arXiv:2206.08420v1 [stat.ME] 16 Jun 2022

Generalised Bayesian Inference for Discrete Intractable Likelihood

Takuo Matsubara[1,3] Jeremias Knoblauch[2] François-Xavier Briol[2,3] Chris. J. Oates[1,3]

[1]Newcastle University, UK
[2]University College London, UK
[3]The Alan Turing Institute, UK

Abstract

Discrete state spaces represent a major computational challenge to statistical inference, since the computation of normalisation constants requires summation over large or possibly infinite sets, which can be impractical. This paper addresses this computational challenge through the development of a novel generalised Bayesian inference procedure suitable for discrete intractable likelihood. Inspired by recent methodological advances for continuous data, the main idea is to update beliefs about model parameters using a discrete Fisher divergence, in lieu of the problematic intractable likelihood. The result is a generalised posterior that can be sampled using standard computational tools, such as Markov chain Monte Carlo, circumventing the intractable normalising constant. The statistical properties of the generalised posterior are analysed, with sufficient conditions for posterior consistency and asymptotic normality established. In addition, a novel and general approach to calibration of generalised posteriors is proposed. Applications are presented on lattice models for discrete spatial data and on multivariate models for count data, where in each case the methodology facilitates generalised Bayesian inference at low computational cost.

1 Introduction

This paper focuses on statistical models for data defined on a discrete set $\mathcal{X}$, whose probability mass function p_θ involves a parameter θ to be inferred. In this setting, there is an urgent need for computational methodology applicable to models that are *intractable*, in the specific sense that

$$p_\theta(\boldsymbol{x}) = \frac{\tilde{p}_\theta(\boldsymbol{x})}{Z_\theta}, \qquad Z_\theta := \sum_{\boldsymbol{x}\in\mathcal{X}} \tilde{p}_\theta(\boldsymbol{x}), \tag{1}$$

where the positive function $\tilde{p}_\theta$ is straightforward to evaluate but direct computation of the normalising constant $Z_\theta \in (0, \infty)$ is impractical. This situation is ubiquitous in the discrete data context, since it is often impractical to compute a sum over a large or infinite discrete set. To limit scope, this paper considers generalised Bayesian inference where, to date, several computational approaches have been proposed. These approaches, which are recalled in Section 2, are mainly applicable in settings where it is possible to simulate data $\boldsymbol{x}$, conditional on the parameter θ. However, in several of the most scientifically important instances of (1), exact (or even approximate) simulation from the model is not practical.

Important examples of statistical models exhibiting these computational challenges include lattice models of spatial data (Moores et al., 2020; Jiang et al., 2021), statistical models for graph-valued data (Lusher et al., 2013), statistical models for multivariate count data (Inouye et al., 2017), and statistical mechanics models for the configuration of many-body systems (Yatsyshin et al., 2022). In each case, the normalising constant involves summation over a set whose cardinality is exponential in the dimension of the lattice, in the size of the nodal set of the graph, or even infinite, rendering both its direct computation and the simulation of data intractable in general.

To circumvent both computation of the normalising constant and simulation from the statistical model, Matsubara et al. (2022) proposed a generalised Bayesian posterior, called *KSD-Bayes*, which is based on a Stein discrepancy. The resulting generalised posterior is consistent and asymptotically normal, and thus shares many of the properties of the standard Bayesian posterior whilst admitting a form which does not require the computation of an intractable normalisation constant. However, a major limitation of KSD-Bayes is the dependence of the generalised posterior on a user-specified symmetric positive definite function, called a kernel, which determines precisely how beliefs are updated. In continuous domains, such as $\mathbb{R}^d$, there are several natural choices of kernel available, and their associated Stein discrepancies have been well-studied (Anastasiou et al., 2021). However, in discrete domains there are often no natural choices of kernel, or when natural choices exists (such as a heat kernel; Chung and Graham, 1997) they can be computationally impractical.

This paper presents *DFD-Bayes*, the first generalised Bayesian inference method tailored to inference with discrete intractable likelihood. The approach is based on a novel discrete version of the Fisher divergence which, in contrast to KSD-Bayes, does not require a kernel to be specified. Further, the DFD-Bayes posterior has computational complexity $O(nd)$, where n is the number of data and d is the data dimension, which compares favourably to the KSD-Bayes computational complexity of $O(n^2 d)$. The DFD-Bayes methodology is supported by asymptotic guarantees, presented in Section 3, and empirical results, in Section 4, demonstrate state-of-the-art performance in the applications considered. Before setting out the proposed methodology, we first we review related work in Section 2.

2 Background

The aim of this section is to briefly review existing Bayesian and generalised Bayesian methodology for intractable statistical models, extending the discussion to include both continuous and discrete data. Frequentist estimation for intractable models is not discussed (we refer the reader to e.g. Hyvärinen, 2005; Takenouchi and Kanamori, 2017).

Approximate Likelihood Faced with an intractable model, a pragmatic approach is simply to employ standard Bayesian inference with a tractable approximation to the likelihood (e.g. Bhattacharyya and Atchade, 2019). A classical example of approximate likelihood is the pseudolikelihood of Besag (1974), which replaces the joint probability mass function of the data with a product of conditional probability mass functions, each of which is sufficiently low-dimensional (or otherwise tractable enough) to permit normalising constants to be computed. Generalisations of this approach are sometimes referred to as composite likelihood (Varin et al., 2011). These approximations are usually model-specific, and analysis of the approximation error may be difficult in general (Lindsay et al., 2011).

Simulation-Based Methods One class of intractable statistical models that has been explored in detail are models for which it is possible to simulate data $\boldsymbol{x}$ conditional on the parameter θ. A well-known approach to inference in this class of models is the exchange algorithm of Møller et al. (2006) and Murray et al. (2006), which constructs a Markov chain on an extended state space for which the standard Bayesian posterior occurs as a marginal. Simulation of the Markov chain requires both exact simulation from the statistical model and evaluation of $\tilde{p}_\theta(\boldsymbol{x})$. Further methodological development has been focused on removing the requirement to evaluate $\tilde{p}_\theta(\boldsymbol{x})$, with approximate Bayesian computation (Marin et al., 2012), Bayesian synthetic likelihood (Price et al., 2018), MMD-Bayes (Cherief-Abdellatif and Alquier, 2020; Pacchiardi and Dutta, 2021) and the posterior boostrap (Dellaporta et al., 2022) emerging as likelihood-free methods, which require only that data can be simulated. Unfortunately, for many statistical models of discrete data, exact simulation (the state-of-the-art being e.g. Propp and Wilson, 1998) from the model is impractical.

Markov Chain-Based Methods Another pragmatic approach is to substitute exact simulations with approximate simulations, such as obtained from a Markov chain that leaves the posterior invariant. This

idea has been demonstrated to work in specific instances; see Liang (2010); Caimo and Friel (2011); Everitt (2012) or the review of Park and Haran (2018). The main drawback of these approaches, as far as this paper is concerned, is that they require the design of a rapidly mixing Markov chain on a possibly large (or infinite) discrete set. As such, these methods require bespoke implementations for each class of statistical model considered, and for many models of interest appropriate Markov chains have yet to be developed. Thus Markov chain-based methods do not represent a general solution to discrete intractable likelihood.

Russian Roulette The pseudo-marginal approach justifies replacing the intractable likelihood $p_\theta(\boldsymbol{x})$ with a positive unbiased estimator $\hat{p}_\theta(\boldsymbol{x})$ of the likelihood in the context of a Metropolis–Hastings algorithm (Andrieu and Roberts, 2009). The practical difficulty of this approach is to construct a positive unbiased estimator. Lyne et al. (2015) proposed the Russian roulette estimator for intractable statistical models, a simulation technique from the physics literature (Carter and Cashwell, 1975) which involves random truncation of the sum (or of an integral in the continuous context) defining the normalising constant. The Russian roulette estimator is unbiased but is not guaranteed to be positive, meaning that post hoc re-weighting of the Markov chain sample path is required. The ergodicity of Russian roulette has not, to the best of our knowledge, been theoretically studied (see the discussion in Wei and Murray, 2017). Further, the mixing time of the Markov chain is known to be sensitive to the variance of $\hat{p}_\theta(\boldsymbol{x})$, which can be large for estimators based on random truncation (especially when there is no clear a priori ordering for the summands, which can occur in the discrete context). As such, the pseudo-marginal approach does not at present represent a general computational solution to intractable likelihood.

Generalised Bayesian Inference Motivated by the absence of generally applicable computational methodology for intractable likelihood, Matsubara et al. (2022) proposed a solution called KSD-Bayes. The setting for this approach is the nascent field of generalised Bayesian inference. Given a prior $\pi(\theta)$, a dataset $\{\boldsymbol{x}_i\}_{i=1}^n \subset \mathcal{X}$, and a constant $\beta > 0$, generalised Bayesian inference updates beliefs using a loss function $D_n(\theta)$, producing a generalised posterior

$$\pi_n^D(\theta) \propto \pi(\theta) \exp(-\beta D_n(\theta)). \tag{2}$$

The standard Bayesian posterior is recovered by the negative log-likelihood $D_n(\theta) = -\sum_{i=1}^n \log p_\theta(\boldsymbol{x}_i)$, while several alternative loss functions have been developed to confer robustness in settings where the statistical model is misspecified (see the survey in Bissiri et al. (2016) for the case of additive loss functions, Jewson et al. (2018) for the case for divergence-based loss functions, and Knoblauch et al. (2022) for further generalisation). KSD-Bayes (Matsubara et al., 2022) is distinguished among existing generalised Bayesian inference methods by its applicability to statistical models involving an intractable normalising constant (see also Section 4.2 of Giummolè et al., 2019). This was achieved by selecting $D_n(\theta)$ to be a Stein discrepancy between the statistical model p_θ and the empirical distribution of the dataset, which can be computed without the normalising constant. Strikingly, a fully conjugate treatment of the continuous exponential family model, and a straight-forward treatment of the discrete exponential family model using Markov chain Monte Carlo, is possible using KSD-Bayes; this in principle provides a solution to many of the aforementioned instances of intractable likelihood. However, the dependence of KSD-Bayes on a user-specified kernel renders the approach unattractive for discrete domains, where there are often no natural choices of kernel, or where natural choices[1] are computationally impractical. Furthermore, the $O(n^2 d)$ computational cost of KSD-Bayes is super-linear in the size of the dataset.

This paper presents general methodology for inferring the parameters of a intractable discrete statistical model. The main idea is to employ a novel discrete Fisher divergence as a loss function in a generalised Bayesian inference context. The resulting *DFD-Bayes* method does not require a choice of kernel, enjoys

[1] A natural choice is the heat kernel, whose origins lie in spectral graph theory (Chung and Graham, 1997). However, computation of the heat kernel requires a $O(D^3)$ cost where $D = \text{card}(\mathcal{X})$, which is often impractical. For example, the classical Ising model on a lattice $\mathcal{X} = \{0, 1\}^d$ has $D = 2^d$, while the Conway–Maxwell–Poisson model of Section 4.1 has $D = \infty$, meaning approximation of the heat kernel would be required.

theoretical guarantees, and can be computed in a cost $O(nd)$ that is linear in the size of the dataset. Full details of the DFD-Bayes approach are provided next.

3 Methodology

This section presents and analyses DFD-Bayes. First, we present a novel discrete formulation of the Fisher divergence in Section 3.1. DFD-Bayes is introduced in Section 3.2, where posterior consistency and a Bernstein–von Mises result are established. Section 3.3 presents a novel approach to calibration of generalised posteriors, which may be of independent interest. Limitations of DFD-Bayes are discussed in Section 3.4.

Notation Denote by $\mathcal{X}$ a countable set in which data are contained, and by Θ the set of permitted values for the parameter θ, where Θ is a Borel subset of $\mathbb{R}^p$ for some $p \in \mathbb{N}$. Probability distributions on $\mathcal{X}$ are identified with their probability mass functions, with respect to the counting measure on $\mathcal{X}$. The i-th coordinate of a function $f : \mathcal{X} \to \mathbb{R}^m$ is denoted by $f_i : \mathcal{X} \to \mathbb{R}$. For a probability distribution q on $\mathcal{X}$ and $m, p \in \mathbb{N}$, denote by $L^p(q, \mathbb{R}^m)$ the vector space of measurable functions $f : \mathcal{X} \to \mathbb{R}^m$ such that $\sum_{i=1}^m \mathbb{E}_{X\sim q}[f_i(X)^p] < \infty$. For $\boldsymbol{z} \in \mathbb{R}^m$, the Euclidean norm is denoted $\|\boldsymbol{z}\|$. A Dirac measure at $x \in \mathcal{X}$ is denoted by δ_x.

3.1 A Discrete Fisher Divergence

The Fisher divergence underpins several frequentist estimators for intractable statistical models, most notably score matching (Hyvärinen, 2005; Yu et al., 2016, 2019; Barp et al., 2019), and has been used in the context of Bayesian model selection (Dawid and Musio, 2015; Shao et al., 2019; Jewson and Rossell, 2021). It is classically defined for continuous domains; for (sufficiently regular) densities p and q on $\mathbb{R}^d$, the Fisher divergence is $\mathrm{FD}(p\|q) = \mathbb{E}_{X\sim q}[\|\nabla \log p(X) - \nabla \log q(X)\|^2]$ where ∇ denotes the gradient operator in $\mathbb{R}^d$. Its main advantage is that it can be computed without knowledge of the normalising constant[2] of p and, furthermore, expectations with respect to p are not required. The Fisher divergence was extended to discrete domains in Lyu (2009); Xu et al. (2022). However, existing work focuses on domains $\mathcal{X}$ of finite cardinality or countable one-dimensional models, and a technical contribution of this paper, which may be of independent interest, is to present an extension of Fisher divergence to discrete domains which may be a countably infinite set in multiple dimensions. The extended divergence satisfies the requirements of a proper local scoring rule and thus complements existing scoring rule methodology developed in the finite domain context in Dawid et al. (2012).

Standing Assumption 1. *Let $\mathcal{X} = S_1 \times \cdots \times S_d$, where each S_i, $i = 1, \ldots, d$, is a countable ordered set with more than one element, and $d \in \mathbb{N}$.*

For any countable ordered set S, precisely one of the following must hold: (i) no smallest or largest elements of S exist; (ii) both a smallest element, $s_{\min}$, and a largest element, $s_{\max}$, exist; (iii) only $s_{\min}$ exists; (iv) only $s_{\max}$ exists. Without loss of generality, we will identify the case (iv) with (iii) by reversing the ordering of S. In addition, it will be useful to extend the domains S_i to include an additional state (not part of the ordering), denoted $\star$, and to this end we let $S_i^\star = S_i \cup \{\star\}$ and $\mathcal{X}^\star = S_1^\star \times \cdots \times S_d^\star$. A function $h : \mathcal{X} \to \mathbb{R}$ extends to a function $h : \mathcal{X}^\star \to \mathbb{R}$ by setting $h(\boldsymbol{x}) = 0$ whenever any of the coordinates of $\boldsymbol{x}$ are equal to $\star$.

Definition 1. *Let S be a countable ordered set. For consecutive elements $r < s < t$ in S we let $s^- := r$ and $s^+ := t$. If both $s_{\min}$ and $s_{\max}$ exist, we let $s_{\min}^- := s_{\max}$ and $s_{\max}^+ := s_{\min}$ or, if only $s_{\min}$ exists, we let $s_{\min}^- := \star$ and $\star^+ = s_{\min}$. For $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathcal{X}$, define $\boldsymbol{x}^{i+} := (x_1, \ldots, x_i^+, \ldots, x_d)$ and $\boldsymbol{x}^{i-} := (x_1, \ldots, x_i^-, \ldots, x_d)$.*

The above structure can be exploited to define operators for $\mathcal{X}$ that are analogous to the gradient and divergence operators for $\mathbb{R}^d$:

[2]The Fisher divergence depends only on $\nabla \log p$, equal to the ratio $(\nabla p)/p$, meaning it is sufficient to know p up to a normalising constant.

Definition 2. *For $h : \mathcal{X} \to \mathbb{R}$, define the forward difference and backward difference operators by*

$$\nabla^+ h(\boldsymbol{x}) := \begin{bmatrix} h(\boldsymbol{x}^{1+}) - h(\boldsymbol{x}) \\ \vdots \\ h(\boldsymbol{x}^{d+}) - h(\boldsymbol{x}) \end{bmatrix}, \qquad \nabla^- h(\boldsymbol{x}) := \begin{bmatrix} h(\boldsymbol{x}) - h(\boldsymbol{x}^{1-}) \\ \vdots \\ h(\boldsymbol{x}) - h(\boldsymbol{x}^{d-}) \end{bmatrix}.$$

For $h : \mathcal{X} \to \mathbb{R}^d$, define the forward divergence and backward divergence operators by $\nabla^+ \cdot h(\boldsymbol{x}) := \sum_{i=1}^d (h_i(\boldsymbol{x}^{i+}) - h_i(\boldsymbol{x}))$ and $\nabla^- \cdot h(\boldsymbol{x}) := \sum_{i=1}^d (h_i(\boldsymbol{x}) - h_i(\boldsymbol{x}^{i-}))$.

Based on Definitions 1 and 2, we can construct a divergence applicable to discrete domains $\mathcal{X}$, which we term a *discrete Fisher divergence*:

Definition 3. *Let p and q be positive probability distributions on $\mathcal{X}$, such that $(\nabla^- p)/p, (\nabla^- q)/q \in L^2(q, \mathbb{R}^d)$. The* discrete Fisher divergence *is defined as*

$$\mathrm{DFD}(p\|q) := \mathbb{E}_{X\sim q}\left[\left\| \frac{\nabla^- p(X)}{p(X)} - \frac{\nabla^- q(X)}{q(X)} \right\|^2\right]. \tag{3}$$

Proposition 1 justifies the name 'divergence' and offers an alternative, computable formula for (3).

Proposition 1. *The discrete Fisher divergence satisfies $\mathrm{DFD}(p\|q) \geq 0$ for any p, q, with equality if and only if $p = q$. Furthermore, it admits the following alternative formula*

$$\mathrm{DFD}(p\|q) = \mathbb{E}_{X\sim q}\left[\left\| \frac{\nabla^- p(X)}{p(X)} \right\|^2 + 2\nabla^+ \cdot \left(\frac{\nabla^- p(X)}{p(X)} \right) + \left\| \frac{\nabla^- q(X)}{q(X)} \right\|^2\right]. \tag{4}$$

The proof is provided in Appendix A.1. Note that $\mathrm{DFD}(p\|q)$ can be computed without the normalising constant of p, analogously to $\mathrm{FD}(p\|q)$ in $\mathbb{R}^d$. From Proposition 1, the discrete Fisher divergence between a model p_θ and an empirical distribution $p_n = \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{x}_i}$ corresponding to data $\{\boldsymbol{x}_i\}_{i=1}^n$, is computed as

$$\mathrm{DFD}(p_\theta\|p_n) \stackrel{\theta}{=} \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^d \left(\frac{p_\theta(\boldsymbol{x}_i^{j-})}{p_\theta(\boldsymbol{x}_i)} \right)^2 - 2\left(\frac{p_\theta(\boldsymbol{x}_i)}{p_\theta(\boldsymbol{x}_i^{j+})} \right) \tag{5}$$

where $\stackrel{\theta}{=}$ indicates equality up to an additive, θ-independent constant.

Remark 1. *The computational cost associated with evaluation of (5) is $O(nd)$, which improves on the $O(n^2 d)$ cost of kernel Stein discrepancy. Furthermore, if $\mathcal{X}$ is a finite set and count data are provided, indicating the number of times each of the elements of $\mathcal{X}$ occurred in the dataset, then the complexity of (5) reduces to $O(d)$, independent of the size of the dataset.*

Remark 2. *The discrete Fisher divergence can also be interpreted as a Stein discrepancy constructed based on an L^2-ball Stein set (Barp et al., 2019). This implies that discrete Fisher divergence is stronger than popular kernel Stein discrepancies; see Appendix C.*

3.2 A Generalised Posterior

We are now in a position to present DFD-Bayes.

Definition 4 (DFD-Bayes)**.** *Given a prior distribution π on Θ, a statistical model $p_\theta : \mathcal{X} \to (0, \infty)$ parametrised by $\theta \in \Theta$, and a dataset $\{\boldsymbol{x}_i\}_{i=1}^n$, the* DFD-Bayes posterior *is*

$$\pi_n^D(\theta) \propto \pi(\theta) \exp\left(-\beta n \,\mathrm{DFD}(p_\theta\|p_n)\right), \tag{6}$$

where $\beta \in (0, \infty)$ is a constant to be specified.

This is clearly a special case of the generalised posterior in (2) where $D_n(\theta) = n\,\mathrm{DFD}(p_\theta \| p_n)$. The role of n in (6) is to ensure correct scaling of the generalised posterior in the large n limit, while the appropriate choice of β is crucial in calibrating the coverage of the generalised posterior at finite sample sizes, and will be discussed in Section 3.3. For the moment, two important properties are highlighted:

Remark 3. *In contrast to standard Bayes posteriors in the presence of intractable likelihood, the DFD-Bayes posterior is directly amenable to standard Markov chain Monte Carlo. This is because* (5) *does not depend on the intractable constant, with the cost of evaluating* (6) *as low as* $O(d)$ *(c.f. Remark 1).*

Remark 4. *DFD-Bayes is invariant to order-preserving transformations of the dataset. This is in contrast to KSD-Bayes, which is not invariant to how the data are represented.*

The asymptotic behaviour of the standard Bayesian posterior is well-understood, with sufficient conditions for posterior consistency and asymptotic normality providing frequentist justification for Bayesian inference in the large data limit. Our attention now turns to establishing analogous conditions for DFD-Bayes. The setting for which we derive our theory is the following:

Standing Assumption 2. *The data* $\{\boldsymbol{x}_i\}_{i=1}^n$ *consist of independent samples from a probability distribution* p *on* $\mathcal{X}$*. The distribution* p *and the statistical model* p_θ *for these data satisfy* $(\nabla^- p)/p, (\nabla^- p_\theta)/p_\theta \in L^2(p, \mathbb{R}^d)$*, for all* $\theta \in \Theta$*.*

The setting of independent data is broad enough to contain important examples of discrete intractable likelihood, including the models studied in Section 4. The other assumption simply ensures that $\mathrm{DFD}(p_\theta \| p_n)$ is well-defined, due to Proposition 1. Under the setting above, a natural first requirement is that the statistical model is identifiable in the large data limit:

Assumption 1. *There exists a unique minimiser* θ_* *of* $\theta \mapsto \mathrm{DFD}(p_\theta \| p)$ *and there exists a sequence* $\{\theta_n\}_{n=1}^\infty$ *such that* θ_n *minimises* $\theta \mapsto \mathrm{DFD}(p_\theta \| p_n)$ *almost surely for all* n *sufficiently large. Further, there exists a bounded convex open set* $U \subseteq \Theta$ *such that* $\theta_* \in U$ *and* $\theta_n \in U$ *almost surely for all* n *sufficiently large.*

The existence of U in Assumption 1 essentially implies that for large enough n, we can restrict our theoretical analysis to a bounded subset $U \subseteq \Theta$. This part of the assumption is not restrictive: it can be enforced by re-parameterising the model p_θ so that its new parameter space is bounded and convex.[3] The existence of $\{\theta_n\}_{n=1}^\infty$ and θ_* is more difficult to assess in practice, since the true data generating distribution p is unknown. That being said, assuming their existence is a mild regularity condition; and common in the asymptotic analysis of Bayesian procedures (see e.g. van der Vaart, 1998, Section 10). It is worth highlighting that Assumption 1 does not require the model family $\{p_\theta \mid \theta \in \Theta\}$ to contain p, which is in contrast to the assumptions needed for the classical Bernstein–von Mises theorem (van der Vaart, 1998, Theorem 10.1). On the other hand, if the model family $\{p_\theta \mid \theta \in \Theta\}$ contains p uniquely, existence of θ_* is immediate since DFD is a divergence and hence $\mathrm{DFD}(p_\theta \| p) = 0$ if and only if $p_\theta = p$.

Our second main requirement is a technical condition on the derivatives and moments of the model, to ensure that the asymptotic limit is well-defined. It is helpful to introduce the shorthand $r_{j-}(\boldsymbol{x}, \theta) := p_\theta(\boldsymbol{x}^{j-})/p_\theta(\boldsymbol{x})$, and to let ∇_θ^s denote s-times differentiation with respect to θ. For a function $g : \Theta \to \mathbb{R}$, the derivative $\nabla_\theta^s g(\theta)$ takes a value in the s-times Cartesian product of $\mathbb{R}^p$.

Assumption 2. *Assume that* $\theta \mapsto p_\theta(\boldsymbol{x})$ *is three times continuously differentiable in* U *for any* $\boldsymbol{x} \in \mathcal{X}$*, and*

$$\mathbb{E}_{X\sim p}\left[\sup_{\theta\in U} \|\nabla_\theta^s r_{j-}(X^{j+}, \theta)\|\right] < \infty \qquad \text{and} \qquad \mathbb{E}_{X\sim p}\left[\sup_{\theta\in U} \|\nabla_\theta^s (r_{j-}(X, \theta)^2)\|\right] < \infty$$

for all $j = 1, \ldots, d$ *and* $s = 1, 2, 3$.

[3] For example, we can re-parameterise any unbounded parameter κ through the logistic function and define the invertible transformation $\theta = (1 + e^{-\kappa})^{-1} \in [0, 1]$.

In contrast to Assumption 1, it is easier to verify Assumption 2, as illustrated in Example 1. It considers the exponential family, a large class of models which encompasses the models in our experiments in Section 4. For example, any model on a space $\mathcal{X}$ of finite cardinality is an exponential family model (Amari, 2016, Ch. 2.2.2).

Example 1 (Exponential Family)**.** *Consider an exponential family model* $p_\theta(\boldsymbol{x}) \propto \exp(\eta(\theta) \cdot T(\boldsymbol{x}) + b(\boldsymbol{x}))$*, where* $\eta : \Theta \to \mathbb{R}^k$*,* $T : \mathcal{X} \to \mathbb{R}^k$ *and* $b : \mathcal{X} \to \mathbb{R}$ *for some* $k \in \mathbb{N}$*. For this model, we have* $r_{j-}(\boldsymbol{x}, \theta) = \exp(\eta(\theta) \cdot (T(\boldsymbol{x}^{j-}) - T(\boldsymbol{x})) + b(\boldsymbol{x}^{j-}) - b(\boldsymbol{x}))$*. Assumption 2 is satisfied if, for* $j = 1, \ldots, d$*, (i)* $\|\eta(\theta)\|$ *and* $\|\nabla_\theta^s \eta(\theta)\|$ *for* $s = 1, 2, 3$ *are bounded over* $\theta \in U$*, (ii)* $\|T(\boldsymbol{x}^{j-}) - T(\boldsymbol{x})\|$ *is bounded over* $\boldsymbol{x} \in \mathcal{X}$*, and (iii)* $\mathbb{E}_{X \sim p}[\exp(b(X^{j-}) - b(X))^2] < \infty$*. The requirements (ii) and (iii) are immediate if* $\mathcal{X}$ *is a finite set.*

The calculations that accompany Example 1 are provided in Appendix D.1. The following theorem establishes that both consistency and a Bernstein–von Mises theorem hold. The former implies that our generalised posterior concentrates around the population minimiser θ_* with probability 1 when $n \to \infty$. The latter establishes that our generalised posterior is normal around θ_* in the same asymptotic limit.

Theorem 1. *Suppose Assumptions 1 and 2 hold. Assume that the prior* π *is positive and continuous at* θ_**. Let* $B_\epsilon(\theta_*) := \{\theta \in \Theta \mid \|\theta - \theta_*\|_2 < \epsilon\}$*. Then for any* $\epsilon > 0$*,*

$$\int_{B_\epsilon(\theta_*)} \pi_n^D(\theta) \mathrm{d}\theta \xrightarrow{\text{a.s.}} 1 \qquad \text{as } n \to \infty. \tag{7}$$

Denote by $\widetilde{\pi}_n^D$ *a density on* $\mathbb{R}^d$ *of a random variable* $\sqrt{n}(\theta - \theta_n)$ *for* $\theta \sim \pi_n^D$*. If* $H_* := \beta \nabla_\theta^2 \operatorname{DFD}(p_\theta \| p)|_{\theta=\theta_*}$ *is nonsingular, then*

$$\int_{\mathbb{R}^p} \left| \widetilde{\pi}_n^D(\theta) - \frac{1}{Z_*} \exp\left(-\frac{1}{2}\theta \cdot H_* \theta\right) \right| \mathrm{d}\theta \xrightarrow{\text{a.s.}} 0 \qquad \text{as } n \to \infty, \tag{8}$$

where Z_* *is the normalising constant of* $\exp\left(-\frac{1}{2}\theta \cdot H_* \theta\right)$*.*

The proof of Theorem 1 is provided in Appendix A.2. The result was established using similar arguments from early work by Hooker and Vidyashankar (2014); Ghosh and Basu (2016) and extended techniques of Miller (2021); Matsubara et al. (2022). See also the recent review of Bochkina (2022).

3.3 A New Approach to Calibration of Generalised Posteriors

The weight β in (2) controls the scale of the generalised posterior, and the selection of an appropriate value for β is critical to ensure the generalised posterior is calibrated. The literature on this topic is underdeveloped, but two existing approaches stand out. The first approach was proposed in the recent review paper of Syring and Martin (2019). It consists of a new stochastic sequential update algorithm for choosing β, such that a 95% highest posterior density region coincides with a 95% confidence interval. Unfortunately, this approach leads to a large computational cost and is therefore often impractical. The second approach is due to Lyddon et al. (2019) and consists in setting β such that the scale of the posterior's asymptotic covariance matrix coincides with that of a frequentist counterpart with correct coverage. Matsubara et al. (2022) numerically showed that this approach is unstable when n is not large enough or when θ is high dimensional. In addition, the second approach does not take the prior π into account, because it depends only the generalised posterior's asymptotic covariance matrix.

In order to remedy some of these issues, the present paper proposes a novel selection criterion for β that can be viewed as a more analytically tractable alternative to Syring and Martin (2019). This criterion is applicable to generalised posteriors beyond DFD-Bayes and may therefore be of independent interest. Our approach consists of two steps: (i) computing minimisers of B "bootstrapped" losses and (ii) estimating an appropriate value of β using the closed-form expression in Theorem 2. In contrast to Syring and Martin (2019), step (ii) is non-iterative and exact. Additionally, computation of each minimiser in step (i) is embarrassingly parallel. Relative to the approach of Lyddon et al. (2019), the advantage of our method is

that it does not rely on asymptotic quantities, takes the prior into account, and maintains numerical stability even if the parameter θ is high-dimensional.

To describe the method we first define the minimiser $\theta_n = \arg\min_{\theta\in\Theta} D_n(\theta)$, where D_n is a specified loss function based on a dataset $\{\boldsymbol{x}_i\}_{i=1}^n$. To make the dependence on β explicit, we denote the posterior π_n^D by $\pi_{n,\beta}^D$. In step (i), bootstrap datasets $\{\boldsymbol{x}_i^{(b)}\}_{i=1}^n$, $b = 1, \ldots, B$, are generated by sampling each $\boldsymbol{x}_i^{(b)}$ uniformly with replacement from the original dataset. Then, for each bootstrap dataset, we compute a minimiser $\theta_n^{(b)} = \arg\min_{\theta\in\Theta} D_n^{(b)}(\theta)$, where the superscript indicates that $D_n^{(b)}$ is based the b^{th} bootstrap dataset. This leads to an empirical measure $\delta_\theta^B = \frac{1}{B}\sum_{b=1}^B \delta(\theta_n^{(b)})$ which approximates the sampling distribution of the estimator θ_n. In step (ii), we choose β to minimise a statistical divergence between $\pi_{n,\beta}^D$ and δ_θ^B. However, this is not straight-forward, since the majority of statistical divergences (e.g. Kullback–Liebler divergence) require the normalising constant of $\pi_{n,\beta}^D$ for every β. Interestingly, this is the same computational challenge posed by intractable likelihood. Our proposed approach is therefore to employ a divergence that circumvents computational of the normalisation constant; here we minimise the Fisher divergence in the continuous domain Θ:

$$\beta_* \in \arg\min_{\beta>0} \frac{1}{n}\sum_{b=1}^B \left\|\nabla \log \pi_{n,\beta}^D\big(\theta_n^{(b)}\big)\right\|^2 + 2\,\mathrm{Tr}\left(\nabla^2 \log \pi_{n,\beta}^D\big(\theta_n^{(b)}\big)\right). \tag{9}$$

This leads to an explicit score-matching estimator for β, circumventing intractability of (2), as will now be established.

Theorem 2. *Consider a generalised posterior $\pi_{n,\beta}^D$ with twice differentiable loss function $D_n : \Theta \to \mathbb{R}$. Suppose that there exists at least one $\theta_n^{(b)}$ s.t. $\nabla_\theta D_n(\theta_n^{(b)}) \neq 0$ and that $\sum_{b=1}^B \nabla_\theta D_n(\theta_n^{(b)}) \cdot \nabla_\theta \log \pi(\theta_n^{(b)}) + \mathrm{Tr}(\nabla_\theta^2 D_n(\theta_n^{(b)})) > 0$. Then β_* in (9) is unique, with*

$$\beta_* = \frac{\sum_{b=1}^B \nabla_\theta D_n(\theta_n^{(b)}) \cdot \nabla_\theta \log \pi(\theta_n^{(b)}) + \mathrm{Tr}(\nabla_\theta^2 D_n(\theta_n^{(b)}))}{\sum_{b=1}^B \|\nabla_\theta D_n(\theta_n^{(b)})\|^2} > 0. \tag{10}$$

The proof is provided in Appendix A.3. The condition in Theorem 2 directly implies existence and positivity of (10). One may replace this condition with more sophisticated condition e.g. on some appropriate local convexity of D_n and $\log \pi$ that can be guaranteed regardless of the computed values of $\{\theta_n^{(b)}\}_{b=1}^B$. However, in practice, computing (10) and verifying the existence and positivity directly is strikingly easier than validating the local convexity of D_n and $\log \pi$.

Note that (10) is straight-forward to compute whenever the loss D_n is amenable to automatic differentiation. For completeness, we also provide an explicit expression in Appendix D.2 for the case of the DFD-Bayes posterior with an exponential family model.

Remark 5. *Step (i) of our algorithm is embarrassingly parallelisable over bootstrap samples. Additionally, each component inside the sum in (10) can also be computed in parallel during step (ii). Overall, the computational cost can therefore be reduced linearly in the number of available cores K, and the total cost of step (ii) is $O(p^2 \times C \times B/K)$, where C is the cost of evaluating $D_n(\theta)$ and $\pi(\theta)$ at θ.*

3.4 Limitations

There are at least two important limitations of the DFD-Bayes methodology, which will now be discussed. First, DFD-Bayes was not derived as an approximation to standard Bayesian inference, and thus the semantics associated with the generalised posterior should not be confused with the semantics of standard Bayesian inference; see Bissiri et al. (2016); Knoblauch et al. (2022) for a detailed discussion of this point. In particular, we need to calibrate DFD-Bayes through the selection of β, which is not a feature of standard Bayesian inference under well-specified models. Although we expect our bootstrap approach to outperform

existing alternative approaches for small sample size n, it is possible that in those cases the bootstrap criterion for selecting β in Section 3.3 will fail, and in these circumstances the generalised posterior will fail to be calibrated.

Second, the generalised posterior may suffer from similar drawbacks to score-based methods for continuous data, including insensitivity to mixing proportions (Wenliang and Kanagawa, 2021). Indeed, for a two-component mixture model $p_\theta(\boldsymbol{x}) = (1-\theta)p_1(\boldsymbol{x}) + \theta p_2(\boldsymbol{x})$, we can compute the ratios

$$\rho_i := \left[\frac{\nabla^- p_\theta(\boldsymbol{x})}{p_\theta(\boldsymbol{x})}\right]_i = 1 - \frac{(1-\theta)p_1(\boldsymbol{x}^{i-}) + \theta p_2(\boldsymbol{x}^{i-})}{(1-\theta)p_1(\boldsymbol{x}) + \theta p_2(\boldsymbol{x})}$$

on which the discrete Fisher divergence is based. Suppose, informally, that the high probability regions R_1 of p_1 and R_2 of p_2 are separated, meaning that $p_2 \approx 0$ on R_1 and $p_1 \approx 0$ on R_2. Then these ratios are approximately independent of the parameter θ on $R_1 \cup R_2$, since $\rho_i \approx 1 - p_1(\boldsymbol{x}^{i-})/p_1(\boldsymbol{x})$ for $\boldsymbol{x} \in R_1$ and $\rho_i \approx 1 - p_2(\boldsymbol{x}^{i-})/p_2(\boldsymbol{x})$ for $\boldsymbol{x} \in R_2$. It follows that $\mathrm{DFD}(p_\theta \| p_n)$ is approximately independent of θ whenever the data $\{\boldsymbol{x}\}_{i=1}^n \subseteq R_1 \cup R_2$. Thus, although DFD-Bayes may be applied to mixture models, supported by the theoretical guarantees of Theorem 1, the inferences for mixing proportions so-obtained are likely to be data-inefficient.

4 Experimental Assessment

To complement the theoretical assessment we now provide a detailed empirical assessment, focusing on three important instances of discrete intractable likelihood. First, in Section 4.1 we consider a relatively simple model for over- and under-dispersed count data, called the Conway–Maxwell–Poisson model. Section 4.2 concerns an application to Ising-type models for discrete spatial data. Finally, we apply DFD-Bayes to perform inference for the parameters of flexible multivariate models for count data in Section 4.3. Source code to reproduce these experiments can be downloaded from `https://github.com/takuomatsubara/Discrete-Fisher-Bayes`.

4.1 Conway–Maxwell–Poisson Model

The first model we consider is a generalisation of the Poisson model for over- and under-dispersed count data, due to Conway and Maxwell (1962). This model is on $\mathcal{X} = \mathbb{N} \cup \{0\}$ (hence $d = 1$ and $\mathrm{card}(\mathcal{X}) = \infty$) and generalises the Poisson distribution through the inclusion of an additional parameter controlling how the data are dispersed. Since the work of Shmueli et al. (2005), this model has been used in a wide range of fields including transport, finance and retail. The model has two parameters $\theta \in \Theta = (0,\infty)^2 \cup ([0,1] \times \{0\})$ (and hence $p = 2$) and its probability mass function is given by $p_\theta(x) = \tilde{p}_\theta(x) Z_\theta^{-1}$ where $\tilde{p}_\theta(x) = (\theta_1)^x (x!)^{-\theta_2}$. The normalising constant is given by $Z_\theta = \sum_{y=0}^\infty \tilde{p}_\theta(y)$, which has no analytical form except for certain special cases of $\theta \in \Theta$, including the case $\theta_2 = 1$ for which the standard Poisson model is recovered.

This model is an ideal test-bed for DFD-Bayes: although the likelihood is formally intractable, it is relatively straightforward to directly approximate the normalising constant[4]. This enables a direct comparison with standard Bayesian inference in the case where the model is well-specified. To this end, we simulated two datasets from the model: (i) an under-dispersed case where $\theta^* = (4, 1.25)$, and (ii) an over-dispersed case where $\theta^* = (4, 0.75)$, shown in Figure 1 (left). Three inference methods were compared: standard Bayesian inference, the KSD-Bayes method of Matsubara et al. (2022), and the DFD-Bayes method we have proposed. The settings of KSD-Bayes are described in Appendix E.1.1. In each case, the prior π was taken to be the chi-squared distribution with 3 degrees of freedom for each of θ_1 and θ_2 independently. A Metropolis–Hastings algorithm was used to sample from all the posteriors; and details can be found in Appendix E.1.2. The weight β in DFD-Bayes and KSD-Bayes was calibrated by our approach described in Section 3.3.

Figure 1 (right) illustrates the posteriors, based on typical datasets of size $n = 2,000$. The estimated value of β_* was 1.91 for DFD-Bayes and 5.04 for KSD-Bayes in the over-dispersed case $\theta_2 = 0.75$, and 0.46

[4]The standard Bayesian inferences reported in this section used the approximation $Z_\theta \approx \sum_{y=0}^{99} \tilde{p}_\theta(y)$ and the associated approximate likelihood.

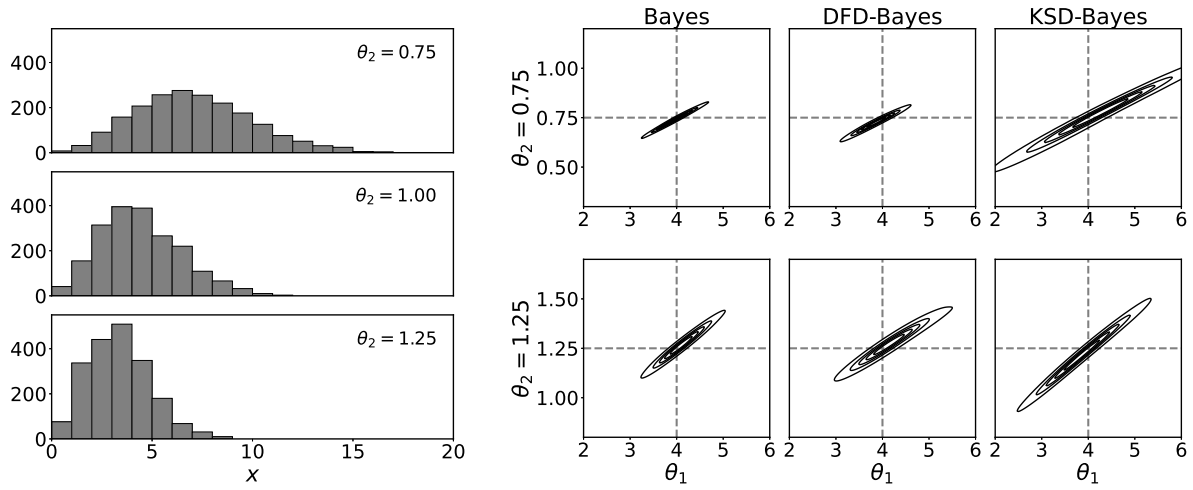

Figure 1: Comparison of standard Bayesian inference with the generalised posteriors from DFD-Bayes and KSD-Bayes on the Conway–Maxwell–Poisson model in the over-dispersed case $\theta_2 = 0.75$ and the under-dispersed case $\theta_2 = 1.25$ for $n = 2,000$.

for DFD-Bayes and 2.51 for KSD-Bayes in the under-dispersed case $\theta_2 = 1.25$. The left panel of Figure 2 displays the distribution of calibrated weight β_* as in Section 3.3 over multiple instances of the dataset, along with the values advocated in Lyddon et al. (2019). For both methods, the calibrated weight is estimated in a stable manner for this example.

The inferences obtained using DFD-Bayes resembled those obtained using standard Bayesian inference, irrespective of whether the data were over- or under-dispersed. Those obtained using KSD-Bayes were more conservative than standard Bayes and DFD-Bayes, although the maximum a posteriori estimator approximated the true parameter well. Note that the credible regions of the generalised posteriors can substantially differ from those of standard Bayesian inference; in our approach a credible region of a generalised posterior is calibrated with reference to the distribution of a corresponding frequentist estimator estimated by bootstrapping, leading to approximately correct frequentist coverage as shown in Figure 2 (middle). Calibration led to improved inference outcomes for both DFD-Bayes and KSD-Bayes. In the KSD-Bayes case for example, the value of $\beta_* \geq 1$ intensified the concentration around the true parameter by placing more importance on the loss than the prior. In addition, our approach to calibration is relatively more conservative than Lyddon et al. (2019) because the prior is taken into account.

There is a stark difference in computational cost between DFD-Bayes and KSD-Bayes[5], as demonstrated in the right panel of Figure 2. Indeed, the computational cost of DFD-Bayes is seen to increase linearly with n, while the cost of KSD-Bayes increases quadratically.

Finally, to assess performance in a real-world data setting, we apply DFD-Bayes to infer the parameters of a Conway–Maxwell–Poisson model using the sales dataset of Shmueli et al. (2005). All relevant details are contained in Appendix E.1.3. Figure 3 compares our fitted model to a standard Bayesian analysis using the Poisson distribution, which is the closest analysis one can perform without confronting an intractable likelihood. As observed in the central panel of Figure 3, the Poisson model is not able to capture over-dispersion of the data, whereas the Conway–Maxwell–Poisson model fitted using DFD-Bayes, shown in the right panel, provides a reasonable fit. The DFD-Bayes posterior (left) appears approximately normal, in line with Theorem 1.

[5]The cost of standard Bayesian inference in this experiment is entirely determined by the accuracy with which the normalisation constant is approximated; since direct approximation of the normalisation constant is infeasible in general, we do not report this cost.

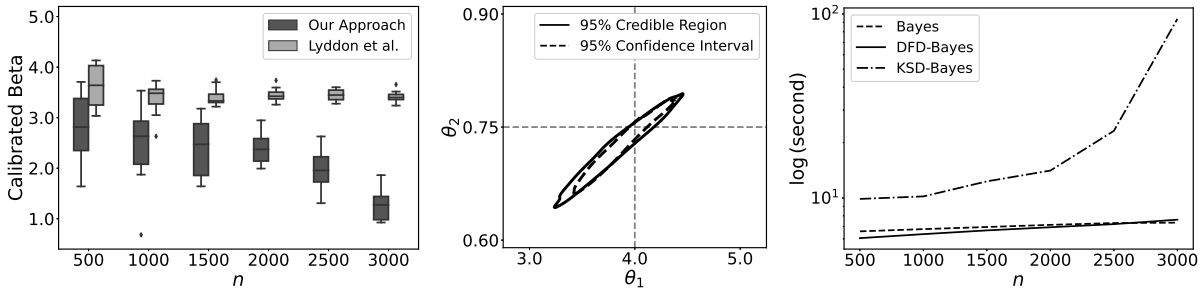

Figure 2: Distribution of β_* across different realisations of the dataset at each data number n for $\theta_2 = 0.75$ (left), comparison of a 95% credible region of the DFD-Bayes posterior and a 95% confidence interval of the frequentist counterpart for $n = 2000$ (centre), and comparison of computational times of each Metropolis–Hastings algorithm (right). The confidence interval was estimated by a 95% highest probability density region of a kernel density estimator constructed at 100 bootstrap minimisers used in our approach to calibrate β.

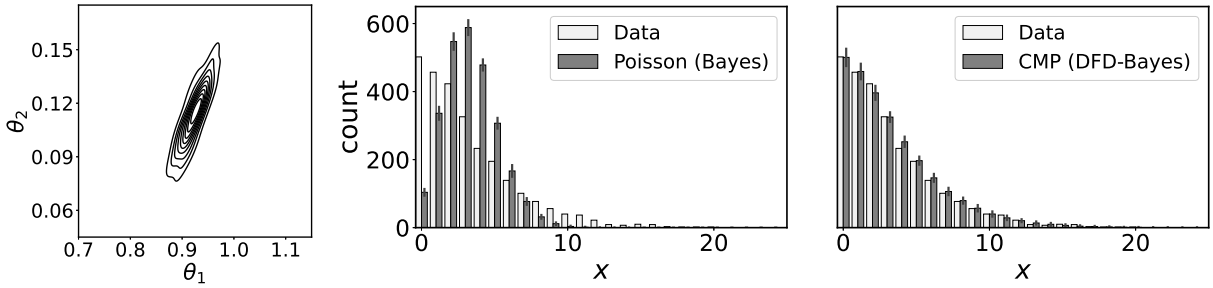

Figure 3: Comparison of DFD-Bayes for the Conway–Maxwell–Poisson model and standard Bayes for the Poisson distribution on the sales data of Shmueli et al. (2005). Left: The generalised posterior distribution produced using DFD-Bayes. Centre: Posterior predictive distribution, at the level of the data, for a Poisson model with standard Bayesian inference performed. Right: Posterior predictive distribution, at the level of the data, for a Conway–Maxwell–Poisson model with DFD-Bayes inference performed. In both cases, error bars indicate one standard deviation of the posterior predictive distribution.

4.2 Ising Model

The aim of this section is to consider a more challenging instance of discrete intractable likelihood, where the data are high-dimensional (i.e. d is large) and the cardinality of each coordinate domain S_i is small. A small cardinality of S_i is particularly interesting, because the intuition that our difference operators arise from discretisation of continuous differential operators fails to hold. This setting is typified by the Ising model (which has $S_i = \{0, 1\}$), variants of which are used to model diverse phenomena, including cellular signaling processes (Jiang et al., 2021), the network structure of the amino-acid sequences (Xue et al., 2012), and animal skin patterns (Zakany et al., 2022). The computational challenge of performing Bayesian inference for Ising-type models has, to-date, principally been addressed using techniques such as pseudo-likelihood (see the recent surveys in Pensar et al., 2017; Bhattacharyya and Atchade, 2019). Unfortunately, these do not necessarily lead to asymptotically exact inference since the correct likelihood is replaced by an approximation.

Let G be an undirected graph on a d-dimensional vertex set and let $\mathcal{N}_i$ denote the neighbours of node i, with self-edges excluded. An Ising model describes a discrete process that assigns each vertex of G either the value 0 or 1, and thus the data domain is $\mathcal{X} = \{0, 1\}^d$. The probability mass function has the exponential

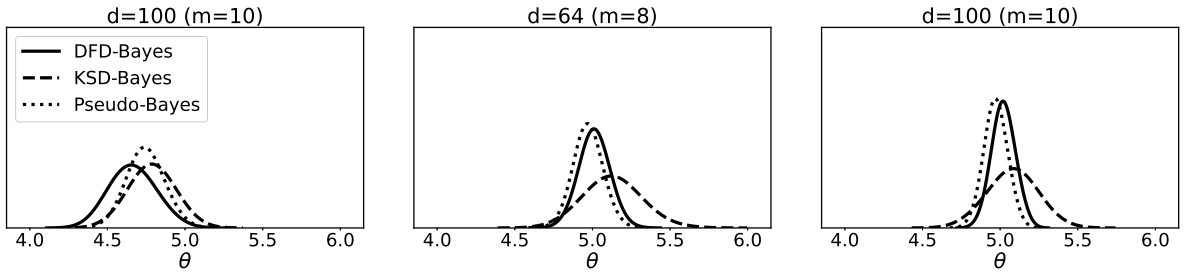

Figure 4: Comparison of approximate Bayesian inference based on pseudo-likelihood, DFD-Bayes and KSD-Bayes, applied to the Ising model with $\theta = 5$ for $n = 1,000$ and $d = 10 \times 10$. For all methods, the value β_* from Section 3.3 was used.

family form

$$p_\theta(\boldsymbol{x}) \propto \exp\left(\frac{1}{\theta}\sum_{i=1}^{d}\sum_{j\in\mathcal{N}_i} x_i x_j\right) \tag{11}$$

where θ is a temperature parameter, controlling the propensity for neighbouring vertices to share a common value. Here we consider the ferromagnetic Ising model, which has $\theta \in (0, \infty)$. To conduct a simulation study we consider the case where G is a $m \times m$ grid. Simulating from Ising models is challenging due to the high-dimensional discrete domain, so here we restrict attention to $m \in \{5, \ldots, 10\}$ to ensure that data were accurately simulated[6]. A total of $n = 1,000$ data points were generated from an Ising model with $\theta = 5$, using an extended run of a Metropolis—Hastings algorithm, the details of which are contained in Appendix E.2.1. A chi-squared prior with degree of freedom 3 was used. Three inference methods were compared: the KSD-Bayes method of Matsubara et al. (2022), the proposed DFD-Bayes method, and standard Bayesian inference based on a the pseudo-likelihood

$$\tilde{p}_\theta(\boldsymbol{x}) = \prod_{i=1}^{d} p_\theta(x_i|\{x_j : j \in \mathcal{N}_i\}),$$

where $p_\theta(x_i|\{x_j : j \in \mathcal{N}_i\})$ is a restriction of the original model (11) to the i-th coordinate x_i under the condition $\{x_j : j \in \mathcal{N}_i\}$ that results in a Bernoulli distribution of x_i for each $i = 1, \cdots, d$ (Besag, 1974). The latter Pseudo-Bayes approach can be viewed as a special case of generalised Bayes inference, since it replaces the original likelihood loss of the model (11) with the pseudo-likelihood loss, and therefore we also applied the proposed calibration procedure to this method. The settings of KSD-Bayes are described in Appendix E.2.2. A Metropolis–Hastings algorithm was also used to sample from all generalised posteriors, the details for which are contained in Appendix E.2.3.

Results are presented for three different datasets of size $n = 1,000$ and dimension $d = 36$ ($m = 6$), $d = 64$ ($m = 8$), and $d = 100$ ($m = 10$) in Figure 4. For the lowest dimension $d = 36$, all the approaches produced similar posteriors. For the higher-dimensional cases, it can be seen that the DFD-Bayes and Pseudo-Bayes posteriors concentrate around the true parameter $\theta = 5$. The KSD-Bayes posterior is more conservative, whilst DFD-Bayes gives a comparable result to Pseudo-Bayes. For $d = 100$, the total computational time required to perform this analysis (including calibration) was 540 seconds for DFD-Bayes, 2,353 seconds for KSD-Bayes, and 1,053 seconds for Pseudo-Bayes each in average over 10 independent experiments, confirming that DFD-Bayes incurs a significantly lower computational cost than both alternatives. The value of the weight obtained through our calibration method for $d = 100$ in Figure 4 was 0.013 for DFD-Bayes, 0.157 for KSD-Bayes, and 0.579 for Pseudo-Bayes. These small values of weight indicated that the calibration worked effectively, preventing the over-concentration of each posterior.

[6]The value of m used in these experiments was not constrained by the computational demand of DFD-Bayes, which scales as $O(m^2)$.

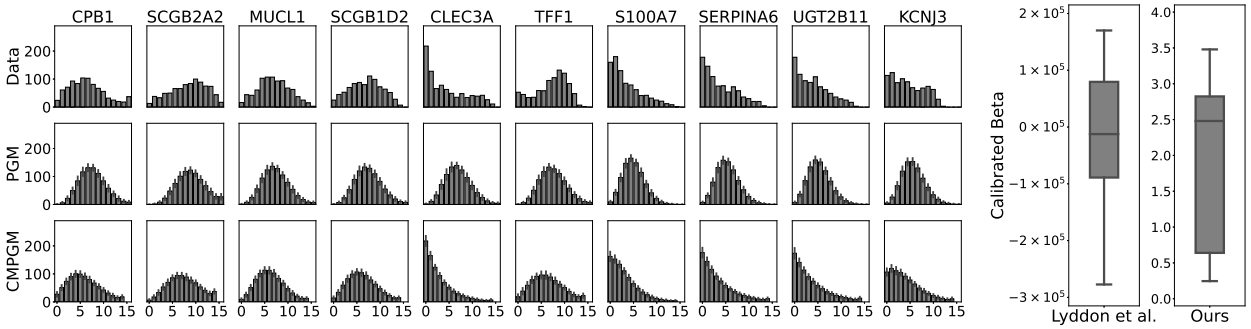

Figure 5: Left: Posterior predictive distributions from the Poisson graphical model and the Conway–Maxwell–Poisson graphical model. Right: Sampling distributions of β_* for the Conway–Maxwell–Poisson graphical model by Lyddon et al. (2019) and by the proposed approach, computed using 10 independent realisations of the dataset.

4.3 Multivariate Count Data

We conclude our numerical experiments with a problem involving multivariate count data. Count data occur in diverse application areas, and variables in such data are rarely independent, yet the literature on statistical modelling of such data is limited. Poisson graphical models have emerged as a powerful tool for modelling such data; see the recent review of Inouye et al. (2017). To the best of our knowledge a Bayesian treatment of Poisson graphical models has yet to be attempted[7], and we speculate that this is due to the computational challenges of the associated intractable likelihood. Our aim here is to assess the suitability of DFD-Bayes for learning the parameters of a Poisson graphical model.

Let G be an undirected graph on a vertex set $\{1, \ldots, d\}$ and let $\mathcal{M}_i$ denote the neighbours of node i that are contained in the set $\{i+1, \ldots, d\}$. A Poisson graphical model has probability mass function

$$p_\theta(\boldsymbol{x}) \propto \exp\left(\sum_{i=1}^{d} \theta_i x_i - \sum_{i=1}^{d} \sum_{j \in \mathcal{M}_i} \theta_{i,j} x_i x_j - \sum_{i=1}^{d} \log(x_i!)\right)$$

where the parameters θ consist of both the linear coefficients $\theta_i \in (-\infty, \infty)$ and the interaction coefficients $\theta_{i,j} \in [0, \infty)$. Our aim is to reproduce an analysis of a breast cancer gene expression dataset described in Inouye et al. (2017), but in a generalised Bayesian framework. For this problem, $n = 878$, $d = 10$, and $p = 64$ which renders the computational cost of $O(n^2 d)$ at every MCMC step and of $O(p^2 n^2 d)$ at calibration associated with KSD-Bayes inefficient. Full details of the dataset are contained in Appendix E.3.1. Independent standard normal priors were employed for each θ_i, and half-normal distributions with scale $(d(d-1)/2)^{-1}$ were employed for each $\theta_{i,j}$. A No-U-Turn Sampler was used to sample from the DFD-Bayes posterior, as described in Appendix E.3.2. The total computational time required to perform this analysis, including calibration, was $1,896$ seconds. Results, in Figure 5, demonstrate that the Poisson graphical model is in fact a poor fit for these data, since the data show signs of being under-dispersed relative to the standard Poisson model. However, in terms of identifying the best parameter values for this model, DFD-Bayes appears to have performed well.

As a possible improvement, and to further stress-test the DFD-Bayes method, we considered fitting a generalisation of the Poisson graphical model that allows for over- and under-dispersion in a manner

[7]Though we note that a pairwise Markov random field whose marginals are close to being Poisson was considered in Roy and Dunson (2020), who employed an approximate likelihood to circumvent the intractable normalising constant.

analogous to Conway and Maxwell (1962). This model takes the form

$$p_\theta(\boldsymbol{x}) \propto \exp\left(\sum_{i=1}^{d}\theta_i x_i - \sum_{i=1}^{d}\sum_{j\in\mathcal{M}_i}\theta_{i,j}x_i x_j - \sum_{i=1}^{d}\theta_{0,i}\log(x_i!)\right)$$

where the additional parameters $\theta_{0,i} \in [0,\infty)$ control the dispersion, with $\theta_{0,i} = 1$ recovering the standard Poisson marginal. This time, $p = 74$ as opposed to $p = 64$ for the Poisson-based model. For this Conway–Maxwell–Poisson graphical model, the same priors as the Poisson graphical model were used for θ_i and $\theta_{i,j}$, and half-normal priors with scale $1/\sqrt{2}$ were used for each $\theta_{0,i}$. Results in Figure 5 demonstrate an improved fit to the dataset. Indeed, the optimal β for the Poisson graphical model was $\beta_* = 0.2150$, which is smaller than the corresponding value $\beta_* = 0.9971$ for the Conway–Maxwell–Poisson graphical model, resulting in a conservative inference outcome when the statistical model is most misspecified and supporting the effectiveness of the proposed approach to calibration.

The right panel of Figure 5 shows the sampling distributions of estimators for the weight β in the context of the Conway–Maxwell–Poisson graphical model, computed using bootstrap resampling of the gene expression dataset. It can be seen that the asymptotic approach proposed in Lyddon et al. (2019) is severely numerically unstable and can even lead to a negative weight, while the approach proposed in Section 3.3 remains stable within a reasonable range between 0 and 3.5. The lack of stability of the approach by Lyddon et al. (2019) arises from the need to invert a covariance matrix of derivatives of the loss, which can easily become numerically singular if the parameter dimension is high. In contrast, note that our approach involves no matrix inversion. This real-data analysis using flexible parametric models highlights the value in being able to perform rapid and automatic (i.e. free from user-specified degrees of freedom) generalised Bayesian inference for discrete intractable likelihood.

5 Conclusion

This paper proposed a novel generalised Bayesian inference procedure for discrete intractable likelihood. The approach, called DFD-Bayes, is distinguished by its lack of user-specified hyperparameters, its suitability for standard Markov chain Monte Carlo algorithms, and its linear (in n, the size of the dataset) computational cost per-iteration of the Markov chain. Furthermore, the generalised posterior is consistent and asymptotically normal. This paper also established a novel approach to calibration of generalised Bayesian posteriors which is computationally efficient (through embarrassing parallelism) and numerically stable, even when the parameter of the statistical model is high-dimensional.

This work focused on independent and identically distributed data, meaning that (for example) regression models were not considered. Relaxing the independence and identical distribution assumptions represents a natural direction for future work, and a road map is provided by recent research in the score-matching literature (Xu et al., 2022).

One of our technical contributions is to present a discrete Fisher divergence applicable to distributions defined on a countably infinite set. This divergence can be regarded as a proper local scoring rule, which complements existing methodology developed in the finite domain context in Dawid et al. (2012). The use of scoring rules as loss functions within a generalised Bayesian framework for continuous data was considered in Giummolè et al. (2019); Pacchiardi and Dutta (2021), and our work can be seen as an analogous approach for discrete data, with particular focus on intractable likelihood.

DFD-Bayes was demonstrated to outperform the comparative approach, KSD-Bayes, in our experiments both in terms of inferential performance and computational cost. However, one of the significant advantages of KSD-Bayes is robustness in the presence of outliers contained in dataset (Matsubara et al., 2022). This is confirmed through an additional experiment on the Ising model in Appendix C.4 Thus, in settings where robust inference is required, the KSD-Bayes approach may be preferred. Future work could however focus on generalising our DFD construction to allow for further robustness as per the diffusion score-matching framework of Barp et al. (2019).

Acknowledgements TM, FXB and CJO were supported by the EPSRC grant EP/N510129/1 and the programme on Data Centric Engineering at the Alan Turing Institute, UK. JK was funded by EPSRC grant EP/L016710/1, the Facebook Fellowship Programme, as well as a Biometrika Fellowship.

References

S. Amari. *Information Geometry and Its Applications*. Springer, 2016.

A. Anastasiou, A. Barp, F.-X. Briol, B. Ebner, R. E. Gaunt, F. Ghaderinezhad, J. Gorham, A. Gretton, C. Ley, Q. Liu, L. Mackey, C. J. Oates, G. Reinert, and Y. Swan. Stein's method meets statistics: A review of some recent developments. *arXiv:2105.03481*, 2021.

C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725, 2009.

A. Barp, F.-X. Briol, A. Duncan, M. Girolami, and L. Mackey. Minimum Stein discrepancy estimators. *Advances in Neural Information Processing Systems*, 32, 2019.

A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer, 2011.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974.

A. Bhattacharyya and Y. Atchade. A Bayesian analysis of large discrete graphical models. *arXiv:1907.01170*, 2019.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 78(5):1103, 2016.

N. Bochkina. Bernstein-von Mises theorem and misspecified models: a review. *arXiv:2204.13614*, 2022.

A. Caimo and N. Friel. Bayesian inference for exponential random graph models. *Social Networks*, 33(1): 41–55, 2011.

L. L. Carter and E. D. Cashwell. *Particle-transport simulation with the Monte Carlo method*. Technical Information Center Energy Research and Development Administration, 1975.

B.-E. Cherief-Abdellatif and P. Alquier. MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. *Proceedings of the 2nd Symposium on Advances in Approximate Bayesian Inference*, pages 1–21, 2020.

F. R. Chung and F. C. Graham. *Spectral graph theory*. American Mathematical Soc., 1997.

R. W. Conway and W. L. Maxwell. A queuing model with state dependent service rates. *Journal of Industrial Engineering*, 12:132–136, 1962.

J. Davidson. *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford University Press, 1994.

A. P. Dawid and M. Musio. Bayesian model selection based on proper scoring rules. *Bayesian Analysis*, 10 (2):479–499, 2015.

A. P. Dawid, S. Lauritzen, and M. Parry. Proper local scoring rules on discrete sample spaces. *The Annals of Statistics*, 40(1):593–608, 2012.

C. Dellaporta, J. Knoblauch, T. Damoulas, and F.-X. Briol. Robust Bayesian inference for simulator-based models via the MMD posterior bootstrap. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 943–970, 2022.

R. Durrett. *Probability: Theory and Examples (4th Edition)*. Cambridge University Press, 2010.

E. M. Elçi, J. Grimm, L. Ding, A. Nasrawi, T. M. Garoni, and Y. Deng. Lifted worm algorithm for the Ising model. *Physical Review E*, 97(4):042126, 2018.

R. G. Everitt. Bayesian parameter estimation for latent markov random fields and social networks. *Journal of Computational and Graphical Statistics*, 21(4):940–960, 2012.

A. Ghosh and A. Basu. Robust Bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, 68:413–437, 2016.

F. Giummolè, V. Mameli, E. Ruli, and L. Ventura. Objective Bayesian inference with proper scoring rules. *Test*, 28(3):728–755, 2019.

G. Hooker and A. N. Vidyashankar. Bayesian model robustness via disparities. *Test*, 23(3):556–584, 2014.

A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

D. I. Inouye, E. Yang, G. I. Allen, and P. Ravikumar. A review of multivariate distributions for count data derived from the Poisson distribution. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(3): e1398, 2017.

J. Jewson and D. Rossell. General Bayesian loss function selection and the use of improper models. *arXiv:2106.01214*, 2021.

J. Jewson, J. Q. Smith, and C. Holmes. Principled Bayesian minimum divergence inference. *Entropy*, 20(6): 442, 2018.

X. Jiang, Q. Li, and G. Xiao. Bayesian modeling of spatial transcriptomics data via a modified Ising model. *arXiv:2104.13957*, 2021.

J. Knoblauch, J. Jewson, and T. Damoulas. An optimization-centric view on bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23(132):1–109, 2022.

F. Liang. A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computation and Simulation*, 80(9):1007–1022, 2010.

B. G. Lindsay, G. Y. Yi, and J. Sun. Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, pages 71–105, 2011.

D. Lusher, J. Koskinen, and G. Robins. *Exponential random graph models for social networks: Theory, methods, and applications*. Cambridge University Press, 2013.

S. P. Lyddon, C. C. Holmes, and S. G. Walker. General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106(2):465–478, 2019.

A.-M. Lyne, M. Girolami, Y. Atchadé, H. Strathmann, and D. Simpson. On Russian roulette estimates for Bayesian inference with doubly-intractable likelihoods. *Statistical Science*, 30(4):443–467, 2015.

S. Lyu. Interpretation and generalization of score matching. *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, page 359–366, 2009.

J.-M. Marin, P. Pudlo, C. P. Robert, and R. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22:1167–1180, 2012.

T. Matsubara, J. Knoblauch, F.-X. Briol, and C. J. Oates. Robust generalised Bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society B: Statistical Methodology*, 2022. To appear.

J. W. Miller. Asymptotic normality, concentration, and coverage of generalized posteriors. *Journal of Machine Learning Research*, 22(168):1–53, 2021.

M. Moores, G. Nicholls, A. Pettitt, and K. Mengersen. Scalable Bayesian inference for the inverse temperature of a hidden Potts model. *Bayesian Analysis*, 15(1):1–27, 2020.

I. Murray, Z. Ghahramani, and D. J. C. MacKay. MCMC for doubly-intractable distributions. *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence*, pages 359–366, 2006.

J. Møller, A. N. Pettitt, R. Reeves, and K. K. Berthelsen. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, 2006.

L. Pacchiardi and R. Dutta. Generalized Bayesian likelihood-free inference using scoring rules estimators. *arXiv:2104.03889*, 2021.

J. Park and M. Haran. Bayesian inference in the presence of intractable normalizing functions. *Journal of the American Statistical Association*, 113(523):1372–1390, 2018.

J. Pensar, H. Nyman, J. Niiranen, and J. Corander. Marginal pseudo-likelihood learning of discrete markov network structures. *Bayesian Analysis*, 12(4):1195–1215, 2017.

L. F. Price, C. C. Drovandi, A. Lee, and D. J. Nott. Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics*, 27(1):1–11, 2018.

J. Propp and D. Wilson. Coupling from the past: a user's guide. *Microsurveys in Discrete Probability*, 41: 181–192, 1998.

A. Roy and D. B. Dunson. Nonparametric graphical model for counts. *Journal of Machine Learning Research*, 21(229):1–21, 2020.

S. Shao, P. E. Jacob, J. Ding, and V. Tarokh. Bayesian model comparison with the Hyvarinen score: computation and consistency. *Journal of the American Statistical Association*, 114(528):1826–1837, 2019.

J. Shi, Y. Zhou, J. Hwang, M. K. Titsias, and L. Mackey. Gradient estimation with discrete Stein operators. *arXiv: 2202.09497*, 2022.

G. Shmueli, T. P. Minka, J. B. Kadane, S. Borle, and P. Boatwright. A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):127–142, 2005.

N. Syring and R. Martin. Calibrating general posterior credible regions. *Biometrika*, 106(2):479–486, 2019.

T. Takenouchi and T. Kanamori. Statistical inference with unnormalized discrete models and localized homogeneous divergences. *Journal of Machine Learning Research*, 18(1):1804–1829, 2017.

A. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*, pages 5–42, 2011.

Y.-W. Wan, G. I. Allen, and Z. Liu. TCGA2STAT: simple TCGA data access for integrated statistical analysis in R. *Bioinformatics*, 32(6):952–954, 11 2015.

C. Wei and I. Murray. Markov Chain Truncation for Doubly-Intractable Inference. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 54:776–784, 2017.

L. K. Wenliang and H. Kanagawa. Blindness of score-based methods to isolated components and mixing proportions. In *NeurIPS Workshop "Your Model is Wrong: Robustness and misspecification in probabilistic modeling"*, 2021.

J. Xu, J. L. Scealy, A. T. A. Wood, and T. Zou. Generalized score matching for regression. *arXiv:2203.09864*, 2022.

L. Xue, H. Zou, and T. Cai. Nonconcave penalized composite conditional likelihood estimation of sparse ising models. *The Annals of Statistics*, 40(3):1403–1429, 2012.

J. Yang, Q. Liu, V. Rao, and J. Neville. Goodness-of-fit testing for discrete distributions via Stein discrepancy. *Proceedings of the 35th International Conference on Machine Learning*, pages 5561–5570, 2018.

P. Yatsyshin, S. Kalliadasis, and A. B. Duncan. Physics-constrained Bayesian inference of state functions in classical density-functional theory. *The Journal of Chemical Physics*, 156(7):074105, 2022.

M. Yu, M. Kolar, and V. Gupta. Statistical inference for pairwise graphical models using score matching. *Advances in Neural Information Processing Systems*, 29, 2016.

S. Yu, M. Drton, and A. Shojaie. Generalized score matching for non-negative data. *Journal of Machine Learning Research*, 20:1–70, 2019.

S. Zakany, S. Smirnov, and M. C. Milinkovitch. Lizard skin patterns and the ising model. *Physical Review Letters*, 128(4):048102, 2022.

Appendices

These appendices are structured as follows: The proofs for all theoretical results are contained in Appendix A, with the proof of an auxiliary result reserved for Appendix B. The relationship between discrete Fisher divergence and Stein discrepancies is explored in Appendix C. Detailed calculations for worked examples are provided in Appendix D. Full details on our numerical experiments are provided in Appendix E

A Proofs of Theoretical Results

This first appendix contains the proof of all theoretical results in the paper, including Proposition 1, Theorem 1 and Theorem 2.

A.1 Proof of Proposition 1

First we introduce three technical lemmas that will be useful:

Lemma 1. *For any $\boldsymbol{x} \in \mathcal{X}$ and $i = 1, \ldots, d$, it holds that $(\boldsymbol{x}^{i-})^{i+} = \boldsymbol{x}$ and $(\boldsymbol{x}^{i+})^{i-} = \boldsymbol{x}$.*

Proof. Since $\mathcal{X} = S_1 \times \cdots \times S_d$ from the Standing Assumption,

$$\boldsymbol{x}^{i-} = (x_1, \cdots, x_i^-, \cdots, x_d), \qquad \boldsymbol{x}^{i+} = (x_1, \cdots, x_i^+, \cdots, x_d). \tag{12}$$

It is thus sufficient to show that $(x_i^-)^+ = x_i$ and $(x_i^+)^- = x_i$ for any $i = 1, \ldots, d$. Consider, therefore, a countable ordered set S with more than one element. Our aim is to establish the identity $(s^-)^+ = s$ and $(s^+)^- = s$ for all $s \in S$. Existence of the least and greatest element $s_{\min}$ and $s_{\max}$ of S determines four qualitatively distinct cases to be checked: (i) neither of them exist; (ii) both of them exists; (iii) only $s_{\min}$ exists; (iv) only $s_{\max}$ exists. Recall that we identify the case (iv) with (iii) without loss of generality by reversing the ordering of S. The identity for (i) & (ii) is trivial since the maps $s \mapsto s^-$ is bijective from S to itself with inverse $s \mapsto s^+$. For case (iii), we have $(s^-)^+ = s$ for $s \neq s_{\min}$ and $(s^+)^- = s$ for all $s \in S$, Recalling the definition $s_{\min}^- = \star$ and $\star^+ = s_{\min}$ completes the argument. □

Lemma 2. *For any $f, g : \mathcal{X} \to \mathbb{R}$ and any $i = 1, \ldots, d$, suppose $\sum_{\boldsymbol{x} \in \mathcal{X}} |f(\boldsymbol{x}) g(\boldsymbol{x}^{i-})| < \infty$, that is, the series is absolutely convergent. Then we have*

$$\sum_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}) g(\boldsymbol{x}^{i-}) = \sum_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}^{i+}) g(\boldsymbol{x}). \tag{13}$$

Proof. Since $\mathcal{X} = S_1 \times \cdots \times S_d$ from the Standing Assumption, the series can be expressed as

$$\sum_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}) g(\boldsymbol{x}^{i-}) = \sum_{x_1 \in S_1} \cdots \sum_{x_i \in S_i} \cdots \sum_{x_d \in S_d} f(x_1, \cdots, x_i, \cdots, x_d) g(x_1, \cdots, x_i^-, \cdots, x_d),$$
$$\sum_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}^{i+}) g(\boldsymbol{x}) = \sum_{x_1 \in S_1} \cdots \sum_{x_i \in S_i} \cdots \sum_{x_d \in S_d} f(x_1, \cdots, x_i^+, \cdots, x_d) g(x_1, \cdots, x_i, \cdots, x_d).$$

Holding the coordinates $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d$ fixed, and exploiting absolute convergence to justify the interchange of summations, the claimed result follows if

$$\sum_{x_i \in S_i} \tilde{f}(x_i) \tilde{g}(x_i^-) = \sum_{x_i \in S_i} \tilde{f}(x_i^+) \tilde{g}(x_i) \tag{14}$$

where $\tilde{f}(x_i) := f(x_1, \cdots, x_i, \cdots, x_d)$ and $\tilde{g}(x_i) := g(x_1, \cdots, x_i, \cdots, x_d)$ are viewed as functions on S_i.

Consider, therefore, an arbitrary countable ordered set S, for which we aim to establish the identity $\sum_{s \in S} h(s) k(s^-) = \sum_{s \in S} h(s^+) k(s)$ for any functions $h, k : S \to \mathbb{R}$ s.t. $\sum_{s \in S} |h(s) k(s^-)| < \infty$. First, notice that any countable set can be indexed using a consecutive subset $I \subseteq \mathbb{Z}$, such that $s_i < s_j$ if and only if $i < j$. The identity therefore can be written as $\sum_{i \in I} h(s_i) k(s_i^-) = \sum_{i \in I} h(s_i^+) k(s_i)$, and will be verified for the three qualitatively distinct cases of index set I described in the proof of Lemma 1:

(i) $I = \mathbb{Z}$. The result is immediate, since $(s_i, s_i^-) = (s_i, s_{i-1})$ and $(s_i^+, s_i) = (s_{i+1}, s_i)$ range over the same set for $i \in I$. The series $\sum_{i\in I} h(s_i^+)k(s_i)$ is absolutely convergent since the sets $\{h(s_i)k(s_i^-)\}_{i\in I}$ and $\{h(s_i^+)k(s_i)\}_{i\in I}$ in the two series are equal.

(ii) $I = \{1, \dots, n\}$ for some $n \in \mathbb{N}$. In this case $s_{\min} = s_1$ and $s_{\max} = s_n$, and it follows from the definition of decrements and increments that

$$\begin{aligned}\sum_{i\in I} h(s_i)k(s_i^-) &= h(s_1)k(s_1^-) + h(s_2)k(s_1) + \cdots + h(s_n)k(s_{n-1}) \\ &= h(s_n^+)k(s_n) + h(s_2)k(s_1) + \cdots + h(s_n)k(s_{n-1}) = \sum_{i\in I} h(s_i^+)k(s_i),\end{aligned}$$

where the sets $\{h(s_i)k(s_i^-)\}_{i\in I}$ and $\{h(s_i^+)k(s_i)\}_{i\in I}$ are again equal.

(iii) $I = \{1, 2, \dots\}$. In this case $s_{\min} = s_1$, and it follows from the definition $s_1^- = \star$ and $k(\star) = 0$ that

$$\begin{aligned}\sum_{i\in I} h(s_i)k(s_i^-) &= \underbrace{h(s_1)k(\star)}_{=0} + h(s_2)k(s_1) + h(s_3)k(s_2) + \cdots \\ &= h(s_2)k(s_1) + h(s_3)k(s_2) + \cdots = \sum_{i\in I} h(s_i^+)k(s_i).\end{aligned}$$

The series $\sum_{i\in I} h(s_i^+)k(s_i)$ is absolutely convergent since the set $\{h(s_i^+)k(s_i)\}_{i\in I}$ is a subset of the absolutely summable set $\{h(s_i)k(s_i^-)\}_{i\in I}$.

This completes the proof. □

Let $F(\mathcal{X}, S)$ denote the set of all functions f of the form $f : \mathcal{X} \to S$.

Lemma 3. *For $p : \mathcal{X} \to (0, \infty)$, the map $\mu_p := (\nabla^- p)/p$ is an injection $\mu : F(\mathcal{X}, (0, \infty)) \to F(\mathcal{X}, \mathbb{R}^d)$.*

Proof. It suffices to show that each value $p(\boldsymbol{x})$, for $\boldsymbol{x} \in \mathcal{X}$, can be explicitly recovered from μ_p. Note that, since p takes values in $(0, \infty)$, the embedding μ_p is well-defined. From the Standing Assumption, we have that $\mathcal{X} = S_1 \times \cdots \times S_d$, where each S_i is a countable ordered set with more than one element. Since the S_i serve only as index sets, we can without loss of generality assume that S_i is a consecutive subset of $\mathbb{Z}$ and that $0 \in S_i$, for each $i = 1, \dots, d$. The idea of the proof is to demonstrate that each of the quantities $p(\boldsymbol{x})$ can be explicitly expressed in terms of μ_p, $p(\mathbf{0})$ and $\{p(\boldsymbol{y}) : \|\boldsymbol{y}\|_1 < \|\boldsymbol{x}\|_1\}$, where $\|\boldsymbol{x}\|_1 := |x_1| + \cdots + |x_d|$. It would then follow from a simple inductive argument that $p(\boldsymbol{x})$ can be expressed in terms of μ_p and $p(\mathbf{0})$. Finally, the constraint that $\sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x}) = 1$ uniquely determines $p(\mathbf{0})$, demonstrating that $p(\boldsymbol{x})$ can be explicitly recovered.

Given $\boldsymbol{x} \in \mathcal{X}$, assume $\boldsymbol{x} \neq \mathbf{0}$, for otherwise the claim will trivially hold. Then let $i \in \{1, \dots, d\}$ be such that $x_i \neq 0$. If $x_i > 0$, then from the definition of $\mu_p(\boldsymbol{x})_i = 1 - p(\boldsymbol{x}^{i-})/p(\boldsymbol{x})$ we have the relation

$$p(\boldsymbol{x}) = \frac{p(\boldsymbol{x}^{i-})}{1 - \mu_p(\boldsymbol{x})_i}$$

where $\|\boldsymbol{x}^{i-}\|_1 = \|\boldsymbol{x}\|_1 - 1$. Conversely, if $x_i < 0$, then using Lemma 1 we have $\mu_p(\boldsymbol{x}^{i+})_i = 1 - p(\boldsymbol{x})/p(\boldsymbol{x}^{i+})$ and we have the relation

$$p(\boldsymbol{x}) = [1 - \mu_p(\boldsymbol{x}^{i+})_i]p(\boldsymbol{x}^{i+})$$

where $\|\boldsymbol{x}^{i+}\|_1 = \|\boldsymbol{x}\|_1 - 1$. The previously described inductive argument completes the proof. □

Now we prove the main result:

Proof of Proposition 1. The assumption that p and q are positive with $(\nabla^- p)/p, (\nabla^- q)/q \in L^2(q, \mathbb{R}^d)$ ensures that $\mathrm{DFD}(p\|q)$ is well-defined, immediately from (3). Expanding the square gives that

$$\mathrm{DFD}(p\|q) = \mathbb{E}_{X\sim q}\left[\left\|\frac{\nabla^- p(X)}{p(X)}\right\|^2 - 2\underbrace{\frac{\nabla^- p(X)}{p(X)}\cdot\frac{\nabla^- q(X)}{q(X)}}_{=:(*)} + \left\|\frac{\nabla^- q(X)}{q(X)}\right\|^2\right].$$

First we will show that $\mathbb{E}_{X\sim q}[(*)] = -\mathbb{E}_{X\sim q}[\nabla^+ \cdot (\nabla^- p(X)/p(X))]$. By definition of ∇^-,

$$\begin{aligned}\mathbb{E}_{X\sim q}[(*)] &= \sum_{j=1}^d \mathbb{E}_{X\sim q}\left[\frac{p(X) - p(X^{j-})}{p(X)} - \frac{p(X) - p(X^{j-})}{p(X)}\frac{q(X^{j-})}{q(X)}\right] \\ &= \sum_{j=1}^d \Bigg\{ \sum_{\boldsymbol{x}\in\mathcal{X}} \frac{p(\boldsymbol{x}) - p(\boldsymbol{x}^{j-})}{p(\boldsymbol{x})} q(\boldsymbol{x}) - \underbrace{\sum_{\boldsymbol{x}\in\mathcal{X}} \frac{p(\boldsymbol{x}) - p(\boldsymbol{x}^{j-})}{p(\boldsymbol{x})} q(\boldsymbol{x}^{j-})}_{(**)} \Bigg\}.\end{aligned}$$

Application of Lemma 2 to the term $(**)$, with $f(\boldsymbol{x}) = (p(\boldsymbol{x}) - p(\boldsymbol{x}^{j-}))/p(\boldsymbol{x})$ and $g(\boldsymbol{x}) = q(\boldsymbol{x})$, and using Lemma 1 to deduce that $(\boldsymbol{x}^{j-})^{j+} = \boldsymbol{x}$, reveals that

$$(**) = \sum_{\boldsymbol{x}\in\mathcal{X}} \frac{p(\boldsymbol{x}^{j+}) - p(\boldsymbol{x})}{p(\boldsymbol{x}^{j+})} q(\boldsymbol{x})$$

for each $j = 1, \ldots, d$. Hence

$$\mathbb{E}_{X\sim q}[(*)] = \mathbb{E}_{X\sim q}\left[\sum_{i=1}^d \frac{p(X) - p(X^{j-})}{p(X)} - \frac{p(X^{j+}) - p(X)}{p(X^{j+})}\right] = -\mathbb{E}_{X\sim q}\left[\nabla^+ \cdot \frac{\nabla^- p(X)}{p(X)}\right],$$

where we have used the definition of $(\nabla^+ \cdot)$ and, again, the fact that $(\boldsymbol{x}^{j-})^{j+} = \boldsymbol{x}$ from Lemma 1. Thus we have established that

$$\mathrm{DFD}(p\|q) = \mathbb{E}_{X\sim q}\left[\left\|\frac{\nabla^- p(X)}{p(X)}\right\|^2 + 2\nabla^+ \cdot \frac{\nabla^- p(X)}{p(X)} + \left\|\frac{\nabla^- q(X)}{q(X)}\right\|^2\right].$$

Finally we verify that $\mathrm{DFD}(p\|q) = 0$ if and only if $p = q$. From Lemma 3 we have the injective embedding $p \mapsto \mu_p := (\nabla^- p)/p$ of a positive density $p : \mathcal{X} \to (0, \infty)$ into $F(\mathcal{X}, \mathbb{R}^d)$. Since $q > 0$, the map $p \mapsto \mu_p$ is also an injection into $L^2(q, \mathbb{R}^d)$, equipped with the canonical norm $\|\nu\|_{L^2(q,\mathbb{R}^d)} := \mathbb{E}_{X\sim q}[\|\nu(X)\|^2]$, $\forall \nu \in L^2(q, \mathbb{R}^d)$. From (3) we recognise that $\mathrm{DFD}(p\|q) = \|\mu_p - \mu_q\|^2_{L^2(q,\mathbb{R}^d)}$ is the squared distance between μ_p and μ_q according to the canonical norm of $L^2(q, \mathbb{R}^d)$. Since $\|\mu_p - \mu_q\|_{L^2(q,\mathbb{R}^d)} = 0$ if and only if $\mu_p = \mu_q$ in $L^2(q, \mathbb{R}^d)$, it follows from injectivity of $p \mapsto \mu_p$ that $\mathrm{DFD}(p\|q) = 0$ if and only if $p = q$, as required. □

A.2 Proof of Theorem 1

This appendix contains the proof of Theorem 1. Miller (2021) provided sufficient conditions for consistency and asymptotic normality of generalised Bayesian posteriors of the form $\pi_n^D(\mathrm{d}\theta) \propto \exp(-nD_n(\theta))\pi(\mathrm{d}\theta)$, where $D_n : \Theta \to \mathbb{R}$ is a loss function that may depend on the data $\{\boldsymbol{x}_i\}_{i=1}^n$. These results can be leveraged to analyse DFD-Bayes, by setting

$$D_n(\theta) \stackrel{\theta}{=} \frac{\beta}{n}\sum_{i=1}^n \left\|\frac{\nabla^- p_\theta(\boldsymbol{x}_i)}{p_\theta(\boldsymbol{x}_i)}\right\|^2 + 2\nabla^+ \cdot \left(\frac{\nabla^- p_\theta(\boldsymbol{x}_i)}{p_\theta(\boldsymbol{x}_i)}\right). \tag{15}$$

These conditions were refined into more applicable forms in Matsubara et al. (2022). While Matsubara et al. (2022) focused on their particular case of losses based on kernelised Stein discrepancies, their argument can be directly applied for essentially any arbitrary loss D_n. We repeat this argument by modifying it so that it can be applied for any loss D_n. Let $B_\epsilon(\theta_*) := \{\theta \in \Theta : \|\theta - \theta_*\| < \epsilon\}$.

Theorem 3. *Let $\Theta \subseteq \mathbb{R}^p$ be Borel. Let $D : \Theta \to \mathbb{R}$ be a fixed measurable function and $\{D_n\}_{n=1}^{\infty}$ be a sequence s.t. $D_n : \Theta \to \mathbb{R}$ is a measurable function dependent of random data $\{\boldsymbol{x}_i\}_{i=1}^n \subset \mathcal{X}$. Let $H_n(\theta) := \nabla_\theta^2 D_n(\theta)$. Suppose that, for some bounded convex open set $U \subseteq \Theta$, the following hold:*

C1 D_n a.s. converges pointwise to D;

C2 D_n is r times continuously differentiable in U and $\limsup_{n\to\infty} \sup_{\theta\in U} \|\nabla_\theta^r D_n(\theta)\| < \infty$ a.s. for $r = 1, 2, 3$;

C3 for all n sufficiently large, $\theta_n \in U$ for any $\theta_n \in \arg\min D_n$ a.s., and a point $\theta_* \in U$ uniquely attains $D(\theta_*) = \inf_{\theta\in\Theta} D(\theta)$.

C4 $H_n(\theta_*) \xrightarrow{a.s.} H_*$ for some nonsingular H_*;

C5 π is continuous and positive at θ_*.

Then, for any $\epsilon > 0$, the generalised posterior $\pi_n^D(\mathrm{d}\theta) \propto \exp(-nD_n(\theta))\pi(\mathrm{d}\theta)$ satisfies

$$\int_{B_\epsilon(\theta_*)} \pi_n^D(\theta)\, \mathrm{d}\theta \xrightarrow{\text{a.s.}} 1.$$

Let $(\theta_n)_{n=1}^{\infty} \subset \Theta$ be a sequence s.t. θ_n minimises D_n for all n sufficiently large. Denote by $\widetilde{\pi}_n^D$ a density on $\mathbb{R}^d$ of the random variable $\sqrt{n}(\theta - \theta_n)$, where $\theta \sim \pi_n^D$. Then

$$\int_{\mathbb{R}^d} \left| \widetilde{\pi}_n^D(\theta) - \frac{1}{Z_*} \exp\left(-\frac{1}{2}\theta \cdot H_*\theta\right) \right| \mathrm{d}\theta \xrightarrow{\text{a.s.}} 0$$

where Z_* is the normalising constant of $\exp(-\frac{1}{2}\theta \cdot H_*\theta)$.

The proof of Theorem 3 is deferred to Appendix B. The main proof of Theorem 1 aims to show that the preconditions C1-C5 of Theorem 3 are satisfied for the particular function D_n in (15), defining the DFD-Bayes generalised posterior.

Proof of Theorem 1. Without loss of generality, we will give the proof for $\beta = 1$ for notational convenience.[8] Let $r_{j-}(\boldsymbol{x}, \theta) := p_\theta(\boldsymbol{x}^{j-})/p_\theta(\boldsymbol{x})$ and $r_{j+}(\boldsymbol{x}, \theta) := p_\theta(\boldsymbol{x})/p_\theta(\boldsymbol{x}^{j+})$ for each $j = 1, \ldots, d$. By the definition of ∇^- and ∇^+, we can write D_n as

$$D_n(\theta) \overset{\theta}{=} \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^d 1 - 2\frac{p_\theta(\boldsymbol{x}_i)}{p_\theta(\boldsymbol{x}_i^{j+})} + \left(\frac{p_\theta(\boldsymbol{x}_i^{j-})}{p_\theta(\boldsymbol{x}_i)}\right)^2 = \frac{1}{n}\sum_{i=1}^n \underbrace{\sum_{j=1}^d 1 - 2r_{j+}(\boldsymbol{x}_i, \theta) + \big(r_{j-}(\boldsymbol{x}_i, \theta)\big)^2}_{=:R(\boldsymbol{x}_i,\theta)}.$$

It is easy to verify that $R(\boldsymbol{x}, \theta) = \|\nabla^- p_\theta(\boldsymbol{x})/p_\theta(\boldsymbol{x})\|^2 + 2\nabla^+ \cdot (\nabla^- p_\theta(\boldsymbol{x})/p_\theta(\boldsymbol{x}))$ from the definition. In what follows we set $D(\theta) := \mathbb{E}_{X\sim p}[R(X, \theta)]$ and verify that preconditions C1-C5 of Theorem 1 are satisfied. Note that C3 holds directly by Assumption 1 and C5 is also assumed directly in Theorem 1.

C1: By the strong law of large number (Durrett, 2010, Theorem 2.5.10),

$$D_n(\theta) = \frac{1}{n}\sum_{i=1}^n R(\boldsymbol{x}_i, \theta) \quad \xrightarrow{\text{a.s.}} \quad \mathbb{E}_{X\sim p}[R(X, \theta)] = D(\theta), \tag{16}$$

[8] To extend the proof to arbitrary $\beta > 0$, simply replace $D_n(\theta) = \mathrm{DFD}(p_\theta \| p_n)$ in all arguments by $D_n(\theta) = \beta\, \mathrm{DFD}(p_\theta \| p_n)$. All the arguments hold immediately since β is a constant.

provided that $\mathbb{E}_{X\sim p}[|R(X,\theta)|] < \infty$ for each $\theta \in \Theta$. Thus we must check that $\mathbb{E}_{X\sim p}[|R(X,\theta)|] < \infty$. By the triangle inequality,

$$\begin{aligned}
\mathbb{E}_{X\sim p}[|R(X,\theta)|] &= \mathbb{E}_{X\sim p}\left[\left|R(X,\theta) + \left\|\frac{\nabla^- p(X)}{p(X)}\right\|^2 - \left\|\frac{\nabla^- p(X)}{p(X)}\right\|^2\right|\right] \\
&\leq \mathbb{E}_{X\sim p}\left[\left|R(X,\theta) + \left\|\frac{\nabla^- p(X)}{p(X)}\right\|^2\right|\right] + \mathbb{E}_{X\sim p}\left[\left\|\frac{\nabla^- p(X)}{p(X)}\right\|^2\right] \\
&\leq \mathbb{E}_{X\sim p}\left[\left|\left\|\frac{\nabla^- p_\theta(X)}{p_\theta(X)}\right\|^2 + 2\nabla^+ \cdot \frac{\nabla^- p_\theta(X)}{p_\theta(X)} + \left\|\frac{\nabla^- p(X)}{p(X)}\right\|^2\right|\right] + \mathbb{E}_{X\sim p}\left[\left\|\frac{\nabla^- p(X)}{p(X)}\right\|^2\right] \\
&= \mathbb{E}_{X\sim p}\left[\left\|\frac{\nabla^- p_\theta(X)}{p_\theta(X)} - \frac{\nabla^- p(X)}{p(X)}\right\|^2\right] + \mathbb{E}_{X\sim p}\left[\left\|\frac{\nabla^- p(X)}{p(X)}\right\|^2\right]
\end{aligned}$$

where the last equality holds from Proposition 1 and both the quantities are finite by Standing Assumption 1. Hence (16) holds for every $\theta \in \Theta$.

C2: From Assumption 2, we have that $r_{j+}(\boldsymbol{x},\theta)$ and $r_{j-}(\boldsymbol{x},\theta)$ are three times continuously differentiable with respect to $\theta \in U$ for all $\boldsymbol{x} \in \mathcal{X}$, and thus $D_n(\theta)$ is three times continuously differentiable with respect to $\theta \in U$. For any $s \in \{1,2,3\}$,

$$\nabla_\theta^s D_n(\theta) = \frac{1}{n}\sum_{i=1}^n \nabla_\theta^s R(\boldsymbol{x}_i,\theta) = \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^d \nabla_\theta^s (r_{j-}(\boldsymbol{x}_i,\theta)^2) - 2\nabla_\theta^s r_{j+}(\boldsymbol{x}_i,\theta). \tag{17}$$

By the triangle inequality, we have an upper bound

$$\sup_{\theta\in U}\|\nabla_\theta^s D_n(\theta)\| \leq \frac{1}{n}\sum_{i=1}^n \underbrace{\sum_{j=1}^d \sup_{\theta\in U}\left\|\nabla_\theta^s (r_{j-}(\boldsymbol{x}_i,\theta)^2)\right\| + 2\sup_{\theta\in U}\|\nabla_\theta^s r_{j+}(\boldsymbol{x}_i,\theta)\|}_{=:G(\boldsymbol{x}_i)}.$$

The quantity $\frac{1}{n}\sum_{i=1}^n G(\boldsymbol{x}_i)$ is a random variable dependent on $\{\boldsymbol{x}_i\}_{i=1}^n$. By the strong law of large numbers (Durrett, 2010, Theorem 2.5.10),

$$\frac{1}{n}\sum_{i=1}^n G(\boldsymbol{x}_i) \xrightarrow{\text{a.s.}} \mathbb{E}_{X\sim p}[G(X)] < \infty$$

provided that $\mathbb{E}_{X\sim p}[|G(X)|] < \infty$. Indeed, this condition holds since from positivity of G

$$\mathbb{E}_{X\sim p}[|G(X)|] = \sum_{j=1}^d \mathbb{E}_{X\sim p}\left[\sup_{\theta\in U}\left\|\nabla_\theta^s (r_{j-}(X,\theta)^2)\right\|\right] + 2\mathbb{E}_{X\sim p}\left[\sup_{\theta\in U}\|\nabla_\theta^s r_{j+}(X,\theta)\|\right],$$

where the right hand side is finite by Assumption 2. Then

$$\limsup_{n\to\infty}\sup_{\theta\in U}\|\nabla_\theta^s D_n(\theta)\| \leq \limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^n G(\boldsymbol{x}_i) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n G(\boldsymbol{x}_i) \overset{\text{a.s.}}{=} \mathbb{E}_{X\sim p}[G(X)] < \infty$$

for any $s \in \{1,2,3\}$, which establishes C2.

C4: Let $h(\boldsymbol{x},\theta) := \nabla_\theta^2 R(\boldsymbol{x},\theta)$. From (17), $H_n(\theta) = \frac{1}{n}\sum_{i=1}^n h(\boldsymbol{x}_i,\theta)$. By the strong law of large number (Durrett, 2010, Theorem 2.5.10), we have $H_n(\theta) \xrightarrow{\text{a.s.}} \mathbb{E}_{X\sim p}[h(X,\theta)]$ provided that $\mathbb{E}_{X\sim p}[\|h(X,\theta)\|] < \infty$. Indeed, this condition holds for all $\theta \in U$, since we have the upper bound

$$\mathbb{E}_{X\sim p}[\|h(X,\theta_*)\|] \leq \mathbb{E}_{X\sim p}\left[\sup_{\theta\in U}\|h(X,\theta)\|\right] \leq \mathbb{E}_{X\sim p}\left[|G(X)|\right] < \infty$$

where the right hand side is bounded by the preceding argument. It remains to verify that $H_* := \lim_{n\to\infty} H_n(\theta_*)$ is equal to $\nabla^2_\theta \,\mathrm{DFD}(p_\theta\|p)|_{\theta=\theta_*}$, from which C4 follows since H_* was assumed to be nonsingular in the statement of Theorem 1. By the Lebesgue's dominated convergence theorem, for each $\theta \in U$,

$$\lim_{n\to\infty} H_n(\theta) = \mathbb{E}_{X\sim p}[\nabla^2_\theta R(\boldsymbol{x},\theta)] = \nabla^2_\theta \mathbb{E}_{X\sim p}[R(\boldsymbol{x},\theta)] = \nabla^2_\theta D(\theta).$$

provided that $\mathbb{E}_{X\sim p}[\sup_{\theta\in U} \|\nabla^2_\theta R(\boldsymbol{x},\theta)\|] < \infty$. This condition holds for all $\theta \in U$ since $\mathbb{E}_{X\sim p}[\sup_{\theta\in U} \|\nabla^2_\theta R(\boldsymbol{x},\theta)\|] \le \mathbb{E}_{X\sim p}[|G(X)|] < \infty$. Since $\theta_* \in U$ in particular, $H_* = \nabla^2_\theta D(\theta)\,|_{\theta=\theta_*} = \nabla^2_\theta \,\mathrm{DFD}(p_\theta\|p)|_{\theta=\theta_*}$, as claimed.

Thus preconditions C1-C5 are satisfied and the result follows from Theorem 3. □

A.3 Proof of Theorem 2

Proof. We first calculate the Fisher divergence between the generalised posterior π^D_n and an empirical distribution δ^B_θ of the bootstrap minimisers $\{\theta^{(b)}_n\}_{b=1}^B$, and then minimise it as a function of the weighting constant β. Recall that the score-matching divergence (Hyvärinen, 2005) is given by

$$\mathrm{D}(\pi^D_n\|\delta^B_\theta) = \frac{1}{B}\sum_{b=1}^B \underbrace{\left\|\nabla_\theta \log \pi^D_n(\theta^{(b)}_n)\right\|^2}_{=(*_1)} + \underbrace{2\,\mathrm{Tr}\left(\nabla^2_\theta \log \pi^D_n(\theta^{(b)}_n)\right)}_{=(*_2)}.$$

The score function of π^D_n is given by

$$\nabla_\theta \log \pi^D_n(\theta) = -\beta \nabla_\theta D_n(\theta) + \nabla_\theta \log \pi(\theta),$$

which is independent of the normalising constant of π^D_n. Similarly, the second derivative is $\nabla^2_\theta \log \pi^D_n(\theta) = -\beta\nabla^2_\theta D_n(\theta) + \nabla^2_\theta \log \pi(\theta)$. Therefore the terms $(*_1)$ and $(*_2)$ in the Fisher divergence can be written as

$$\begin{aligned}(*_1) &= \beta^2 \|\nabla_\theta D_n(\theta^{(b)}_n)\|^2 - 2\beta \nabla_\theta D_n(\theta^{(b)}_n)\cdot \nabla_\theta \log \pi(\theta^{(b)}_n) + \|\nabla_\theta \log \pi(\theta^{(b)}_n)\|^2 \\ (*_2) &= -\beta\,\mathrm{Tr}\left(\nabla^2_\theta D_n(\theta^{(b)}_n)\right) + \mathrm{Tr}\left(\nabla^2_\theta \log \pi(\theta^{(b)}_n)\right)\end{aligned}$$

Now consider minimising the Fisher divergence $\mathrm{D}(\pi^D_n\|\delta^B_\theta)$ with respect to the weighting constant β. Plugging the terms $(*_1)$ and $(*_2)$ in the Fisher divergence, we have

$$\mathrm{D}(\pi^D_n\|\delta^B_\theta) = \frac{1}{B}\sum_{b=1}^B \beta^2 \|\nabla_\theta D_n(\theta^{(b)}_n)\|^2 - 2\beta \nabla_\theta D_n(\theta^{(b)}_n)\cdot \nabla_\theta \log \pi(\theta^{(b)}_n) - 2\beta\,\mathrm{Tr}\left(\nabla^2_\theta D_n(\theta^{(b)}_n)\right) + C$$

where we denote any term independent of β by C in this proof. Exchanging the order of the summation and the constant β, the Fisher divergence turns out to be a quadratic function of β as follows:

$$\begin{aligned}\mathrm{D}(\pi^D_n\|\delta^B_\theta) &= \beta^2 \underbrace{\frac{1}{B}\sum_{b=1}^B \|\nabla_\theta D_n(\theta^{(b)}_n)\|^2}_{=(a)} - 2\beta \underbrace{\frac{1}{B}\sum_{b=1}^B \nabla_\theta D_n(\theta^{(b)}_n)\cdot \nabla_\theta \log \pi(\theta^{(b)}_n) + \mathrm{Tr}\left(\nabla^2_\theta D_n(\theta^{(b)}_n)\right)}_{=(b)} + C \\ &= a\beta^2 - 2b\beta + C = a\left(\beta - \frac{b}{a}\right)^2 - \frac{b^2}{4a^2} + C\end{aligned}$$

where the last equality follows from completing the square. Therefore the Fisher divergence $\mathrm{D}(\pi^D_n\|\delta^B_\theta)$ is minimised at $\beta_* = b/a$, that is,

$$\beta_* = \frac{\sum_{b=1}^B \nabla_\theta D_n(\theta^{(b)}_n)\cdot \nabla_\theta \log \pi(\theta^{(b)}_n) + \mathrm{Tr}\left(\nabla^2_\theta D_n(\theta^{(b)}_n)\right)}{\sum_{b=1}^B \|\nabla_\theta D_n(\theta^{(b)}_n)\|^2},$$

as claimed, where the denominator and numerator are positive immediately from the first and second assumption respectively, which assures that $\beta_* > 0$. □

B Proof of Theorem 3: Simplified Conditions for Miller (2021)

Before showing that the preconditions C1-C5 of Theorem 3 are sufficient for (Miller, 2021, Theorem 4), we introduce the following lemma on a.s. uniform convergence used in the proof.

Lemma 4. *(a.s. uniform convergence) Suppose that the preconditions C1 and C2 in Theorem 3 holds for $r = 1$. Then D_n a.s. converges uniformly to D on the bounded convex open set U in Theorem 3.*

Proof. Davidson (1994, Theorem 21.8) showed that $D_n \xrightarrow{a.s.} D$ uniformly on U if and only if (a) $D_n \xrightarrow{a.s.} D$ pointwise on U and (b) $\{D_n\}_{n=1}^{\infty}$ is strongly stochastically equicontinuous on U. The condition (a) is immediately implied by the precondition C1 of Theorem 3 and hence the condition (b) is shown in the remainder. By Davidson (1994, Theorem 21.10), $\{D_n\}_{n=1}^{\infty}$ is strongly stochastically equicontinuous on U if there exists a stochastic sequence $\{\mathcal{L}_n\}_{n=1}^{\infty}$ independent of θ s.t.

$$|D_n(\theta) - D_n(\theta')| \leq \mathcal{L}_n \|\theta - \theta'\|_2, \quad \forall \theta, \theta' \in U \qquad \text{and} \qquad \limsup_{n\to\infty} \mathcal{L}_n < \infty \text{ a.s.}$$

Since D_n is continuously differentiable on the set U by the precondition C2 of Theorem 3 with $r = 1$, the mean value theorem yields that

$$|D_n(\theta) - D_n(\theta')| \leq \sup_{\theta\in U} \|\nabla_\theta D_n(\theta)\|_2 \|\theta - \theta'\|_2, \quad \forall \theta, \theta' \in U.$$

Again by the precondition C2 of Theorem 3 with $r = 1$, we have $\limsup_{n\to\infty} \sup_{\theta\in U} \|\nabla_\theta D_n(\theta)\|_2 < \infty$ a.s. Therefore, setting $\mathcal{L}_n = \sup_{\theta\in U} \|\nabla_\theta D_n(\theta)\|_2$ concludes the proof. □

We now show that (Miller, 2021, Theorem 4) holds a.s. under the preconditions C1-C5 of Theorem 3, which in turn implies Theorem 3 directly. A main argument in the proof is essentially same as that of Matsubara et al. (2022) but that is modified here to allow for an arbitrary loss D_n.

Proof. In order to apply (Miller, 2021, Theorem 4), we first extend π and D_n from Θ to $\mathbb{R}^p$ by setting $\pi(\theta) = 0$ and $D_n(\theta) = \sup_{\theta\in\Theta} |D_n(\theta)| + 1$ for all $\theta \in \mathbb{R}^p \setminus \Theta$, so that we have $\pi : \mathbb{R}^p \to \mathbb{R}$, $D_n : \mathbb{R}^p \to \mathbb{R}$ and $\pi_n^D : \mathbb{R}^p \to \mathbb{R}$. Note that in Miller (2021, Theorem 4), $\{D_n\}_{n=1}^{\infty}$ is regarded as a sequence of deterministic functions, while here $\{D_n\}_{n=1}^{\infty}$ is a sequence of stochastic functions dependent of random data $\{X_i\}_{i=1}^{n}$. It will be shown that Miller (2021, Theorem 4) holds a.s. for the stochastic sequence $\{D_n\}_{n=1}^{\infty}$. We hence verify the following prerequisites (1)–(6) of (Miller, 2021, Theorem 4) a.s. hold. Recall that $H_n(\theta) = \nabla_\theta^2 D_n(\theta)$ and $H_* = \lim_{n\to\infty} H_n(\theta_*)$ from Theorem 3:

1. the prior density π is continuous at θ_* and $\pi(\theta_*) > 0$.
2. $\theta_n \overset{a.s.}{\to} \theta_*$.
3. the Taylor expansion $D_n(\theta) = D_n(\theta_n) + (1/2)(\theta - \theta_n) \cdot H_n(\theta_n)(\theta - \theta_n) + r_n(\theta - \theta_n)$ holds on U a.s. where r_n is the reminder term.
4. the remainder r_n of the Taylor expansion satisfies that $|r_n(\theta)| \leq C\|\theta\|_2^3$, $\forall \theta \in B_\epsilon(0)$ a.s. for all n sufficiently large and some $\epsilon > 0$.
5. $H_n(\theta_n) \overset{a.s.}{\to} H_*$, $H_n(\theta_n)$ is symmetric for all n sufficiently large and H_* is positive definite.
6. $\liminf_{n\to\infty} \left(\inf_{\theta\in\mathbb{R}^p \setminus \mathcal{B}_\epsilon(\theta_n)} D_n(\theta) - D_n(\theta_n) \right) > 0$ a.s. for any $\epsilon > 0$.

Part (1): The precondition C5 of Theorem 3.

Part (2): The strong consistency $\theta_n \overset{a.s.}{\to} \theta_*$ is shown by an argument similar to van der Vaart (1998, Theorem 5.7) or essentially same as Matsubara et al. (2022, Lemma 3). First, it follows from Lemma 4 that $D_n \overset{a.s.}{\to} D$ uniformly on U under the conditions of Theorem 3. Thus, for all n sufficiently large, we can take

$\delta > 0$ s.t. $|D_n(\theta) - D(\theta)| < \delta/2$ a.s. over $\theta \in U$, which in turn leads to (a) $D(\theta) < D_n(\theta) + \delta/2$ and (b) $D_n(\theta) < D(\theta) + \delta/2$ a.s. over $\theta \in U$. Then applying both (a) and (b), the following bound on $D(\theta_n)$ holds for all n sufficiently large:

$$D(\theta_n) \overset{(a)}{<} D_n(\theta_n) + \delta/2 \overset{(*)}{\leq} D_n(\theta_*) + \delta/2 \overset{(b)}{<} D(\theta_*) + \delta \quad \text{a.s.} \tag{18}$$

where the second inequality $(*)$ follows from the fact that θ_n is the minimiser of D_n. Since $\inf_{\theta \in \mathbb{R}^p} D(\theta) = \inf_{\theta \in \Theta} D(\theta)$ is uniquely attained at $\theta_* \in U$ by Theorem 3 (3), for any $\epsilon > 0$ we have $D(\theta) - D(\theta_*) > 0$ for all $\theta \in \mathbb{R}^p \setminus B_\epsilon(\theta_*)$. Given an arbitrary $\epsilon > 0$, let $\delta = \inf_{\theta \in \Theta \setminus B_\epsilon(\theta_*)} D(\theta) - D(\theta_*) > 0$. It then follows from (18) that, for all n sufficiently large,

$$D(\theta_n) < \inf_{\theta \in \mathbb{R}^p \setminus B_\epsilon(\theta_*)} D(\theta) \quad \text{a.s.}$$

This implies that $\theta_n \in B_\epsilon(\theta_*)$ a.s. for any $\epsilon > 0$ arbitrary small for all n sufficiently large. Therefore $\theta_n \overset{a.s.}{\to} \theta_*$ by definition of convergence.

Part (3): From the precondition C2 of Theorem 3, D_n is 3 times continuously differentiable over U. Noting that $\nabla_\theta D_n(\theta) = 0$ at a minimiser θ_n of D_n, the Taylor expansion of D_n around the minimiser θ_n gives that

$$D_n(\theta) = D_n(\theta_n) + \frac{1}{2}(\theta - \theta_n) \cdot H_n(\theta_n)(\theta - \theta_n) + r_n(\theta - \theta_n)$$

where r_n is the remainder of the Taylor expansion.

Part (4): Since r_n is the remainder of the Taylor expansion, we have an upper bound

$$|r_n(\theta - \theta_n)| \leq \frac{1}{6} \sup_{\theta \in U} \|\nabla^3_\theta D_n(\theta)\|_2 \|\theta - \theta_n\|_2^3, \quad \forall \theta \in U.$$

The precondition C2 of Theorem 3 guarantees that $\limsup_{n \to \infty} \sup_{\theta \in U} \|\nabla^3_\theta D_n(\theta)\|_2 < \infty$ a.s. It is thus possible to take some positive constant C s.t. $(1/6) \sup_{\theta \in U} \|\nabla^3_\theta D_n(\theta)\|_2 \leq C$ a.s. for all n sufficiently large. For all n sufficiently large, there exists some open ϵ-neighbour $B_\epsilon(\theta_n)$ contained in the open set U since $\theta_n \in U$. Combining these two facts concludes that

$$|r_n(\theta - \theta_n)| \leq C\|\theta - \theta_n\|_2^3, \quad \forall \theta \in B_\epsilon(\theta_n) \quad \Longrightarrow \quad |r_n(\theta)| \leq C\|\theta\|_2^3, \quad \forall \theta \in B_\epsilon(0)$$

holds for some $\epsilon > 0$.

Part (5): We first show that $\|H_n(\theta_n) - H_*\|_2 \overset{a.s.}{\to} 0$. By the triangle inequality,

$$\|H_n(\theta_n) - H_*\|_2 \leq \|H_n(\theta_n) - H_n(\theta_*)\|_2 + \|H_n(\theta_*) - H_*\|_2 .$$

For the first term, it follows from the mean value theorem that

$$\|H_n(\theta_n) - H_n(\theta_*)\|_2 \leq \sup_{\theta \in U} \|\nabla_\theta H_n(\theta)\|_2 \|\theta_n - \theta_*\|_2 = \sup_{\theta \in U} \|\nabla^3_\theta D_n(\theta)\|_2 \|\theta_n - \theta_*\|_2.$$

The precondition C2 of Theorem 3 guarantees that $\limsup_{n \to \infty} \sup_{\theta \in U} \|\nabla^3_\theta D_n(\theta)\|_2 < \infty$ a.s. It is thus possible to take some positive constant C' s.t. $\|H_n(\theta_n) - H_n(\theta_*)\|_2 \leq C' \|\theta_n - \theta_*\|_2$ for all n sufficiently large. Then we have $\|H_n(\theta_n) - H_n(\theta_*)\|_2 \overset{a.s.}{\to} 0$ by the preceding part (2) $\theta_n \overset{a.s.}{\to} \theta_*$. For the second term, it is directly implied by the precondition C4 of Theorem 3 that $\|H_n(\theta_*) - H_*\|_2 \overset{a.s.}{\to} 0$. Combining these two facts concludes that $\|H_n(\theta_n) - H_*\|_2 \overset{a.s.}{\to} 0$. We next show that $H_n(\theta_n)$ is symmetric. The (i, j) entry of $H_n(\theta) = \nabla^2_\theta D_n(\theta)$ is given by the partial derivative $(\partial^2 / \partial\theta_i \partial\theta_j) D_n(\theta)$ with respect to i-th and j-th entry of θ. Since D_n is twice continuously differentiable by the precondition C2 of Theorem 3, the Schwartz's theorem implies that the commutation $(\partial^2 / \partial\theta_i \partial\theta_j) D_n(\theta) = (\partial^2 / \partial\theta_j \partial\theta_i) D_n(\theta)$ holds and therefore $H_n(\theta)$ is symmetric for any $\theta \in \Theta$. Finally we show positive definiteness of H_*. For all n sufficiently large, $H_n(\theta_n)$

is positive semi-definite by the fact that θ_n is the minimiser of D_n and accordingly the limit H_* is positive semi-definite. Then H_* is positive definite since H_* is nonsingular by the precondition C4 of Theorem 3.

Part (6): It holds for any sequence $a_n, b_n \in \mathbb{R}$ that $\liminf_{n\to\infty}(a_n - b_n) \geq \liminf_{n\to\infty} a_n + \liminf_{n\to\infty}(-b_n)$. Furthermore from the property that $\liminf_{n\to\infty}(-b_n) = -\limsup_{n\to\infty} b_n$, we have $\liminf_{n\to\infty}(a_n - b_n) \geq \liminf_{n\to\infty} a_n - \limsup_{n\to\infty} b_n$. Applying this, we have

$$\liminf_{n\to\infty}\left(\inf_{\theta\in\mathbb{R}^p\setminus B_\epsilon(\theta_n)} D_n(\theta) - D_n(\theta_n)\right) = \underbrace{\liminf_{n\to\infty}\inf_{\theta\in\mathbb{R}^p\setminus B_\epsilon(\theta_n)} D_n(\theta)}_{=:(*_1)} - \underbrace{\limsup_{n\to\infty} D_n(\theta_n)}_{=:(*_2)}.$$

For the first term $(*_1)$, it is obvious from the way of extending D_n from Θ to $\mathbb{R}^p$ that

$$(*_1) = \liminf_{n\to\infty}\inf_{\theta\in\mathbb{R}^p\setminus B_\epsilon(\theta_n)} D_n(\theta) \geq \liminf_{n\to\infty}\inf_{\theta\in\Theta\setminus B_\epsilon(\theta_n)} D_n(\theta) \quad \text{a.s.}$$

For any set $A \subset \mathbb{R}^p$ and function $g : \mathbb{R}^p \to \mathbb{R}$, define $\inf_{\theta\in A\setminus B_\epsilon(\theta_n)} g(\theta) := \sup_{\theta\in A} g(\theta)$ if $A \setminus B_\epsilon(\theta_n)$ is empty. Decomposing Θ into two sets U and $\Theta \setminus U$ leads to

$$(*_1) \geq \liminf_{n\to\infty}\inf_{\theta\in\Theta\setminus B_\epsilon(\theta_n)} D_n(\theta) \geq \min\Big(\underbrace{\liminf_{n\to\infty}\inf_{\theta\in U\setminus B_\epsilon(\theta_n)} D_n(\theta)}_{=:(*_{11})}, \underbrace{\liminf_{n\to\infty}\inf_{\theta\in\Theta\setminus(U\cup B_\epsilon(\theta_n))} D_n(\theta)}_{=:(*_{12})}\Big) \quad \text{a.s.}$$

For the term $(*_{11})$, since $D_n \overset{a.s.}{\to} D$ uniformly on U by Lemma 4 and $\theta_n \overset{a.s.}{\to} \theta_*$ by the preceding part (2),

$$(*_{11}) = \liminf_{n\to\infty}\inf_{\theta\in U\setminus B_\epsilon(\theta_n)} D_n(\theta) = \lim_{n\to\infty}\inf_{\theta\in U\setminus B_\epsilon(\theta_n)} D_n(\theta) = \inf_{\theta\in U\setminus B_\epsilon(\theta_*)} D(\theta) \quad \text{a.s.}$$

For the term $(*_{12})$, since the global minimiser θ_n of D_n is contained in U a.s. for all n sufficiently large by the precondition C3 of Theorem 3,

$$(*_{12}) = \liminf_{n\to\infty}\inf_{\theta\in\Theta\setminus(U\cup B_\epsilon(\theta_n))} D_n(\theta) > \liminf_{n\to\infty}\inf_{\theta\in U} D_n(\theta) = \inf_{\theta\in U} D(\theta) = D(\theta_*) \quad \text{a.s.}$$

where the second equality follows from the a.s. uniform convergence of D_n on U by Lemma 4. For the second term $(*_2)$, again since $D_n \overset{a.s.}{\to} D$ uniformly on U and $\theta_n \overset{a.s.}{\to} \theta_*$, we have

$$(*_2) = \limsup_{n\to\infty} D_n(\theta_n) = \lim_{n\to\infty} D_n(\theta_n) = D(\theta_*) \quad \text{a.s.}$$

The original term $(*_1) - (*_2)$ is lower bounded by $(*_1) - (*_2) \geq \min((*_{11}) - (*_2), (*_{12}) - (*_2))$ a.s., and both the term $(*_{11}) - (*_2)$ and $(*_{12}) - (*_2)$ are then further lower bounded by

$$(*_{11}) - (*_2) = \inf_{\theta\in U\setminus B_\epsilon(\theta_*)} D(\theta) - D(\theta_*) > 0 \quad \text{and} \quad (*_{12}) - (*_2) > D(\theta_*) - D(\theta_*) = 0 \quad \text{a.s.},$$

where the first inequality follows from the precondition C3 of Theorem 3 indicating that $\inf_{\theta\in\Theta} D(\theta)$ is uniquely attained at $\theta_* \in U$. Therefore we have $(*_1) - (*_2) \geq \min((*_{11}) - (*_2), (*_{12}) - (*_2)) > 0$ a.s., which concludes the proof. □

C Relation to Stein Discrepancies

Fisher divergences can be related to a more general class of divergences called Stein discrepancies. Since their introduction, Stein discrepancies have demonstrated utility over a range of statistical applications, including hypothesis testing, parameter estimation, variational inference, and post-processing of Markov chain Monte Carlo; see Anastasiou et al. (2021) for a review.

This appendix clarifies the sense in which discrete Fisher divergence can be seen as a special case of a discrete Stein discrepancy with an L^2-based Stein set. The continuous case was previously covered by Theorem 2 in Barp et al. (2019). As a consequence, we deduce that the discrete Fisher divergence is stronger than the popular class of Stein discrepancies based on reproducing kernels.

C.1 Background on Stein Discrepancies

Let $\mathcal{X}_*$ be a locally compact Hausdorff space. For a set $\mathcal{H}$ of functions $f : \mathcal{X}_* \to \mathbb{R}^d$, an operator $S_p : \mathcal{H} \to L^1(p, \mathbb{R}^m)$ depending on a probability distribution p on $\mathcal{X}_*$ is called a *Stein operator* if $\mathbb{E}_{X\sim p}[S_p[h](X)] = 0$ for all $h \in \mathcal{H}$. In these circumstances, we refer to $\mathcal{H}$ as a *Stein set*. The next proposition defines a particular Stein operator that arises naturally when considering discrete domains $\mathcal{X}_* = \mathcal{X}$, where we recall that $\mathcal{X}$ is a countable space in Standing Assumption 1. The reader is referred to Shi et al. (2022) for discussion of alternative Stein operators in the discrete context.

Proposition 2. *Let p be a positive probability distribution on $\mathcal{X}$, such that $(\nabla^- p)/p \in L^2(p, \mathbb{R}^d)$. Define an operator S_p, acting on functions $h \in L^2(p, \mathbb{R}^d)$, by*

$$S_p[h](\boldsymbol{x}) := \frac{\nabla^- p(\boldsymbol{x})}{p(\boldsymbol{x})} \cdot h(\boldsymbol{x}) + \nabla^+ \cdot h(\boldsymbol{x}). \tag{19}$$

Then it holds that $\mathbb{E}_{X\sim p}[S_p[h](X)] = 0$.

Proof. By positivity of p and Cauchy–Schwarz, observe that

$$\begin{aligned}\sum_{\boldsymbol{x}\in\mathcal{X}} |p(\boldsymbol{x}^{i-})h_i(\boldsymbol{x})| = \sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x})\frac{p(\boldsymbol{x}^{i-})}{p(\boldsymbol{x})}|h_i(\boldsymbol{x})| &= \mathbb{E}_{X\sim p}\left[\frac{p(X^{i-})}{p(X)}|h_i(X)|\right] \\ &\leq \mathbb{E}_{X\sim p}\left[\frac{p(X^{i-})^2}{p(X)^2}\right]\mathbb{E}_{X\sim p}\left[h_i(X)^2\right] < \infty\end{aligned} \tag{20}$$

where the first and second term are implied to be finite for any $i = 1, \ldots, d$ since $h \in L^2(p, \mathbb{R}^d)$ and $(\nabla^- p)/p \in L^2(p, \mathbb{R}^d)$ which implies $[\nabla^- p(\boldsymbol{x})/p(\boldsymbol{x})]_i = 1 - p(\boldsymbol{x}^{i-})/p(\boldsymbol{x})$ is square integrable with respect to p.

Now, using the definition of ∇^- and $\nabla^+\cdot$, the Stein operator S_p can be simplified as

$$S_p[h](\boldsymbol{x}) = \sum_{i=1}^{d} h_i(\boldsymbol{x}^{i+}) - \frac{p(\boldsymbol{x}^{i-})}{p(\boldsymbol{x})}h_i(\boldsymbol{x}). \tag{21}$$

The expectation of interest can then be expressed as

$$\mathbb{E}_{X\sim p}[S_p[h](X)] = \sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x})S_p[h](\boldsymbol{x}) = \sum_{i=1}^{d}\sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x})h_i(\boldsymbol{x}^{i+}) - \sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x}^{i-})h_i(\boldsymbol{x}),$$

where we have used the absolute convergence of the series, established in (20), to justify the re-ordering of terms. The result is then immediate from Lemma 2. □

The Stein operator (19) can be considered a discrete analogue of the Langevin Stein operator for continuous domains; see Yang et al. (2018).

Given a Stein operator S_p and Stein set $\mathcal{H}$, the *Stein discrepancy* between probability distributions p and q on $\mathcal{X}$ is defined as the maximum deviation between expectations of the test functions $S_p[h]$ for $h \in \mathcal{H}$:

$$\mathrm{SD}(p\|q) := \sup_{h\in\mathcal{H}} |\mathbb{E}_{X\sim q}\left[S_p[h](X)\right] - \mathbb{E}_{X\sim p}\left[S_p[h](X)\right]| = \sup_{h\in\mathcal{H}} |\mathbb{E}_{X\sim q}\left[S_p[h](X)\right]| \tag{22}$$

The final equality follows from Proposition 2, and our discussion in this appendix implicitly assumes all relevant quantities are well-defined. The Stein discrepancy is computable[9] without knowing the normalising constant of p since it depends on p only through the ratio $(\nabla^- p)/p$, in a similar manner to discrete Fisher divergence in the main text.

[9]That is, the expectations do not involve the normalising constant; whether the supremum over the Stein set is computable depends on how the Stein set is selected.

C.2 Discrete Fisher Divergence as a Stein Discrepancy

We now establish that the discrete Fisher divergence, introduced in the main text, is in fact a Stein discrepancy, corresponding to the Stein operator in Proposition 2 and a Stein set equal to the unit ball of $L^2(q, \mathbb{R}^d)$. This observation will allow us to conclude, in Appendix C.3, that discrete Fisher divergence is stronger than popular kernel Stein discrepancies.

Proposition 3. *Let p and q be positive distributions on $\mathcal{X}$, such that $(\nabla^- p)/p, (\nabla^- q)/q \in L^2(q, \mathbb{R}^d)$. Consider a Stein discrepancy whose Stein operator is* (19) *and whose Stein set is $\mathcal{H} = \{h : \mathcal{X} \to \mathbb{R}^d \mid \sum_{i=1}^d \mathbb{E}_{X\sim q}[h_i(X)^2] \leq 1\}$. Then*

$$\mathrm{SD}(p\|q) = \sqrt{\mathrm{DFD}(p\|q)}. \tag{23}$$

Proof. From (19) and (22), we have that

$$\mathrm{SD}(p\|q) = \sup_{h\in\mathcal{H}} \left| \mathbb{E}_{X\sim q}\left[\frac{\nabla^- p(X)}{p(X)} \cdot h(X) - \frac{\nabla^- q(X)}{q(X)} \cdot h(X) \right] \right|.$$

Note that $L^2(q, \mathbb{R}^d)$ is a Hilbert space when equipped with the inner product $\langle f, g\rangle_{L^2(q,\mathbb{R}^d)} := \mathbb{E}_{X\sim q}[f(X) \cdot g(X)]$. Thus we can view $\mathrm{SD}(p\|q)$ as the maximum of the inner product

$$\mathrm{SD}(p\|q) = \sup_{h\in\mathcal{H}} \left| \left\langle \frac{\nabla^- p}{p} - \frac{\nabla^- q}{q}, h \right\rangle_{L^2(q,\mathbb{R}^d)} \right|, \tag{24}$$

which is well-defined since $u := (\nabla^- p)/p - (\nabla^- q)/q \in L^2(q, \mathbb{R}^d)$. Let $\|\cdot\|_{L^2(q,\mathbb{R}^d)}$ denote the norm of $L^2(q, \mathbb{R}^d)$, so that $\mathcal{H}$ is the set of $f \in L^2(q, \mathbb{R}^d)$ for which $\|f\|_{L^2(q,\mathbb{R}^d)} \leq 1$. By the Cauchy–Schwarz inequality, the inner product in (24) attains its supremum at $h = u/\|u\|_{L^2(q,\mathbb{R}^d)} \in \mathcal{H}$. Therefore

$$\mathrm{SD}(p\|q) = \sup_{h\in\mathcal{H}} |\langle u, h\rangle_{L^2(q,\mathbb{R}^d)}| = \|u\|_{L^2(q,\mathbb{R}^d)} = \sqrt{\mathbb{E}_{X\sim q}\left[\left\| \frac{\nabla^- p(X)}{p(X)} - \frac{\nabla^- q(X)}{q(X)} \right\|^2 \right]},$$

which concludes the proof. □

C.3 The Fisher Divergence Dominates the Kernel Stein Discrepancy

A popular choice of Stein set $\mathcal{H}$, that can lead to a closed form Stein discrepancy, is the unit ball of a reproducing kernel Hilbert space. The resulting *kernel Stein discrepancy* was recently considered in the discrete context in Yang et al. (2018). In this appendix we establish that our discrete Fisher divergence, introduced in the main text, is a stronger notion of divergence than kernel Stein discrepancy. This may render the discrete Fisher divergence more statistically efficient in applications where a statistical model is well-specified, in addition to the computational advantage (Remark 1) and the non-reliance on a user-specified kernel discussed in the main text.

A symmetric, positive definite function $k : \mathcal{X}_* \times \mathcal{X}_* \to \mathbb{R}$ is called a kernel. For every kernel k, there exists a unique associated Hilbert space of real-valued functions on $\mathcal{X}_*$, called a reproducing kernel Hilbert space, denoted $\mathcal{H}_k$; see e.g. Berlinet and Thomas-Agnan (2011) for background. Let $\mathcal{H}_k^d := \mathcal{H}_k \times \cdots \times \mathcal{H}_k$, that is, a space of functions $h : \mathcal{X}_* \to \mathbb{R}^d$ whose each i-th output-coordinate $h_i : \mathcal{X}_* \to \mathbb{R}$ belongs to $\mathcal{H}_k$. Yang et al. (2018) studied the Stein discrepancy for a discrete space $\mathcal{X}_* = \mathcal{X}$, of finite cardinality only, using the Stein operator (19) and a Stein set $\mathcal{H} = \{h \in \mathcal{H}_k^d : \sum_{i=1}^d \|h_i\|_{\mathcal{H}_k}^2 \leq 1\}$. Here we first establish that the Stein set $\{h \in \mathcal{H}_k^d : \sum_{i=1}^d \|h_i\|_{\mathcal{H}_k}^2 \leq 1\}$ constructed from $\mathcal{H}_k^d$ is contained in another Stein set $\{h \in L^2(q, \mathbb{R}^d) : \|h\|_{L^2(q,\mathbb{R}^d)}^2 \leq 1\}$ constructed from $L^2(q, \mathbb{R}^d)$ for any general domain $\mathcal{X}_*$, under a standard condition on the reproducing kernel. This in turn shows that the discrete Fisher divergence dominates the kernel Stein discrepancy.

Proposition 4. *Let q be a probability distribution on $\mathcal{X}_*$. Let $k : \mathcal{X}_* \times \mathcal{X}_* \to \mathbb{R}$ be a kernel such that $k(\boldsymbol{x}, \boldsymbol{x}) \leq 1$ for all $\boldsymbol{x} \in \mathcal{X}_*$. Then the unit ball of $\mathcal{H}_k^d$ is contained in the unit ball of $L^2(q, \mathbb{R}^d)$.*

Proof. First let $f : \mathcal{X}_* \to \mathbb{R}^d$ be any element of $\mathcal{H}_k^d$, where its i-th output-coordinate $f_i : \mathcal{X}_* \to \mathbb{R}$ belongs to $\mathcal{H}_k$ each. From the reproducing property of $\mathcal{H}_k$, followed by the Cauchy–Schwartz inequality, the norm of f in $L^2(q, \mathbb{R}^d)$ is upper bounded as follows:

$$\begin{aligned}
\|f\|^2_{L^2(q,\mathbb{R}^d)} = \sum_{i=1}^d \mathbb{E}_{X\sim q}[f_i(X)^2] &= \sum_{i=1}^d \mathbb{E}_{X\sim q}[\langle f_i(\cdot), k(X,\cdot)\rangle^2_{\mathcal{H}_k}] \\
&\leq \sum_{i=1}^d \mathbb{E}_{X\sim q}\left[\|f_i\|^2_{\mathcal{H}_k}\|k(X,\cdot)\|^2_{\mathcal{H}_k}\right] = \sum_{i=1}^d \mathbb{E}_{X\sim q}\left[\|f_i\|^2_{\mathcal{H}_k} k(X,X)\right] \\
&= \left(\sum_{i=1}^d \|f_i\|^2_{\mathcal{H}_k}\right) \mathbb{E}_{X\sim q}\left[k(X,X)\right] = \|f\|^2_{\mathcal{H}_k^d}\, \mathbb{E}_{X\sim q}\left[k(X,X)\right].
\end{aligned}$$

The continuous embedding of $\mathcal{H}_k^d$ in $L^2(q, \mathbb{R}^d)$ therefore holds, and moreover the embedding constant is at most one, since $\mathbb{E}_{X\sim q}[k(X, X)] \leq 1$ due to the assumption that $k(\boldsymbol{x}, \boldsymbol{x}) \leq 1$ for all $\boldsymbol{x} \in \mathcal{X}_*$. In particular, it follows that the unit ball of $\mathcal{H}_k^d$ is contained in the unit ball of $L^2(q, \mathbb{R}^d)$. □

Built upon Proposition 4, we can immediately show the discrete Fisher divergence dominates the kernel Stein discrepancy for the case where $\mathcal{X}_* = \mathcal{X}$.

Proposition 5. *Let p and q be positive distributions on $\mathcal{X}$, such that $(\nabla^- p)/p, (\nabla^- q)/q \in L^2(q, \mathbb{R}^d)$. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel such that $k(\boldsymbol{x}, \boldsymbol{x}) \leq 1$ for all $\boldsymbol{x} \in \mathcal{X}$. Let $\mathcal{S}_p$ be a Stein operator in* (19)*. Then the kernel Stein discrepancy, denoted* SD_k*, satisfies* $\mathrm{SD}_k(p\|q) \leq \sqrt{\mathrm{DFD}(p\|q)}$.

Proof. From (22) (which in turn relies on Proposition 2), it is straightforward to see that

$$\mathrm{SD}_k(p\|q) = \sup_{\|h\|_{\mathcal{H}_k^d} \leq 1} |\mathbb{E}_{X\sim q}\left[S_p[h](X)\right]| \leq \sup_{\|h\|_{L^2(q,\mathbb{R}^d)} \leq 1} |\mathbb{E}_{X\sim q}\left[S_p[h](X)\right]| = \sqrt{\mathrm{DFD}(p\|q)},$$

where the inequality follows from Proposition 4 immediately and the final equality is Proposition 3. □

This argument is not restricted to the discrete case but is immediately applicable for the continuous case. One of the most common Stein operator for a continuous domain $\mathcal{X}_* = \mathbb{R}^d$ is

$$\mathcal{S}_p[h](\boldsymbol{x}) = \nabla \log p(\boldsymbol{x}) \cdot h(\boldsymbol{x}) + \nabla \cdot h(\boldsymbol{x}) \tag{25}$$

The Fisher divergence $\mathrm{FD}(p\|q) = \mathbb{E}_{X\sim q}[\|\nabla \log p(X) - \nabla \log q(X)\|^2]$ for densities p, q on $\mathcal{X}_*$ dominates the kernel Stein discrepancy constructed from the above Stein operator and the kernel on $\mathcal{X}_*$.

Proposition 6. *Let $\mathcal{X}_* = \mathbb{R}^d$. Let p and q be positive continuously differentiable densities on $\mathcal{X}_*$, such that $\nabla \log p, \nabla \log q \in L^2(q, \mathbb{R}^d)$. Let $k : \mathcal{X}_* \times \mathcal{X}_* \to \mathbb{R}$ be a kernel such that $k(\boldsymbol{x}, \boldsymbol{x}) \leq 1$ for all $\boldsymbol{x} \in \mathcal{X}_*$. Let $\mathcal{S}_p$ be a Stein operator in* (25) *Then the kernel Stein discrepancy, denoted* SD_k*, satisfies* $\mathrm{SD}_k(p\|q) \leq \sqrt{\mathrm{FD}(p\|q)}$.

Proof. We repeat the same argument as Proposition 5. From (Barp et al., 2019, Theorem 2), $\mathrm{FD}(p\|q)$ can be written as the Stein discrepancy constructed by the Stein set $\{h \in L^2(q, \mathbb{R}^d) : \|h\|^2_{L^2(q,\mathbb{R}^d)} \leq 1\}$. Then from (22) (which in turn relies on Proposition 2), it is straightforward to see that

$$\mathrm{SD}_k(p\|q) = \sup_{\|h\|_{\mathcal{H}_k^d} \leq 1} |\mathbb{E}_{X\sim q}\left[S_p[h](X)\right]| \leq \sup_{\|h\|_{L^2(q,\mathbb{R}^d)} \leq 1} |\mathbb{E}_{X\sim q}\left[S_p[h](X)\right]| = \sqrt{\mathrm{FD}(p\|q)},$$

where the inequality follows from Proposition 4 immediately and the final equality is Proposition 3. □

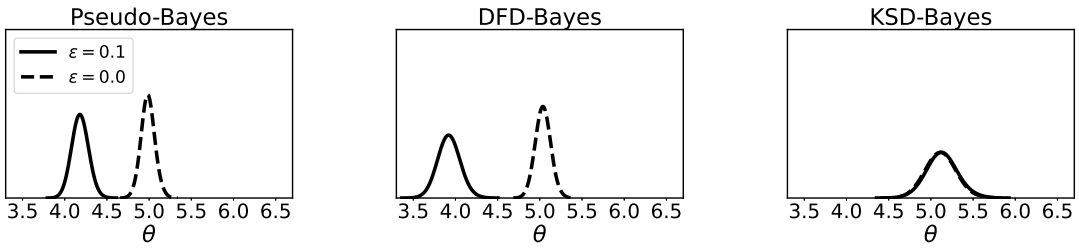

Figure 6: Posteriors of Pseudo-Bayes (left), DFD-Bayes (centre), and KSD-Bayes (right) for the Ising model in the presence of outlier with $\epsilon = 0.1$ and no outlier with $\epsilon = 0.0$.

C.4 Robustness of the Kernel Stein Discrepancy

Appendix C.3 indicates statistical efficiency of the discrete Fisher divergence over the kernel Stein discrepancy. If one's model is well-specified, minimising the discrete Fisher divergence leads us to a correct model faster than the kernel Stein discrepancy. However this does not mean that the use of the discrete Fisher divergence is always better than the kernel Stein discrepancy. In particular, the kernel Stein discrepancy can be equipped with strong robustness by choosing an appropriate kernel. To demonstrate this, we compare three posteriors of Pseudo-Bayes, DFD-Bayes, and KSD-Bayes for the same Ising model as Section 4.2 with $d = 100$ ($m = 10$) in a setting where dataset contains extreme outliers with a proportion ϵ.

We approximately draw 1000 samples $\{x_i\}_{i=1}^{1000}$ from the Ising model p_θ with $\theta = 5$ by the same Metropolis–Hastings algorithm as Section 4.2. To study the robustness of the posteriors, we replaced a proportion $\epsilon = 0.1$ of the data with the vector $(1, 1, \cdots, 1)$ corresponding to the extreme value in $\mathcal{X}$ that is rarely drawn from the model. Matsubara et al. (2022) showed that KSD-Bayes can satisfy strong qualitative robustness called "global bias-robustness" by choosing a kernel appropriately. For this example, we use the same choice of kernel as Matsubara et al. (2022) below:

$$k(\boldsymbol{x}, \boldsymbol{x}') = m(\boldsymbol{x}) \exp\left(-\frac{1}{d}\sum_{i=1}^{d} \mathbb{1}(x_i - x_i')\right) m(\boldsymbol{x})$$

where $m(x) = \sigma(90 - |\sum_i x_i|)$ based on a sigmoid function $\sigma(t) = (1 + \exp(-t))^{-1}$. This is indeed a proper choice of kernel, and the function $m(\boldsymbol{x})$ in the definition of kernel is designed to restrict the influence of extreme data whose norm is closer to or larger than 90.

In Figure 6 demonstrated that KSD-Bayes offered a correct inference outcome even when the dataset contains outliers, being less affected by the outliers. On the other hand, the Pseudo-Bayes and DFD-Bayes posteriors placed the majority of the probability mass on smaller θ than the correct value $\theta = 5$. The extreme value $(1, 1, \cdots, 1)$ of the outliers is more likely to be drawn from the model of $\theta \ll 1$; the posteriors of Pseudo-Bayes and DFD-Bayes were thus pulled in the direction of smaller θ.

D Calculations for Worked Examples

D.1 Assumption 2 for Example 1

The aim of this section is to establish when Assumption 2 is satisfied for the exponential family model in Example 1. For better presentation, let $T_{j-}(\boldsymbol{x}) := T(\boldsymbol{x}^{j-}) - T(\boldsymbol{x})$ and $b_{j-}(\boldsymbol{x}) := b(\boldsymbol{x}^{j-}) - b(\boldsymbol{x})$ to see that $r_{j-}(\boldsymbol{x}, \theta) = \exp(\eta(\theta) \cdot T_{j-}(\boldsymbol{x}) + b_{j-}(\boldsymbol{x}))$. In addition, let $T_{j+}(\boldsymbol{x}) := T(\boldsymbol{x}) - T(\boldsymbol{x}^{j+})$ and $b_{j+}(\boldsymbol{x}) := b(\boldsymbol{x}) - b(\boldsymbol{x}^{j+})$

to see that $r_{j-}(\boldsymbol{x}^{j+}, \theta) = \exp(\eta(\theta) \cdot T_{j+}(\boldsymbol{x}) + b_{j+}(\boldsymbol{x}))$. It is straightforward to see that, for any $\boldsymbol{x} \in \mathcal{X}$,

$$\begin{aligned}
\nabla_\theta r_{j-}(\boldsymbol{x}^{j+}, \theta) &= \nabla_\theta \eta(\theta) \cdot T_{j+}(\boldsymbol{x}) \, \exp(\eta(\theta) \cdot T_{j+}(\boldsymbol{x}) + b_{j+}(\boldsymbol{x})) \\
&= \nabla_\theta \eta(\theta) \cdot T_{j+}(\boldsymbol{x}) \, \exp(\eta(\theta) \cdot T_{j+}(\boldsymbol{x})) \, \exp(b_{j+}(\boldsymbol{x})) \\
\nabla_\theta (r_{j-}(\boldsymbol{x}, \theta)^2) &= 2 r_{j-}(\boldsymbol{x}, \theta) \, \nabla_\theta r_{j-}(\boldsymbol{x}, \theta) \\
&= 2 \nabla_\theta \eta(\theta) \cdot T_{j-}(\boldsymbol{x}) \, \exp(2\eta(\theta) \cdot T_{j-}(\boldsymbol{x})) \, \exp(2 b_{j-}(\boldsymbol{x}))
\end{aligned}$$

By assumption, $T_{j-}(\boldsymbol{x})$ is bounded over all $\boldsymbol{x} \in \mathcal{X}$, which in turn shows that $T_{j+}(\boldsymbol{x}) = T_{j-}(\boldsymbol{x}^{j+})$ is bounded over all $\boldsymbol{x} \in \mathcal{X}$ since $\boldsymbol{x}^{j+} \in \mathcal{X}$. Further, by assumption, $\sup_{\theta \in U} \|\nabla_\theta \eta(\theta)\| < \infty$ and $\sup_{\theta \in U} \|\eta(\theta)\| < \infty$. Let M be a constant that upper bounds all the terms $\sup_{\boldsymbol{x} \in \mathcal{X}} \|T_{j-}(\boldsymbol{x})\|$, $\sup_{\boldsymbol{x} \in \mathcal{X}} \|T_{j+}(\boldsymbol{x})\|$, $\sup_{\theta \in U} \|\nabla_\theta \eta(\theta)\|$ and $\sup_{\theta \in U} \|\eta(\theta)\|$. Then we have

$$\begin{aligned}
\sup_{\theta \in U} \|\nabla_\theta r_{j-}(\boldsymbol{x}^{j+}, \theta)\| &\leq M^2 \exp\left(M^2\right) \exp(b_{j+}(\boldsymbol{x})), \\
\sup_{\theta \in U} \|\nabla_\theta (r_{j-}(\boldsymbol{x}, \theta)^2)\| &\leq 2M^2 \exp\left(2M^2\right) \exp(2 b_{j-}(\boldsymbol{x})).
\end{aligned}$$

Taking the expectations,

$$\mathbb{E}_{X \sim p}\left[\sup_{\theta \in U} \|\nabla_\theta r_{j-}(X^{j+}, \theta)\|\right] \leq M^2 \exp(M^2) \mathbb{E}_{X \sim p}\left[\exp(b_{j+}(X))\right], \tag{26}$$

$$\mathbb{E}_{X \sim p}\left[\sup_{\theta \in U} \|\nabla_\theta (r_{j-}(X, \theta)^2)\|\right] \leq 2M^2 \exp(2M^2) \mathbb{E}_{X \sim p}\left[\exp(2 b_{j-}(X))\right]. \tag{27}$$

By assumption $\mathbb{E}_{X \sim p}\left[\exp(2 b_{j-}(X))\right] = \mathbb{E}_{X \sim p}\left[\exp(b_{j-}(X))^2\right] < \infty$, and we now argue that this also implies $\mathbb{E}_{X \sim p}\left[\exp(b_{j+}(X))\right] < \infty$. Indeed, from Lemma 2,

$$\begin{aligned}
\mathbb{E}_{X \sim p}\left[\exp(b_{j+}(X))\right] &= \sum_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{x}) \exp(b(\boldsymbol{x}) - b(\boldsymbol{x}^{j+})) = \sum_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{x}^{j-}) \exp(b(\boldsymbol{x}^{j-}) - b(\boldsymbol{x})) \\
&= \sum_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{x}) \frac{p(\boldsymbol{x}^{j-})}{p(\boldsymbol{x})} \exp(b(\boldsymbol{x}^{j-}) - b(\boldsymbol{x})) = \mathbb{E}_{X \sim p}\left[\frac{p(X^{j-})}{p(X)} \exp(b_{j-}(X))\right].
\end{aligned}$$

Now, using the Cauchy–Schwartz inequality,

$$\mathbb{E}_{X \sim p}\left[\exp(b_{j+}(X))\right] \leq \mathbb{E}_{X \sim p}\left[\frac{p(X^{j-})^2}{p(X)^2}\right] \mathbb{E}_{X \sim p}\left[\exp(2 b_{j-}(X))\right]. \tag{28}$$

Existence of the first term in (28) is implied by the Standing Assumption $(\nabla^- p)/p \in L^2(p, \mathbb{R}^d)$, while existence of the second term in (28) was assumed. Therefore we have shown that (26) and (27) exist. Repeating an essentially identical argument, it is straightforward to see also that $\mathbb{E}_{X \sim p}\left[\sup_{\theta \in U} \|\nabla_\theta^s r_{j-}(X^{j+}, \theta)\|\right] < \infty$ and $\mathbb{E}_{X \sim p}\left[\sup_{\theta \in U} \|\nabla_\theta^s (r_{j-}(X, \theta)^2)\|\right] < \infty$ for $s = 2, 3$ as claimed.

D.2 Derivatives of (10) for Example 1

Automatic differentiation is an attractive and promising choice to compute (10) whenever it is available. Nonetheless, it is still straightforward for a majority of parametric models to compute the loss derivatives used in (10). This section aims to demonstrate a form of loss derivatives for a model in Example 1. The optimal β of (10) depends on the first and second derivative of a loss D specified by users. Consider the discrete Fisher divergence D_n that this paper established. The discrete Fisher divergence $D_n(\theta) = \mathrm{DFD}(p_\theta \| p_n)$ between a model p_θ in Example 1 and an empirical distribution p_n of data $\{\boldsymbol{x}_i\}_{i=1}^n$ is given as

$$D_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} \left(r_{j-}(\boldsymbol{x}, \theta)\right)^2 - 2 r_{j-}(\boldsymbol{x}^{j+}, \theta)$$

For simplicity, let $\eta(\theta) = \theta$ here. Then $r_{j-}(\boldsymbol{x}, \theta) = \exp(\theta \cdot T_{j-}(\boldsymbol{x}) + b_{j-}(\boldsymbol{x}))$ and $r_{j-}(\boldsymbol{x}^{j+}, \theta) = \exp(\theta \cdot T_{j+}(\boldsymbol{x}) + b_{j+}(\boldsymbol{x}))$ using the notations in Appendix D.1. Therefore the derivatives are

$$\begin{aligned}
\nabla_\theta r_{j-}(\boldsymbol{x}, \theta) &= T_{j-}(\boldsymbol{x}) \exp(\theta \cdot T_{j-}(\boldsymbol{x}) + b_{j-}(\boldsymbol{x})), \\
\nabla_\theta^2 r_{j-}(\boldsymbol{x}, \theta) &= T_{j-}(\boldsymbol{x}) \otimes T_{j-}(\boldsymbol{x}) \exp(\theta \cdot T_{j-}(\boldsymbol{x}) + b_{j-}(\boldsymbol{x})), \\
\nabla_\theta r_{j-}(\boldsymbol{x}^{j+}, \theta) &= T_{j+}(\boldsymbol{x}) \exp(\theta \cdot T_{j+}(\boldsymbol{x}) + b_{j+}(\boldsymbol{x})), \\
\nabla_\theta^2 r_{j-}(\boldsymbol{x}^{j+}, \theta) &= T_{j+}(\boldsymbol{x}) \otimes T_{j+}(\boldsymbol{x}) \exp(\theta \cdot T_{j+}(\boldsymbol{x}) + b_{j+}(\boldsymbol{x}))
\end{aligned}$$

where $\otimes$ denotes outer product. Built upon these components, we have the required first derivatives of $D_n(\theta)$

$$\begin{aligned}
\nabla_\theta D_n(\theta) &= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} \nabla_\theta \left(r_{j-}(\boldsymbol{x}, \theta)^2 \right) - 2 \nabla_\theta r_{j-}(\boldsymbol{x}^{j+}, \theta) \\
&= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} 2 r_{j-}(\boldsymbol{x}, \theta) \nabla_\theta r_{j-}(\boldsymbol{x}, \theta) - 2 \nabla_\theta r_{j-}(\boldsymbol{x}^{j+}, \theta) \\
&= \frac{2}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} T_{j-}(\boldsymbol{x}) \exp(\theta \cdot T_{j-}(\boldsymbol{x}) + b_{j-}(\boldsymbol{x}))^2 - T_{j+}(\boldsymbol{x}) \exp(\theta \cdot T_{j+}(\boldsymbol{x}) + b_{j+}(\boldsymbol{x}))
\end{aligned}$$

as well as the second derivative of $D_n(\theta)$

$$\begin{aligned}
\nabla_\theta^2 D_n(\theta) &= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} \nabla_\theta \left(2 r_{j-}(\boldsymbol{x}, \theta) \nabla_\theta r_{j-}(\boldsymbol{x}, \theta) \right) - \nabla_\theta \left(2 \nabla_\theta r_{j-}(\boldsymbol{x}^{j+}, \theta) \right) \\
&= \frac{2}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} \nabla_\theta r_{j-}(\boldsymbol{x}, \theta) \otimes \nabla_\theta r_{j-}(\boldsymbol{x}, \theta) + r_{j-}(\boldsymbol{x}, \theta) \nabla_\theta^2 r_{j-}(\boldsymbol{x}, \theta) - \nabla_\theta^2 r_{j-}(\boldsymbol{x}^{j+}, \theta) \\
&= \frac{2}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} 2 T_{j-}(\boldsymbol{x}) \otimes T_{j-}(\boldsymbol{x}) \exp(\theta \cdot T_{j-}(\boldsymbol{x}) + b_{j-}(\boldsymbol{x}))^2 - T_{j+}(\boldsymbol{x}) \otimes T_{j+}(\boldsymbol{x}) \exp(\theta \cdot T_{j+}(\boldsymbol{x}) + b_{j+}(\boldsymbol{x})).
\end{aligned}$$

Plugging these derivatives $\nabla_\theta D_n(\theta)$ and $\nabla_\theta^2 D_n(\theta)$ and a given $\nabla_\theta \log \pi(\theta)$ in (10), the optimal β is computed.

D.3 Assumption 2 for Poisson, Ising, and Conway-Maxwell-Poisson Models

Assumption 2 for the Poisson and Ising models used in the experiments can be verified as a special case of Example 1. Any Poisson model can be written in the form

$$p_\theta(x) \propto \exp\left(\log(\theta_1)\, x - \sum_{k=1}^{x} \log(k) \right).$$

This falls into a class of exponential family in Example 1 by setting $\eta(\theta) = \log(\theta)$, $T(x) = x$, and $b(x) = -\sum_{k=1}^{x} \log(k)$. This gives that $T(x-1) - T(x) = -1$ and $b(x-1) - b(x) = \log(x)$. The condition in Example 1 is then satisfied provided that $\mathbb{E}_{X \sim p}[\exp(\log(X))^2] = \mathbb{E}_{X \sim p}[X^2] < \infty$, i.e. p has a second moment. Similarly, any Ising model can be written in the form

$$p_\theta(x) \propto \exp(\theta \cdot T(\boldsymbol{x}))$$

where $T : \mathcal{X} \to \mathbb{R}^k$ is a vector of summary statistics that define the model. For Ising models, $\mathcal{X}$ is of finite cardinality and $T(\boldsymbol{x})$ is hence bounded for any $\boldsymbol{x} \in \mathcal{X}$. The conditions in Example 1 are then automatically satisfied.

The Conway-Maxwell-Poisson model falls into a class of exponential family but it is beyond the simplified case of Example 1. Nonetheless, Assumption 2 is still verifiable. Recall that the Conway-Maxwell-Poisson model has the form $p_\theta(x) \propto (\theta_1)^x (x!)^{-\theta_2}$ whose ratio function is given by $r_{j-}(x,\theta) = p_\theta(x-1)/p_\theta(x) = x^{\theta_2}/\theta_1$ where $\theta_1, \theta_2 \in (0,\infty)$. The derivative of the ratio with respect to $\theta = (\theta_1, \theta_2)$ is then given by

$$\nabla_\theta r_{j-}(x+1,\theta) = \left(-\frac{(x+1)^{\theta_2}}{\theta_1^2}, \frac{(x+1)^{\theta_2}\log(x+1)}{\theta_1} \right), \quad \nabla_\theta (r_{j-}(x,\theta))^2 = \left(-\frac{x^{2\theta_2}}{\theta_1^3}, \frac{x^{2\theta_2}\log x}{\theta_1^2} \right).$$

Note that the term $x^{2\theta_2}\log x$ in $\nabla_\theta (r_{j-}(x,\theta))^2$ is well-defined even at $x=0$ since it converges to 0 as $x \to 0$ if $\theta_2 > 0$ despite the individual term $\log x$ alone is not well-defined for $x = 0$. Let M_1 and M_2 be the infimum value of θ_1 and the supremum value of θ_2 for (θ_1, θ_2) in the bounded set U to see that

$$\sup_{\theta \in U} \|\nabla_\theta r_{j-}(x+1,\theta)\| = \left| \frac{(x+1)^{M_2}}{M_1^2} \right| + \left| \frac{(x+1)^{M_2}\log(x+1)}{M_1} \right|,$$
$$\sup_{\theta \in U} \|\nabla_\theta (r_{j-}(x,\theta))^2\| = \left| \frac{x^{2M_2}}{M_1^3} \right| + \left| \frac{x^{2M_2}\log x}{M_1^2} \right|.$$

We can derive the same quantity up to constants in the power exponent of each term for the second and third derivative. Then Assumption 2 imposes that expectations of these quantities with respect to the data generating distribution $x \sim p$ are finite. For example, the expectations for the first derivatives are

$$\mathbb{E}_{X\sim p}\left[\sup_{\theta \in U} \|\nabla_\theta r_{j-}(X+1,\theta)\| \right] = \frac{1}{M_1^2}\mathbb{E}_{X\sim p}\left[\left| (x+1)^{M_2} \right| \right] + \frac{1}{M_1}\mathbb{E}_{X\sim p}\left[\left| (x+1)^{M_2}\log(x+1) \right| \right],$$
$$\mathbb{E}_{X\sim p}\left[\sup_{\theta \in U} \|\nabla_\theta (r_{j-}(x,\theta))^2\| \right] = \frac{1}{M_1^3}\mathbb{E}_{X\sim p}\left[\left| x^{2M_2} \right| \right] + \frac{1}{M_1^2}\mathbb{E}_{X\sim p}\left[\left| x^{2M_2}\log x \right| \right],$$

where the boundedness is translated into the moment condition of p as above.

E Details of Experimental Assessment

This appendix contains full details for the experiments that were reported in the main text.

E.1 Conway–Maxwell–Poisson Model

E.1.1 Settings for KSD-Bayes

KSD-Bayes is a generalised posterior constructed by taking a kernel Stein discrepancy as a loss function; see (Matsubara et al., 2022). The approach requires us to specify a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, based on which the kernel Stein discrepancy is constructed. In these experiments, we adopted a kernel recommended by Yang et al. (2018) for the kernel Stein discrepancy in discrete domains $\mathcal{X}$ given by

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{1}{d}\sum_{i=1}^{d} \mathbb{1}(x_i = x_i') \right)$$

where $\mathbb{1}$ is an indicator function, taking values in $\{0,1\}$. The effect of kernel choice is difficult to predict in the discrete context; for example, Yang et al. (2018) found that the closely related kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^{d} \mathbb{1}(x_i = x_i')$, can perform poorly in moderate-to-high dimensions d when employed in a Stein discrepancy. General principles for kernel choice in the discrete setting have not yet been established. Thus, one of the advantages of DFD-Bayes is absence of any user-specified parameters of the method.

E.1.2 Markov Chain Monte Carlo

A Metropolis–Hasting algorithm was employed to sample from the standard Bayesian posterior, as well as KSD-Bayes and DFD-Bayes. For computational convenience, the parametrisation $\tilde{\theta}_1 = \log(\theta_1)$ and $\tilde{\theta}_2 = \log(\theta_2)$ was applied so that parameters are defined on an unbounded domain $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2) \in \mathbb{R}^2$. An isotropic Gaussian random walk proposal with covariance $\sigma^2 I$ was employed, with $\sigma = 0.1$ used for all experiments. The convergence of the Markov chain was diagnosed using univariate Gelman–Rubin statistics for each θ_1 and θ_2 computed from 10 independent chains. In total, 500 samples were obtained from each chain by thinning 5,000 samples, all after an initial burn-in of length 5,000. In all cases, the univariate Gelman–Rubin statistics were below 1.02, respectively for θ_1 and θ_2.

E.1.3 Sales Dataset of Shmueli et al. (2005)

This dataset consists of quarterly sales figures for a particular item of clothing, taken across the different stores of a large national retailer. The original dataset is publicly available at `https://www.stat.cmu.edu/COM-Poisson/Sales-dat` see Shmueli et al. (2005). Quarterly sales at each store can be small and result in a large proportion of 0 entries in the dataset, so that the Conway–Maxwell–Poisson model has a clear advantage against the standard Poisson model.

To obtain a maximum *a posteriori* estimate for the parameters of the Conway–Maxwell–Poisson model for this sales dataset, Shmueli et al. (2005) considered a prior π defined by

$$\pi(\theta) \propto \theta_1^{a-1} \exp(-b\theta_2) \left(\sum_{j=1}^{\infty} \theta_1^j / (j!)^{\theta_2} \right)^{-c} \kappa(a, b, c) \tag{29}$$

where (a, b, c) is the hyper-parameter and $\kappa(a, b, c)$ is the normalising constant of π. The motivation to use this prior is conjugacy, since the resulting posterior takes the same form as (29). However, the prior itself contains the intractable terms $(\sum_{j=1}^{\infty} \theta_1^j / (j!)^{\theta_2})^{-c}$ and $\kappa(a, b, c)$. To avoid this additional intractability, which is not a focus of the present work, we considered a simpler chi-squared prior distribution in the main text.

E.2 Ising Model

E.2.1 Simulating Data from the Ising Model

Samples from the Ising model were obtained using the same Metropolis–Hasting algorithm used in Yang et al. (2018). First, all coordinates x_i of $\boldsymbol{x}$ were randomly initialised to either -1 or 1 with equiprobability 1/2. Then, at each iteration, we randomly select one coordinate x_i of $\boldsymbol{x}$ and flip the value of x_i either from -1 to 1 or from 1 to -1, where the flipped value $\tilde{x}_i$ is accepted with probability $\min(1, \exp(-2\tilde{x}_i \sum_{j \in \mathcal{N}_i} x_j / \theta))$ and otherwise rejected. For the experiments in this paper we ran $n = 1,000$ chains in parallel, in each case taking the final state at iteration $100,000$. This algorithm was used due to its implementational simplicity, rather than its efficiency, and we note that more sophisticated Markov chain Monte Carlo algorithms are available (e.g. Elçi et al., 2018).

E.2.2 Settings for KSD-Bayes

The same choice of kernel as Appendix E.1.1 is used.

E.2.3 Markov Chain Monte Carlo

The same Metropolis–Hasting algorithm as Appendix E.1.2 was used, in this case in dimension $p = 1$ with proposal standard deviation $\sigma = 0.1$. The convergence of the Markov chain was again diagnosed using univariate Gelman–Rubin statistics computed from 10 independent chains. In total, 100 samples were obtained after thinning from 2000 samples, with an initial burn-in of length 2000. In all cases, the univariate Gelman–Rubin statistics were below 1.002.

E.3 Multivariate Count Data

E.3.1 Description of the Dataset

The original data were gathered by the Cancer Genome Atlas Program, run by the National Cancer Institute in the United States, who have built large-scale genomic profiles of cancer patients with the aim to discover the genetic substructures of cancer (Wan et al., 2015). It contains molecular profiles of biological samples of more than 30 cancer types e.g. measured via RNA sequencing technology. The raw data were pre-processed using the TCGA2STAT software developed by Wan et al. (2015). Inouye et al. (2017) studied a subset of these data relevant to breast cancer, consisting of a total count of each gene profile found in biological samples. They applied a "log-count" transform, a common preprocessing technique for RNA sequencing data, for every datum, that is a floor function of a log transformed value of the datum. Gene profiles were then sorted by variance of the counts in descending order, with the top 10 gene profiles constituting the final dataset. The preprocessed data studied in Inouye et al. (2017) can be found in `https://github.com/davidinouye/sqr-graphical-models`.

E.3.2 Markov Chain Monte Carlo

The Metropolis-Hasting Markov Chain Monte Carlo was applied for this experiment. The detail for the Conway–Maxwell–Poisson graphical model is described first as the Poisson graphical model is the special case. For computational convenience, we work with the square of the interaction and dispersion parameters, i.e. $\tilde{\theta}_{i,j} := \theta_{i,j}^2$ and $\tilde{\theta}_{0,i} = \theta_{0,i}$, which modify the model as

$$p_\theta(\boldsymbol{x}) \propto \exp\left(\sum_{i=1}^{d} \theta_i x_i - \sum_{i=1}^{d} \sum_{j \in \mathcal{M}_i} \tilde{\theta}_{i,j}^2 x_i x_j - \sum_{i=1}^{d} \tilde{\theta}_{0,i}^2 \log(x_i!)\right)$$

The domain of each original parameter $\theta_{i,j}$ and $\theta_{0,j}$ is $[0, \infty)$. With this modification, $\tilde{\theta}_{i,j}$ and $\tilde{\theta}_{0,i}$ can be extended to $\mathbb{R}$, making the model $p_\theta(\boldsymbol{x})$ differentiable with respect to $\theta \in \mathbb{R}^p$. The derivatives of the corresponding DFD-Bayes posterior is then available to implement an efficient gradient-based Markov chain Monte Carlo method. We place a standard normal distribution as a prior on each θ_i, a normal distribution with mean 0 and scale $(d(d-1)/2)^{-1}$ as a prior on each $\tilde{\theta}_{i,j}$, and a standard normal distribution as a prior on each $\tilde{\theta}_{0,i}$, that corresponds to the original priors of each θ_i, $\theta_{i,j}$, and $\theta_{0,j}$. The small scale of the half normal distribution prior on $\tilde{\theta}_{i,j}$ was chosen to suppress rapid increase of the quadratic term $x_i x_j$ as opposed to the linear term x_i in the first summation. After the Markov chain finished, the absolute value was taken for the sampled values of $\tilde{\theta}_{i,j}$ and $\tilde{\theta}_{0,i}$ to convert them as the original parameters $\theta_{i,j}$ and $\theta_{0,j}$. The same setting is applied for the Poisson graphical model by fixing the dispersion parameter $\tilde{\theta}_{0,i} = \theta_{0,i} = 1$.

A No-U-Turn Sampler was used to approximate the DFD-Bayes posterior of both the models. In total, 100 points were obtained thinning from $5,000$ samples, with an initial burn-in of length $5,000$. The posterior predictive of each model $p_\theta(\boldsymbol{x})$ was computed by generating $500,000$ samples from $p_\theta(\boldsymbol{x})$ at every θ sampled from the DFD-Bayes posterior. Each $500,000$ predictive samples were thinned to 878 points to make it comparable with the original data of $n = 878$. The number of bootstrap minimisers B used to calibrate β for this experiment was $B = 100$.