}} + f(w,u)\ .\end{aligned}$$

We also recall the risk $\psi_{\mathcal{S}_n,\mathcal{I}_n}$ of the auxiliary optimisation defined in Theorem 7.1.

Lemma E.1 (GMT on discrete sets)**.** *Let $\mathcal{I}_d\subseteq\mathbb{R}^d$, $\mathcal{I}_n\subseteq\mathbb{R}^n$ be discrete sets, and f be finite on $\mathcal{I}_d\times\mathcal{I}_n$. Then for all $c\in\mathbb{R}$,*

$$\mathbb{P}\big(\Psi^{\xi}_{\mathcal{I}_d,\mathcal{I}_n}\geq c\big)\ \geq\ \mathbb{P}\big(\psi_{\mathcal{I}_d,\mathcal{I}_n}\geq c\big)\ .$$

Proof of Lemma E.1. Similar to the proof for the standard GMT (see e.g. proof of Lemma A.1.1 of Thrampoulidis (2016)), the proof relies on an application of Gordon's Gaussian comparison inequality (see e.g. Corollary 3.13 of Ledoux and Talagrand (1991)) applied to two suitably defined Gaussian processes. Consider the two centred Gaussian processes indexed on the set $\mathcal{I}_d\times\mathcal{I}_n$:

$$\begin{aligned}Y_{w,u} &:= w^\top\mathbf{H}u + \sum\nolimits_{l=1}^{M}\xi_l\,\|w\|_{\Sigma^{(l)}}\|u\|_{\tilde{\Sigma}^{(l)}}\ ,\\ X_{w,u} &:= \sum\nolimits_{l=1}^{M}\big(\|w\|_{\Sigma^{(l)}}\mathbf{h}_l^\top\big(\tilde{\Sigma}^{(l)}\big)^{1/2}u + w^\top\big(\Sigma^{(l)}\big)^{1/2}\mathbf{g}_l\|u\|_{\tilde{\Sigma}^{(l)}}\big)\ .\end{aligned}$$

To compare their second moments, we use the independence of $\mathbf{H}$ and $\{\xi_l\}_{l\leq M}$ as well as the independence of $(\mathbf{h}_l,\mathbf{g}_l)_{l\leq M}$: For $w,w'\in\mathcal{I}_d$ and $u,u'\in\mathcal{I}_n$, we have

$$\mathbb{E}[Y_{w,u}Y_{w',u'}]-\mathbb{E}[X_{w,u}X_{w',u'}]$$

$$
\begin{aligned}
&\overset{(a)}{=} \mathbb{E}[w^\top \mathbf{H} u (w')^\top \mathbf{H} u'] + \sum_{l=1}^M \|w\|_{\Sigma^{(l)}} \|w'\|_{\Sigma^{(l)}} \|u\|_{\tilde{\Sigma}^{(l)}} \|u'\|_{\tilde{\Sigma}^{(l)}} \\
&\quad - \sum_{l=1}^M \left(\|w\|_{\Sigma^{(l)}} \|w'\|_{\Sigma^{(l)}} \, u^\top \tilde{\Sigma}^{(l)} u' + w^\top \Sigma^{(l)} w' \, \|u\|_{\tilde{\Sigma}^{(l)}} \|u'\|_{\tilde{\Sigma}^{(l)}} \right) \\
&\overset{(b)}{=} \sum_{l=1}^M \Big(w^\top \Sigma^{(l)} w' \, u^\top \tilde{\Sigma}^{(l)} u' + \|w\|_{\Sigma^{(l)}} \|w'\|_{\Sigma^{(l)}} \|u\|_{\tilde{\Sigma}^{(l)}} \|u'\|_{\tilde{\Sigma}^{(l)}} \\
&\qquad\qquad - \|w\|_{\Sigma^{(l)}} \|w'\|_{\Sigma^{(l)}} \, u^\top \tilde{\Sigma}^{(l)} u' - w^\top \Sigma^{(l)} w' \|u\|_{\tilde{\Sigma}^{(l)}} \|u'\|_{\tilde{\Sigma}^{(l)}} \Big) \\
&= \sum_{l=1}^M \left(\|w\|_{\Sigma^{(l)}} \|w'\|_{\Sigma^{(l)}} - w^\top \Sigma^{(l)} w' \right) \left(\|u\|_{\tilde{\Sigma}^{(l)}} \|u'\|_{\tilde{\Sigma}^{(l)}} - u^\top \tilde{\Sigma}^{(l)} u' \right) . \qquad \text{(E.1)}
\end{aligned}
$$

In (a), we have used that ξ_l's, $\mathbf{h}_l$'s and $\mathbf{g}_l$'s are all standard Gaussians; in (b), we have used

$$
\begin{aligned}
\mathbb{E}[w^\top \mathbf{H} u (w')^\top \mathbf{H} u'] &= \sum_{i,i'=1}^n \sum_{j,j'=1}^d w_i w'_{i'} u_j u'_{j'} \, \mathbb{E}[H_{ij} H_{i'j'}] \\
&= \sum_{l=1}^M \sum_{i,i'=1}^n \sum_{j,j'=1}^d w_i \Sigma^{(l)}_{ii'} w'_{i'} u_j \tilde{\Sigma}^{(l)}_{jj'} u'_{j'} \\
&= \sum_{l=1}^M w^\top \Sigma^{(l)} w' \, u^\top \tilde{\Sigma}^{(l)} u' \, .
\end{aligned}
$$

By the positive semi-definiteness of $\Sigma^{(l)}$ and $\tilde{\Sigma}^{(l)}$, (E.1) is non-negative, and equals to zero when $w = w'$. This shows that the Gaussian processes $(Y_{w,u})_{w \in \mathcal{I}_d, u \in \mathcal{I}_n}$ and $(X_{w,u})_{w \in \mathcal{I}_d, u \in \mathcal{I}_n}$ verify the conditions of the Gaussian comparison inequality (Corollary 3.13 of Ledoux and Talagrand (1991)) and therefore for any real sequence $(\lambda_{w,u})_{w \in \mathcal{I}_d, u \in \mathcal{I}_n}$,

$$
\mathbb{P}\big(\cap_{w \in \mathcal{I}_d} \cup_{v \in \mathcal{I}_n} \{ Y_{w,u} \geq \lambda_{w,u} \} \big) \;\geq\; \mathbb{P}\big(\cap_{w \in \mathcal{I}_d} \cup_{v \in \mathcal{I}_n} \{ X_{w,u} \geq \lambda_{w,u} \} \big) \, .
$$

Choosing $\lambda_{w,u} = -f(w,u) + c$ yields that

$$
\mathbb{P}\Big(\min_{w \in \mathcal{I}_d} \max_{v \in \mathcal{I}_n} (Y_{w,u} + f(w,u)) \geq c \Big) \;\geq\; \mathbb{P}\Big(\min_{w \in \mathcal{I}_d} \max_{v \in \mathcal{I}_n} (X_{w,u} + f(w,u)) \geq c \Big) \, .
$$

Noting that the two min-max quantities correspond to $\Psi^{\xi}_{\mathcal{I}_d, \mathcal{I}_n}$ and $\psi_{\mathcal{I}_d, \mathcal{I}_n}$ concludes the proof. □

The next result extends Lemma E.1 to compact sets.

Lemma E.2 (GMT for compact sets)**.** *Suppose $\mathcal{S}_d \subset \mathbb{R}^p$ and $\mathcal{S}_n \subset \mathbb{R}^n$ are compact and f is continuous on $\mathcal{S}_d \times \mathcal{S}_n$. Then for all $c \in \mathbb{R}$,*

$$
\mathbb{P}(\Psi^{\xi}_{\mathcal{S}_d, \mathcal{S}_n} \geq c) \;\geq\; \mathbb{P}(\psi_{S_p, S_n} \geq c) \, .
$$

Proof of Lemma E.2. The proof is almost identical to the proof of standard GMT results for compact sets, now that we have established Lemma E.1: We show by a compactness argument that both losses only change a little when replacing $\mathcal{S}_d$ and $\mathcal{S}_n$ by their δ-nets $\mathcal{S}^\delta_p$ and $\mathcal{S}^\delta_n$, induced by the Euclidean norms on $\mathbb{R}^n$ and $\mathbb{R}^d$ respectively. The only difference from their proof is that we use a slightly different concentration inequality. Therefore we only set up the essential notation, highlight the differences and refer interested readers to the proof of Theorem 3.2.1 of Thrampoulidis (2016), found in Pg 185-187.

First fix some $\epsilon > 0$. Since f is continuous and thereby uniformly continuous on the compact set $\mathcal{S}_p^\delta \times \mathcal{S}_n^\delta$, there exists some $\delta = \delta(\epsilon) > 0$ such that for all $(w,u),(w',u') \in \mathcal{S}_d \times \mathcal{S}_n$ with $\|(w,u)-(w',u')\| \le \delta$, we have $\|f(w,u)-f(w',u')\| \le \epsilon$. Use this δ to form the δ-nets $\mathcal{S}_p^\delta$ and $\mathcal{S}_n^\delta$. We also write $\|\bullet\|_{op}$ as the operator norm of a matrix, and write

$$S \;\coloneqq\; \max_{1\le l\le M} \max\{\|\Sigma^{(l)}\|_{\mathrm{op}}\,,\, \|\tilde{\Sigma}^{(l)}\|_{\mathrm{op}}\} \quad \text{and} \quad K \;\coloneqq\; \max\Big\{ \sup_{w\in\mathcal{S}_d} \|w\|\,,\, \sup_{u\in\mathcal{S}_n} \|u\| \Big\}\,.$$

K is bounded since $\mathcal{S}_d$ and $\mathcal{S}_n$ are compact, and for $w \in \mathcal{S}_d$, $u \in \mathcal{S}_n$ and $l \le M$, we have

$$\|w\| \;\le\; K\,, \qquad \|w\|_{\Sigma^{(l)}} \;\le\; SK\,, \qquad \|u\| \;\le\; K\,, \qquad \|u\|_{\tilde{\Sigma}^{(l)}} \;\le\; SK\,.$$

Then by the same argument as the proof of Theorem 3.2.1 of Thrampoulidis (2016), there exists $w_1 \in \mathcal{S}_d$, $w_1' \in \mathcal{S}_p^\delta$ with $\|w_1 - w_1'\| \le \delta$ and $u_1 \in \mathcal{S}_n^\delta$ such that

$$\begin{aligned}\Delta_\Psi^\xi \;&\coloneqq\; \min\nolimits_{w\in\mathcal{S}_p^\delta} \max\nolimits_{u\in\mathcal{S}_n^\delta} L_\Psi^\xi(w,u) \;-\; \min\nolimits_{w\in\mathcal{S}_d} \max\nolimits_{u\in\mathcal{S}_n} L_\Psi^\xi(w,u) \\ &\le L_\Psi^\xi(w_1',u_1) - L_\Psi^\xi(w_1,u_1)\,.\end{aligned}$$

Computing the difference gives

$$\begin{aligned}\Delta_\Psi^\xi \;&\le (w_1'-w_1)^\top \mathbf{H} u_1 + \sum_{l=1}^M \xi_l \left(\|w_1'\|_{\Sigma^{(l)}} - \|w_1\|_{\Sigma^{(l)}}\right)\|u_1\|_{\tilde{\Sigma}^{(l)}} + \left(f(w_1',u_1) - f(w_1,u_1)\right) \\ &\le \delta\|\mathbf{H}\| K \;+\; SK \textstyle\sum_{l=1}^M |\xi_l|\, \|w_1' - w_1\|_{\Sigma^{(l)}} \;+\; |f(w_1',u_1) - f(w_1,u_1)| \\ &\le \delta K \|\mathbf{H}\| + \delta S^2 K \textstyle\sum_{l=1}^M |\xi_l| + \epsilon\,.\end{aligned}$$

We seek to control $\|\mathbf{H}\|$ and $\sum_{1=1}^M |\xi_l|$ via concentration inequalities. Let $\mathrm{vec}(\mathbf{H})$ denote the $\mathbb{R}^{pn}$-valued vector formed from the entries of $\mathbf{H}$, and $\Sigma_{\mathbf{H}} \coloneqq \mathrm{Var}[\mathrm{vec}(\mathbf{H})]$. Then we can express, for some $\mathbb{R}^{pn}$-valued standard Gaussian vector η,

$$\|\mathbf{H}\|^2 \;=\; \|\mathrm{vec}(\mathbf{H})\|^2 \;=\; \eta^\top \Sigma_{\mathbf{H}}\, \eta\,.$$

Then by a Chernoff bound, we have that for any $t > 0$,

$$\mathbb{P}(\|\mathbf{H}\| \ge t) \;\le\; \inf\nolimits_{a>0} e^{-at^2} \mathbb{E}\big[e^{a\|\mathbf{H}\|^2}\big] \;=\; \inf\nolimits_{a>0} e^{-at^2} \mathbb{E}\big[e^{a\,\eta^\top \Sigma_{\mathbf{H}} \eta}\big]$$

Applying the formula of the moment-generating function of a Gaussian quadratic form (see e.g. Rencher and Schaalje (2008)) followed by setting $a = \frac{1}{4\|\Sigma_{\mathbf{H}}\|_{op}}$, we obtain

$$\begin{aligned}\mathbb{P}(\|\mathbf{H}\| \ge t) \;&\le\; \inf\nolimits_{a>0} \frac{e^{-at^2}}{\sqrt{\det(I_{pn} - 2a\Sigma_{\mathbf{H}})}} \;\le\; \frac{e^{-t^2/(4\|\Sigma_{\mathbf{H}}\|_{op})}}{\sqrt{\det(I_{pn} - \frac{1}{2\|\Sigma_{\mathbf{H}}\|_{op}}\Sigma_{\mathbf{H}})}} \\ &\le 2^{pn/2}\, e^{-t^2/(4\|\Sigma_{\mathbf{H}}\|_{op})}\,. \qquad \text{(E.2)}\end{aligned}$$

On the other hand, a standard concentration result on univariate Gaussians yields

$$\mathbb{P}(|\xi_l| > t) \;\le\; 2e^{-t^2/2}\,.$$

Taking a union bound, we obtain that for any $t > 0$,

$$\mathbb{P}\big(\Delta_\Psi^\xi \le \delta K t + \delta S^2 K M t + \epsilon \big) \ge 1 - 2^{pn/2}\, e^{-t^2/(4\|\Sigma_{\mathbf{H}}\|_{op})} - 2Me^{-t^2/2} ,$$

and therefore for any $c \in \mathbb{R}$ and $t > 0$,

$$\begin{aligned}
&\mathbb{P}\big(\min_{w\in\mathcal{S}_d} \max_{u\in\mathcal{S}_n} L_\Psi^\xi(w,u) \ge c - \delta K t - \delta S^2 K M t - \epsilon \big) \\
&\quad \ge \mathbb{P}\big(\min_{w\in\mathcal{S}_p^\delta} \max_{u\in\mathcal{S}_n^\delta} L_\Psi^\xi(w,u) \ge c \big) - 2^{pn/2}\, e^{-t^2/(4\|\Sigma_{\mathbf{H}}\|_{op})} - 2e^{-t^2/2} .
\end{aligned} \tag{E.3}$$

A similar argument as in the proof of Theorem 3.2.1 of Thrampoulidis (2016) shows that, there exists $w_2 \in \mathcal{S}_p^\delta$, $u_2 \in \mathcal{S}_d$ and $u_2' \in \mathcal{S}_n^\delta$ with $\|u_2 - u_2'\| \le \delta$ such that

$$\begin{aligned}
&\min_{w\in\mathcal{S}_p^\delta} \max_{u\in\mathcal{S}_n^\delta} L_\psi(w,u) - \min_{w\in\mathcal{S}_d} \max_{u\in\mathcal{S}_n} L_\psi(w,u) \ge L_\psi(w_2,u_2') - L_\psi(w_2,u_2) \\
&\quad = \sum_{l=1}^M \Big(\|w_2\|_{\Sigma^{(l)}} \mathbf{h}_l^\top \big(\tilde{\Sigma}^{(l)}\big)^{1/2} (u_2' - u_2) + w_2^\top \big(\Sigma^{(l)}\big)^{1/2} \mathbf{g}_l \big(\|u_2'\|_{\tilde{\Sigma}^{(l)}} - \|u_2\|_{\tilde{\Sigma}^{(l)}} \big) \Big) \\
&\qquad + \big(f(w_2,u_2') - f(w_2,u_2) \big) \\
&\quad \ge -\delta S^2 K \sum_{l=1}^M (\|\mathbf{h}_l\| + \|\mathbf{g}_l\|) - \epsilon .
\end{aligned}$$

Applying (E.2) to each $\|\mathbf{h}_l\|$ and $\|\mathbf{g}_l\|$ yields that, for any $t > 0$ and $1 \le l \le M$,

$$\mathbb{P}(\|\mathbf{h}_l\| \ge t) \le 2^{n/2} e^{-t^2/4} \qquad \text{and} \qquad \mathbb{P}(\|\mathbf{g}_l\| \ge t) \le 2^{p/2} e^{-t^2/4} .$$

Taking another union bound, we get that for any $t > 0$,

$$\begin{aligned}
&\mathbb{P}\big(\min_{w\in\mathcal{S}_d} \max_{u\in\mathcal{S}_n} L_\psi(w,u) \ge c + 2\delta S^2 K M t + \epsilon \big) \\
&\le \mathbb{P}\big(\min_{w\in\mathcal{S}_d^\delta} \max_{u\in\mathcal{S}_n^\delta} L_\psi(w,u) \ge c \big) + 2^{n/2} M e^{-t^2/4} + 2^{p/2} M e^{-t^2/4} .
\end{aligned} \tag{E.4}$$

Now by Lemma E.1, we have

$$\mathbb{P}\big(\min_{w\in\mathcal{S}_n^\delta} \max_{u\in\mathcal{S}_d^\delta} L_\psi(w,u) \ge c \big) \le \mathbb{P}\big(\min_{w\in\mathcal{S}_p^\delta} \max_{u\in\mathcal{S}_n^\delta} L_\Psi^\xi(w,u) \ge c \big) .$$

Combining this with (E.3) and (E.4) yields

$$\begin{aligned}
&\mathbb{P}\big(\min_{w\in\mathcal{S}_n} \max_{u\in\mathcal{S}_d} L_\psi(w,u) \ge c + 2\delta S^2 K M t + \epsilon \big) \\
&\le \mathbb{P}\big(\min_{w\in\mathcal{S}_n} \max_{u\in\mathcal{S}_d} L_\Psi^\xi(w,u) \ge c - \delta K t - \delta S^2 K M t - \epsilon \big) \\
&\qquad + 2^{n/2} M e^{-t^2/4} + 2^{p/2} M e^{-t^2/4} + 2^{np/2}\, e^{-t^2/(4\|\Sigma_{\mathbf{H}}\|_{op})} + 2e^{-t^2/2} .
\end{aligned}$$

The above holds for all $\epsilon > 0$ and $t > 0$. Set $t = \delta^{-1/2}$, take $\epsilon \to 0$ and choosing a sequence $\delta(\epsilon) \to 0$, we obtain that

$$\mathbb{P}\big(\min_{w\in\mathcal{S}_d} \max_{u\in\mathcal{S}_n} L_\psi(w,u) \ge c \big) \le \mathbb{P}\big(\min_{w\in\mathcal{S}_d} \max_{u\in\mathcal{S}_n} L_\Psi^\xi(w,u) \ge c \big) ,$$

i.e. $\mathbb{P}(\Psi_{\mathcal{S}_d,\mathcal{S}_n}^\xi \ge c) \ge \mathbb{P}(\psi_{S_p,S_n} \ge c)$. □

We are now ready to prove Theorem 7.1 and Corollary 7.2.

Proof of Theorem 7.1. The proof is almost identical to the proof of Theorem 3.3.1 of Thrampoulidis (2016) given the GMT result from Lemma E.2, and we focus on highlighting the differences. To prove the first bound in (i), we first apply Lemma E.2 to obtain that for all $c \in \mathbb{R}$,

$$\mathbb{P}\big(\min_{w\in\mathcal{S}_n} \max_{u\in\mathcal{S}_d} L_\Psi(w,u) + \sum_{l=1}^M \xi_l \|w\|_{\Sigma^{(l)}} \|u\|_{\tilde{\Sigma}^{(l)}} \le c\big) \;\le\; \mathbb{P}(\psi_{S_n,S_d} \le c)\;,$$

where $(\xi_l)_{l\le M}$ is a collection of univariate standard Gaussians independent of $\mathbf{H}$. First notice that, by conditioning on the event $\cap_{l\le M}\{\xi_l \ge 0\}$, we have that

$$\begin{aligned}\mathbb{P}(\Psi_{\mathcal{S}_p,\mathcal{S}_n} \le c) \;&= \mathbb{P}\big(\min_{w\in\mathcal{S}_n} \max_{u\in\mathcal{S}_d} L_\Psi(w,u) \le c\big)\\ &\le \mathbb{P}\big(\min_{w\in\mathcal{S}_n} \max_{u\in\mathcal{S}_d} L_\Psi(w,u) + \sum_{l=1}^M \xi_l \|w\|_{\Sigma^{(l)}} \|u\|_{\tilde{\Sigma}^{(l)}} \le c \,\big|\, \xi_1,\dots,\xi_M \le 0\big)\end{aligned}$$

which holds almost surely. Since ξ_l's are all independent and symmetric about zero, and there are 2^M possibilities for the signs of $(\xi_1,\dots,\xi_M)$, we obtain that

$$\begin{aligned}&\tfrac{1}{2^M}\,\mathbb{P}(\Psi_{\mathcal{S}_p,\mathcal{S}_n} \le c)\\ &\le \tfrac{1}{2^M}\mathbb{P}\big(\min_{w\in\mathcal{S}_n} \max_{u\in\mathcal{S}_d} L_\Psi(w,u) + \sum_{l=1}^M \xi_l \|w\|_{\Sigma^{(l)}} \|u\|_{\tilde{\Sigma}^{(l)}} \le c \,\big|\, \xi_1,\dots,\xi_M \le 0\big)\\ &\le \mathbb{P}\big(\min_{w\in\mathcal{S}_n} \max_{u\in\mathcal{S}_d} L_\Psi(w,u) + \sum_{l=1}^M \xi_l \|w\|_{\Sigma^{(l)}} \|u\|_{\tilde{\Sigma}^{(l)}} \le c\big)\\ &\le \mathbb{P}(\psi_{S_n,S_d} \le c)\;,\end{aligned}$$

which gives the desired statement.

The proof of the bound in (ii) is exactly the same as the proof of Theorem 3.3.1(ii) of Thrampoulidis (2016): It relies on the ability to apply a min-max theorem or a min-max inequality for swapping minimum and maximum under the stated convex-concave assumptions, as well as the invariance of the random term of the loss under a sign change. Both hold for our losses L_Ψ and L_ψ, since $\mathbf{H}$ in our L_Ψ is still zero-mean Gaussian, L_ψ is a linear sum of independent mean-zero Gaussian terms and all additional matrices $\Sigma^{(l)}$ and $\tilde{\Sigma}^{(l)}$ are positive semi-definite. We refer readers to the proof of Theorem 3.3.1(ii) of Thrampoulidis (2016) for a detailed derivation, and note that the only difference in our result is in that the coefficient from the first bound in (i) is now 2^M instead of 2.

The proof of (iii) is also exactly the same as the proof of Theorem 3.3.1(iii) of Thrampoulidis (2016), which only relies on the three assumptions, the statements (i) and (ii) proved above and a union bound. We again refer readers to the proof of Theorem 3.3.1(iii) of Thrampoulidis (2016) for a detailed derivation. □

Proof of Corollary 7.2. The result follows directly from Theorem 7.1(ii); see Corollary 3.3.2 of Thrampoulidis (2016). □