On Margins and Derandomisation in PAC-Bayes

Felix Biggs
Centre for Artificial Intelligence
Department of Computer Science
University College London
United Kingdom

Benjamin Guedj
Centre for Artifical Intelligence
Department of Computer Science
University College London and Inria
United Kingdom and France

Abstract

We give a general recipe for derandomising PAC-Bayesian bounds using margins, with the critical ingredient being that our randomised predictions concentrate around some value. The tools we develop straightforwardly lead to margin bounds for various classifiers, including linear prediction—a class that includes boosting and the support vector machine—single-hidden-layer neural networks with an unusual erf activation function, and deep ReLU networks. Further, we extend to partially-derandomised predictors where only some of the randomness is removed, letting us extend bounds to cases where the concentration properties of our predictors are otherwise poor.

1 INTRODUCTION

PAC-Bayesian[1] generalisation bounds have recently seen a resurgence of interest after the comparative successes of a series of papers applying them to deep neural networks, beginning with Dziugaite and Roy (2017, 2018); Neyshabur et al. (2018), and Letarte et al. (2019); Zhou et al. (2019). One can use these to understand where to apply techniques and motivate new learning algorithms (as for example Foret et al., 2021), as well as provide certification for a given predictor and address the more ambitious goal of understanding generalisation.

One particularly useful aspect of PAC-Bayesian results compared to standard PAC/VC results is that they are non-uniform: the tightness of the guarantee on the generalisation error depends on the specific predictor chosen, not merely on its performance on the training set. This is necessary in cases where our broad class can easily overfit—as for example with many neural networks architectures, which were shown by Zhang et al. (2021) to be able to fit random training labels—since any guarantee must then selectively favour predictors which are reasonable given real data. If our strategic choice of learner turns out to be a good match in practice to the data-generating distribution, the bound should reflect this.

But how to measure this match based only on the training data? A common approach is that we should not just take into account the train error of a given predictor but also its *confidence*. One way to formalise this is the concept of a margin, introduced to bound the error of the perceptron (Novikoff, 1962) and later used to motivate the support vector machine (Cortes and Vapnik, 1995). A confident predictor with a large margin on a given example will be locally robust to parameter (and data) perturbations. If this is true across the dataset our bounds should reflect this and be tighter. From the perspective of Occam's razor this robustness leads to a large set of valid perturbations giving near-equivalent (in terms of dataset outputs) predictors; some predictor in this set is thus likely to be close to a "simple" prediction rule of the kind that we should generally favour (Schapire et al., 1998).

Remarkably, this idea of parameter robustness—as measured by margins—can be formalised through the lens of PAC-Bayes, which more typically bounds the loss of randomised predictors (although there are notable exceptions, see related work in Section 2). After picking a derandomised prediction rule, we can construct a (weighted) class of "proxy" predictors that approximate

[1] PAC-Bayes theory originates in the seminal papers from Shawe-Taylor and Williamson (1997), McAllester (1998) and McAllester (1999), and was further formalised by Catoni (2007), among others – we refer to Guedj (2019) for a recent overview.

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

this rule, with the size and diversity of this class growing with the allowed margin. Since a larger such class is more likely to overlap strongly with our PAC-Bayesian prior, tighter bounds are obtained for larger margins. This idea has been used informally by Langford and Seeger (2001) and Langford and Shawe-Taylor (2003): here we formalise and extend it considerably.

A critical ingredient in this process is the construction of randomised classifiers that have favourable concentration properties—simply, that the parameters are robust to perturbations—so that their deviations from a central derandomised prediction rule are bounded with high probability. The insight that these deviations need only be controlled with high probability rather than certainty is crucial in obtaining better rates and simplifying proofs, and opens the door to the application of powerful concentration of measure results.

We go further and introduce the idea of partially-derandomised predictors, which remove only some of the randomness from the proxy predictors. This enables us to obtain bounds in further cases which would otherwise be difficult to address. We hope that these tools can be further developed to address situations where classical bounding techniques have not worked well (such as in deep neural networks), and that our derived corollaries—such as that for linear prediction—can be used in practice for the provision of self-certifying predictors and model selection.

Contributions and structure. In Section 2 we discuss and formalise the derandomisation of PAC-Bayesian bounds using margins and averaging, introduce the slight generalisation of sub-Gaussian random variables that enables many of our results, and compare to the covering number (or γ-ball) approach. Corollaries of these methods include margin bounds for the following:

Section 3 L_2 and L_1 normed linear prediction; in the L_2 "hard-margin" case this improves on the bound of Bartlett and Shawe-Taylor (1998) and matches the lower bound of Grønlund et al. (2020); the other bounds match the state of the art with simpler proofs.

Section 4 linear predictors with a randomised (and potentially learned) feature space.

Section 5 one-hidden-layer neural networks with erf activations. This involves an interesting new randomised predictor taking the form of a mixture distribution. The introduction of partial-derandomisation also enables bounds with the final two layers derandomised, but the initial layers having randomised weights.

Section 6 deep ReLU networks, a slight improvement on and with proof ideas drawing from that given by Neyshabur et al. (2018).

Finally, in Section 7, we summarise and provide an overview of our results. Related work for specific applications is discussed in the corresponding section, with general work on derandomising bounds in Section 2.

Notation. We will consider classification of i.i.d. examples from a distribution, $\mathcal{D}$, on some product space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, by vector-valued predictors in a function space $\mathcal{H} \subset \widehat{\mathcal{Y}}^{\mathcal{X}}$. For binary classification $\mathcal{Y} = \{+1, -1\}$, $\widehat{\mathcal{Y}} = \mathbb{R}$ and we take the sign of the output as our prediction, while for multi-class prediction, $\mathcal{Y} = [c] := \{1, \ldots, c\}$, $\widehat{\mathcal{Y}} = \mathbb{R}^c$ and the maximum argument is the prediction.

The multi-class margin, $M : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ is the mapping

$$M(f, (x, y)) := f(\boldsymbol{x})[y] - \max_{y' \neq y} f(\boldsymbol{x})[y']$$

where by $f(x)[y]$ we indicate the yth component of $f(x)$. In a slight abuse of notation we also define the binary margin $M(f, (x, y)) := yf(x)$.

We define the margin error $L_\gamma(f) := \mathbb{P}_{z\sim\mathcal{D}}\{M(f, z) \leq \gamma\}$, also writing $L(f) := L_0(f)$ for the misclassification loss or probability of error, and $\hat{L}_\gamma(h) := m^{-1}|\{(x, y) \in S : yh(x) < \gamma\}|$ for the empirical margin error (defined for some sample $S \sim \mathcal{D}^m$ and margin $\gamma \geq 0$). We will also use the abbreviation $\mathbb{P}_{\mathcal{D}}(A) = \mathbb{P}_{z\sim\mathcal{D}}(A)$ and similar for the expectation.

As is common in the PAC-Bayes literature, when considering distributions over predictors $P \in \mathcal{M}_1(\mathcal{H})$ (by which we denote the space of probability measures on $\mathcal{H}$), we write $L(P) := \mathbb{E}_{f\sim P} L(f)$ interchangeably with the above, and analogously with other margin errors[2].

2 PAC-BAYES FOR APPROXIMATIONS

In Section 2.1 we discuss a measure of difference when substituting or "approximating" one distribution over prediction functions by another through a coupling method, and from this derive (partially) derandomised PAC-Bayesian margin bounds in Section 2.2. In Section 2.3 we discuss a powerful concentration-based method for constructing approximating distributions, and in Section 2.4 relate our techniques to covering number methods.

Technical overview. PAC-Bayesian analysis makes a different set of assumptions to Bayesian analysis; the

[2]When considering functions in $\mathcal{H}$ parameterised by some space Θ (of which $\mathcal{H}$ may be a quotient), we will be somewhat loose with interchanging $\mathcal{M}_1(\mathcal{H})$ with $\mathcal{M}_1(\Theta)$; this will not affect any results in practice, as the Kullback-Liebler divergence between distributions on the Θ upper bounds that of their image distributions on $\mathcal{H}$.

guarantees obtained are simultaneously true for any data-generating distribution and PAC-Bayes prior, P_0. This prior serves as a reference measure for the PAC-Bayes posterior P, in a Kullback-Liebler divergence complexity term, $\mathrm{KL}(P, P_0)$. Poor choice of P_0, or an overly-concentrated P will lead to valid but trivial bounds.

If we wish to use a predictive distribution Q concentrated on a single predictor h, we must replace the loss terms by approximating them by a class, $\mathcal{P}$, of randomised "proxy" functions, each of which has margins differing by less than $\gamma > 0$ with high probability $\geq 1-\epsilon$ from h. The PAC-Bayes bounds then obtained can use the minimising proxy from the prior, $\kappa = \min_{P\in\mathcal{P}} \mathrm{KL}(P, P_0)$, leading to bounds of the overall form: with probability greater than $1-\delta$,

$$L(Q) \leq O\left(\hat{L}_\gamma(Q) + \epsilon + \frac{\kappa(Q, P_0; \epsilon, \gamma) + \log(1/\delta)}{m}\right).$$

This ϵ high probability term (which we will usually set $\epsilon \in O(m^{-1})$) is a crucial difference from the covering number approach, and considerably simplifies proofs, since for any proxy $P \in \mathcal{P}$ we only need concentration-of-measure (in the margins) around h. A second innovation is to leave open the possibility of using a partially-derandomised predictive distribution, Q, through a coupling method.

Related work. Work on derandomised PAC-Bayesian bounds is as old as the field; particularly when relating to the average prediction. Bounds holding with high probability over a sampled predictor (directly drawn from the PAC-Bayes posterior) appear *e.g.* in Catoni (2007), Alquier and Biau (2013); Guedj and Alquier (2013).

2.1 Approximation of predictive distributions

Let $P, Q \in \mathcal{M}_1(\mathcal{H})$ be distributions on prediction functions in $\mathcal{H}$, as generally considered in PAC-Bayes. We denote by $\Pi(P, Q) \subset \mathcal{M}_1(\mathcal{H}\times\mathcal{H})$ the set of product distributions with marginals P and Q (also known as couplings between P and Q). For each of these distributions, the margins of these functions are sets of real variables indexed by $\mathcal{Z}$ (and can equivalently be viewed as real-valued stochastic processes on $\mathcal{Z}$).

We define the upper γ-approximate variation of these margins (defined as a relaxation of the total variation distance) using the relative margin at z, $d_z(f,g) = M(f,z) - M(g,z)$ as (abbreviating $\mathbb{P}_{(f,g)\sim\pi}$ as $\mathbb{P}_\pi$ here and henceforth, with a similar notation for the expectation),

$$\mathrm{UAV}_\gamma(P, Q) := \inf_{\pi\in\Pi(P,Q)} \sup_{z\in\mathcal{Z}} \mathbb{P}_\pi \left\{d_z(f,g) > \gamma/2\right\}.$$

In the case of the binary margin, $M(f,(x,y)) = yf(x)$ with $y \in \{+1,-1\}$, so the above is symmetric under interchange of P and Q; this is not true in general. We therefore define the symmetrised version, the γ-approximate variation on $\mathcal{Z}$, as

$$\mathrm{AV}_\gamma(P, Q) := \max(\mathrm{UAV}_\gamma(P, Q), \mathrm{UAV}_\gamma(Q, P)).$$

We say P and Q (γ,ϵ)-approximate each other on $\mathcal{Z}$ if $\mathrm{AV}_\gamma(P, Q) \leq \epsilon$. This is a margin-based generalisation of the total variation distance: the zero-margin approximate variation, $\mathrm{AV}_0(P, Q)$, is equal to the total variation distance between the distributions on margins, $\delta_{\mathrm{TV}}(M(P,z), M(Q,z)) \leq \delta_{\mathrm{TV}}(P, Q)$. However the total variation distance is too strict to yield non-vacuous bounds in most cases where $P \neq Q$. Approximation implies the possibility of substituting one margin loss for another at the cost of these terms and a margin.

Lemma 1. *Let $\gamma > 0, \epsilon \geq 0$; if $P, Q \in \mathcal{M}_1(\mathcal{H})$ with $AV_\gamma(P, Q) \leq \epsilon$, then for any data distribution $\mathcal{D}$, $L(P) \leq L_{\gamma/2}(Q) + \epsilon$ and $L_{\gamma/2}(Q) \leq L_\gamma(P) + \epsilon$.*

Proof. For any events A, B, $\mathbb{P}(A) \leq \mathbb{P}(B) + \mathbb{P}(\bar{B} \cap A)$; and for any coupling $\pi \in \Pi(P, Q)$ we have

$$\begin{aligned} L(P) &= \mathbb{E}_{f\sim P}\mathbb{P}_{\mathcal{D}}\left\{M(f,z) \leq 0\right\} \\ &\leq \mathbb{E}_{g\sim Q}\mathbb{P}_{\mathcal{D}}\left\{M(g,z) \leq \gamma/2\right\} \\ &\quad + \mathbb{E}_\pi\mathbb{P}_{\mathcal{D}}\left\{M(g,z) > \gamma/2 \wedge M(f,z) \leq 0\right\} \\ &\leq L_{\gamma/2}(Q) + \mathbb{E}_{\mathcal{D}}\mathbb{P}_\pi\left\{d_z(f,g) > \gamma/2\right\}. \end{aligned}$$

Replacing the expectation with its pointwise bound and taking the infimum over couplings, we find that $L(P) \leq L_{\gamma/2}(Q) + \mathrm{UAV}_\gamma(P, Q)$. An analogous process follows for the other side, with the order of Q and P reversed. □

Remark. A less sophisticated analysis could have used the bound $|M(f,z) - M(g,z)| \leq 2\max_{y\in\mathcal{Y}} |f(x)[y] - g(x)[y]|$ instead, leading to similar PAC-Bayes bounds. This definition of AV improves constants in some derived bounds and removes a factor of c, the number of classes. We also note that the coupling need not be the same on both sides of the bounds, although we do not use this in later proofs.

2.2 PAC-Bayes bounds with approximations

Lemma 1 can be used to derive a type of PAC-Bayesian bound for a predictive distribution Q, as follows. First we define the (γ,ϵ)-approximating KL projection onto a prior $P_0 \in \mathcal{M}_1(\mathcal{H})$ (defined independently of the data), of (γ,ϵ)-approximations to Q.

Definition 1. Approximating KL-Projection*: Given some prior distribution P_0 on $\mathcal{H}$, the (γ,ϵ)-approximate*

projection of Q onto P_0 is

$$\kappa(Q, P_0; \gamma, \epsilon) := \min_{P \in \mathcal{P}} \mathrm{KL}(P, P_0)$$

where $\mathcal{P} = \{P \in \mathcal{M}^1(\mathcal{H}),\ \ \mathrm{AV}_\gamma(P, Q) \leq \epsilon\}$.

This can be viewed as the "closest" (in the KL sense) *proxy* P, to our prior that approximates Q sufficiently well. In practice, we will restrict the family of proxies, $\mathcal{P}$, to some more tractable set and construct P explicitly. The notion can be used in combination with many PAC-Bayesian bounds, replacing the usual KL divergence with a new complexity term and the losses with margin losses. We give the following formulations as examples.

Theorem 1. *Given data distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, prior $P_0 \in \mathcal{M}^1(\mathcal{H})$, $\gamma > 0, \epsilon \in [0, \frac{1}{2}]$ and $\delta \in (0, 1)$, the following hold each with probability $\geq 1 - \delta$ over $S \sim D^m$, for all $Q \in \mathcal{M}^1(\mathcal{H})$*

"small-kl"

$$\mathrm{kl}(\hat{L}_\gamma(Q) + \epsilon : L(Q) - \epsilon) \leq \frac{\kappa(Q, P_0; \gamma, \epsilon) + \log \frac{2\sqrt{m}}{\delta}}{m},$$

"interpolating" *(given that $\hat{L}_\gamma(Q) = 0$)*

$$L(Q) \leq \frac{\kappa(Q, P_0; \gamma, \epsilon) + \log \frac{1}{\delta}}{m} + 4\epsilon \log \frac{1}{\epsilon},$$

where $\mathrm{kl}(q : p) = q \log \frac{q}{p} + (1 - q) \log \frac{1-q}{1-p}$ *if* $p \geq q$ *and otherwise* 0 *(this formulation is monotonic in q and thus one-sided).*

Proof. The standard PAC-Bayesian bounds (we refer to Appendix A) with loss function $\ell_{\gamma/2}$ are true for the minimising P in Definition 1. We can then use Lemma 1 to replace the losses with those w.r.t. Q.

In the final step, we use the following formulation of the Catoni bound using Germain et al. (2009, Proposition 2.1), valid if ϵ is independent of the sample S and $\hat{L}_\gamma(Q) = 0$:

$$\mathrm{kl}(\epsilon : L(Q) - \epsilon) \leq \frac{1}{m}\left(\mathrm{KL}\left(P, P_0\right) + \log \frac{1}{\delta}\right)$$

and adapt it in the same way as the previous bounds to κ. We then use the lower bound $\mathrm{kl}(\epsilon : p - \epsilon) \geq p + 4\epsilon \log \epsilon$, valid for all $\epsilon \in [0, \frac{1}{2}], p \in [0, 1]$. This is proved in Lemma 5 in supplementary material. □

2.3 Sub-Gaussian derandomisation

One simple case to which the above bounds can be applied, often in a dimension-independent way, is that of total derandomisation by *averaging*: for some P, we set $Q = \delta(F_P)$, a point mass measure on the P-aggregate function $F_P(x) := \mathbb{E}_{f \sim P} f(x)$. If the score does not vary too much under P, as defined by a sub-Gaussian condition, derandomised PAC-Bayes bounds follow directly through our framework. This is a formalisation of a proof idea, used for example by Langford and Shawe-Taylor (2003); by making this connection explicit we contribute a clearer understanding of PAC-Bayes derived margin bounds.

First we define the idea of sub-Gaussian random functions, defined here in a slightly more general way to accommodate "partial-derandomisation".

Definition 2. *We say a coupling $\pi \in \Pi(P, Q)$ is σ^2-sub-Gaussian on $\mathcal{Z}$ if*

$$\mathbb{E}_\pi \exp(t(f(x)[y] - g(x)[y])) \leq \exp(t^2\sigma^2/2)$$

and $E_{f \sim P} f(x)[y] = E_{g \sim Q} g(x)[y]$, for all $t \in \mathbb{R}$, $(x, y) \in \mathcal{Z}$. The square bracket indicates the yth index if the output is multi-dimensional; in the scalar case we remove it.

We will further stretch this definition and call a *single* distribution, P, σ^2-sub-Gaussian, if the trivial coupling $\pi = P \otimes \delta(F_P)$ is. Sub-Gaussianity implies bounds on the approximate variation:

Lemma 2. *If $\pi \in \Pi(P, Q)$ is σ^2-sub-Gaussian on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, for binary and multi-class classification respectively,*

$$\mathrm{AV}_\gamma(P, Q) \leq \begin{cases} \exp(-\gamma^2/8\sigma^2) & \text{for } \mathcal{Y} = \mathbb{R}, \\ \exp(-\gamma^2/16\sigma^2) & \mathcal{Y} = \mathbb{R}^c. \end{cases}$$

Proof. Considering the zero-mean random variable $X = f(x)[y] - g(x)[y]$ for $(f, g) \sim \pi$ (σ^2-sub-Gaussian) and fixed $(x, y) \in \mathcal{Z}$, the Chernoff bound (see Boucheron et al., 2013, for example, for a thorough introduction to sub-Gaussianity), immediately implies

$$\mathbb{P}(X > t) \vee \mathbb{P}(-X > t) \leq e^{t^2/2\sigma^2}$$

for all $t > 0$. In the binary margin case, $M(f, z) = yf(x)$ which is either $f(x)$ or $-f(x)$; setting $t = \gamma/2$ in the above therefore gives the bound.

In the multi-class case we consider the upper bound obtained by letting y' achieve the maximum margin for g; then $M(f, z) \leq f(x)[y] - f(x)[y']$, so

$$\begin{aligned} &\mathbb{P}_\pi\left\{M(f, z) - M(g, z) > \frac{\gamma}{2}\right\} \\ &\leq \mathbb{P}_\pi\left\{f(x)[y] - f(x)[y'] - g(x)[y] + g(x)[y'] > \frac{\gamma}{2}\right\}. \end{aligned}$$

Since both $f(x)[y] - g(x)[y]$ and $f(x)[y'] - g(x)[y']$ are σ^2-sub-Gaussian, their sum is $2\sigma^2$-sub-Gaussian and the bound follows by repeating the process on with signs reversed. □

Other concentration assumptions. We note that although we only discuss sub-Gaussian concentration, it is possible to require other concentration properties, for example sub-exponential ones; our framework easily accommodates these. Sub-Gaussianity is only the simplest way to ensure such concentration, and we primarily consider it in our later results as it already leads to simple proofs of margin bounds in multiple settings.

2.4 Relation to covering

Here we discuss how our bounds can be used to derive a standard covering approach as a sub-case; we show that this leads to certain problems, which are circumvented by the concentration approach. A further consequence is that covering-based bounds usually lead to "uniform" bounds which are subject to problems discussed in Nagarajan and Kolter (2019). All the bounds we provide in later sections are non-uniform and avoid these pitfalls.

By setting $\epsilon = 0$ with certain choices of prior we can obtain a fairly standard "covering" approach: call N_γ a γ-net of $\mathcal{H}$, if for any $f \in \mathcal{H}$, there exists $g \in N_\gamma$ such that $|M(f,z) - M(g,z)| \leq \gamma$ for all $z \in \mathcal{Z}$. If we choose a prior supported everywhere on a $\gamma/2$-net for $\mathcal{H}$, we can achieve $\mathrm{AV}_\gamma(P,Q) = 0$ for any $Q \in \mathcal{M}_1(\mathcal{H})$, including Q supported on just a single predictor. The simplest approach chooses P_0 as uniform on these points so that

$$\kappa(Q, P_0; \gamma, 0) \leq \log |N_{\gamma/2}|$$

where $|N_{\gamma/2}|$ is the cardinality of the net. A more sophisticated choice of non-uniform prior enables structural risk minimisation-type covering number bounds.

However, such bounds will typically be dependent on the dimension of the parameter space, as demonstrated by the following proposition (proved in Appendix B) for linear classification. Our bounds in the following sections will avoid this dimension-dependence.

Theorem 2. *Consider binary classification with functions $f_w = \langle w, x\rangle$ and $x \in \mathbb{R}^d, \|x\|_2 \leq 1$. For any prior P_0 on weights in $\mathbb{R}^d$, there exists a prediction distribution Q supported on $\|w\|_2 \leq 1$ such that*

$$\kappa(Q, P_0; \gamma, \epsilon = 0) \geq \Omega(d).$$

3 LINEAR PREDICTION BOUND

Here we demonstrate our framework in action by deriving generalisation bounds for linear predictors. These bounds essentially follow from an initial Gaussian assumption combined with the sharp (sub-Gaussian) concentration of the predictor output around its mean. We hope they can be useful for self-certification in the low data regime, and for model (or kernel) selection without a validation set.

L_2-normed linear predictors. This situation has been considered by a large number of papers, from Bartlett and Shawe-Taylor (1998, Theorem 1.7, using Fat-Shattering) in the fast-rate or interpolating case, to Bartlett and Mendelson (2002, Theorem 22, using Rademacher complexity) in the "soft-margin" case. McAllester (2003) presents alternative bounds in the "soft-margin" case, and is itself an attempt to find a expression for the implicit PAC-Bayesian result of Langford and Shawe-Taylor (2003). We give bounds for both cases, through a proof similar to the method of Langford and Shawe-Taylor (2003), but using a different base PAC-Bayesian bound which makes solving the interpolating hard-margin scenario of Bartlett and Shawe-Taylor (1998) more straightforward.

Theorem 3. *In the binary classification setting with $\mathcal{X}$ a Hilbert space with $\|x\|_2 \leq R$, and $\delta \in (0,1)$, with probability $\geq 1-\delta$ over $S \sim D^m$, for all linear predictors $F_w(x) = \langle w, x\rangle$ with $\|w\|_2 \leq 1$ and all $\gamma > 0$ ("soft-margin"),*

$$L(F_{\boldsymbol{w}}) \leq \hat{L}_\gamma(F_{\boldsymbol{w}}) + \sqrt{\frac{\hat{L}_\gamma(F_{\boldsymbol{w}}) \cdot \Delta}{m}} + \frac{\Delta + \sqrt{\Delta} + 2}{m},$$

where we define $\Delta := 2\log(2/\delta) + 9(R/\gamma)^2 \log m$. Additionally, under the same conditions and probability, provided $\gamma_\star = \max\{\gamma > 0 : \hat{L}_\gamma(F_w) = 0\}$ exists ("hard-margin"),

$$L(F_w) \leq \frac{8(R/\gamma_\star)^2 \log m + \log(1/\delta)}{m}.$$

Proof. Without loss of generality assume $R = 1$. To consider a free choice of margin γ, we note the scaling property $\mathbf{1}\{M(F_{\boldsymbol{w}}, z) \leq \gamma\} = \mathbf{1}\{M(F_{(\theta/\gamma)\boldsymbol{w}}, z) \leq \theta\}$. This suggests approximating the mean predictor $F_{(\theta/\gamma)\boldsymbol{w}}$ by the distribution P over functions $f = \langle \boldsymbol{u}, \boldsymbol{x}\rangle$ for $\boldsymbol{u} \sim \mathcal{N}((\theta/\gamma)\boldsymbol{w}, \boldsymbol{I})$. Choosing a data-free prior P_0 of a similar form, but with $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ gives a divergence $\mathrm{KL}(P, P_0) = \frac{1}{2}\|(\theta/\gamma)\boldsymbol{w}\|^2 = \theta^2/2\gamma^2$.

P is 1-sub-Gaussian, so by Lemma 2, $\mathrm{AV}_\theta(P, \delta(F_{(\theta/\gamma)\boldsymbol{w}})) \leq \exp(-\theta^2/8) = \epsilon$. Plugging into (the hard-margin) Theorem 1 we obtain for a fixed $\theta^2 = 8\log m$ and all $\gamma_\star$ such that $\hat{L}_{\gamma_\star}(F_{\boldsymbol{w}}) = 0$,

$$\begin{aligned} L(F_{\boldsymbol{w}}) &\leq \frac{\theta^2/2\gamma_\star^2 + \log\frac{1}{\delta}}{m} + \frac{1}{2}\theta^2 \exp(-\theta^2/8) \\ &\leq \frac{4(1 + 1/\gamma_\star^2)\log m + \log\frac{1}{\delta}}{m}. \end{aligned}$$

By the assumptions on $\|\boldsymbol{w}\|_2$ and R, we have $\gamma_\star \leq 1$ to prove the second statement.

Repeating the above but replacing the use of the hard-margin bound with the small-kl formulation in Theorem 1 (and using the vacuity of the bound when $\gamma > 1$), we have the tight bound

$$\begin{aligned}&\mathrm{kl}(\hat{L}_\gamma + m^{-1} : L - m^{-1}) \\ &\leq \frac{1}{m}\left(4(R/\gamma)^2 \log m + \frac{1}{2}\log m + \log\frac{2}{\delta}\right) \leq \frac{\Delta}{2m} \quad (1)\end{aligned}$$

with probability $\geq 1-\delta$. To relax the above we use the lower bound $\mathrm{kl}(q:p) \geq (p-q)^2/(2p)$ for $p > q$ from McAllester (2003) to show $L \leq \hat{L}_\gamma + 2m^{-1} + \sqrt{(\hat{L}_\gamma + m^{-1}) \cdot \Delta/m} + \Delta/m$ which completes the proof. □

In the "hard margin" case Theorem 3 improves on Bartlett and Shawe-Taylor (1998) by a factor of $O(\log m)$, matching the lower bound of Grønlund et al. (2020, Theorem 4). In the "soft-margin" case Theorem 3 is of the same order as the state-of-the-art bound given by Grønlund et al. (2020) but with explicitly stated constants. We emphasise the extreme simplicity of our proof compared to that given for Theorem 2 in Grønlund et al. (2020) (in Section 2, p.3-7), and that these are the tightest explicitly-stated bounds for the problem to our knowledge.

Our soft-margin result in Theorem 3 also improves upon McAllester (2003), which is proved via a similar method, but using a different base PAC-Bayes bound that leads to a weaker result. We present our soft-margin result in the more straightforward form given by the papers above so these comparisons can be easily made, even though Equation (1) is technically tighter. We discuss these differences at length and give full forms of existing results in Appendix C.

We note that the soft-margin formulation of the bound is true universally across $\gamma > 0$, allowing the bound can be optimised for γ in $O(m)$ time. If the margin is large for most examples, we can choose γ so that $\hat{L}_\gamma$ is small and thus the Δ/m term (which is of the same order as the hard-margin bound) dominates. Since the minimum margin can be sensitive to outliers, this bound will often be tighter than the hard-margin one.

The margin only appears in the bounds of Theorem 3 in a "normalised" form, γ/R, otherwise scaling the data would affect the bound. However, we note that the bound can sometimes be decreased by normalising the data (as this maximises the margin for every data point), so we recommend this when using such predictors.

Finally we also acknowledge the result of Hanneke and Kontorovich (2021, Theorem 1) which gives a *algorithm*-dependent hard-margin bound specifically for the SVM output, and eliminates the $\log m$ factor. This is provably optimal in the algorithm-dependent (as ours is in the general) case, which is shown in Grønlund et al. (2020, Theorem 5).

L_1/L_∞-normed linear predictors. Theorem 3 is a bound under L_2 norms for $\mathcal{X}$ and w, applying to situations such as the SVM. For completeness we provide in Appendix D a bound for linear classification under different norm constraints, where the L_1 norm of the weights and L_∞ norm of the features is restricted, as in boosting.

These results are essentially the same as the k-th margin bound of Gao and Zhou (2013), or the central result of Langford and Seeger (2001), but proved through our framework instead, which we believe provides a unifying perspective. The fundamental proof idea is to approximate our predictor by a randomised, unweighted, sum of features, as originally proposed by Schapire et al. (1998); the boundedness of these features leads to sub-Gaussian concentration around their mean, similarly to in Theorem 3.

4 PARTIAL DERANDOMISATION

Bounds of a similar form to Theorem 3 can also be used in another interesting situation: where before linear prediction we apply a feature map, as commonly done in the SVM. If $\phi \in \Phi \subset \{f : \mathcal{X} \to \mathcal{X}^\dagger\}$ (so that $\mathcal{X}^\dagger$ is a Hilbert space and $\mathcal{X}$ an arbitrary set) is the map, our predictor is of the form $\langle w, \phi(x)\rangle$. Theorem 3 then applies with only the modification that R is a bound on $\|\phi(x)\|_2$ instead of $\|x\|_2$.

In certain cases we may wish to *learn* these (perhaps randomised) features in parallel with w. In this case the usual PAC-Bayesian analysis would generally fail without making both w and the map ϕ random. The generality of coupling and approximations as outlined in Section 2 here comes to the fore; we can "partially derandomise" or derandomise w while ϕ is still randomised.

More formally, let $Q^\Phi \in \mathcal{M}_1(\Phi)$ be a probability measure on feature maps so that the posterior Q is a distribution on functions of the form $f(\phi(x))$ for $\phi \sim Q^\Phi$ and deterministic $f : \mathcal{X}^\dagger \to \mathcal{Y}$. The approximating P distribution can then take the form $g(\phi'(x))$ for $g \sim P^g$ and $\phi' \sim Q^\Phi$, the same random feature map. Provided the P^g and Q^Φ distributions are independent, the KL divergence from prior P_0 separates into terms like $\mathrm{KL}(P, P_0) = \mathrm{KL}(P^g, P_0^g) + \mathrm{KL}(Q^\Phi, P_0^\Phi)$. Using this fact and that such mappings do not affect the sub-Gaussianity of our predictors, we obtain the following results, analagous to Theorem 3, but applicable under learned and potentially randomised feature maps.

Lemma 3. *There is a 1-sub-Gaussian coupling between functions defined by $h(x) = \langle w, \phi(x)\rangle$ and $h'(x) = \langle w+g, \phi'(x)\rangle$ where $g \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and $\phi, \phi' \sim Q^\Phi$ independent of g, provided $\|\phi\|_2 \leq 1$ almost surely $[Q^\Phi]$.*

Proof. We use a coupling π such that $\phi = \phi'$, so that

$$\begin{aligned}\mathbb{E}_{(h,h')\sim\pi} &\exp(t(h(x) - h'(x))) \\ &= \mathbb{E}_g \mathbb{E}_{\phi\sim Q^\Phi} \exp(t(\langle w, \phi\rangle - \langle w+g, \phi\rangle)) \\ &= \mathbb{E}_{\phi\sim Q^\Phi} \exp(t^2\|\phi\|_2^2/2) \leq \exp(t^2/2)\end{aligned}$$

where we use the moment generating function of a standard multivariate Gaussian. □

Theorem 4. *In the binary classification setting, let Φ be a space of bounded functions $\phi : \mathcal{X} \to \mathcal{X}^\dagger$ where $\mathcal{X}^\dagger$ is a Hilbert space with $\|\phi\|_2 \leq 1$ everywhere. For any prior $P_0^\Phi \in \mathcal{M}_1(\Phi)$ and $\delta \in (0,1)$, with probability $\geq 1-\delta$ over $S \sim D^m$, for all prediction distributions Q of the form $f(x) = \langle w, \phi(x)\rangle$ with $\|w\|_2 \leq 1, \phi \sim Q^\Phi \in \mathcal{M}_1(\Phi)$,*

$$L(Q) \leq \hat{L}_\gamma(Q) + \sqrt{\frac{\hat{L}_\gamma(F_{\boldsymbol{w}}) \cdot \Delta}{m}} + \frac{\Delta + \sqrt{\Delta} + 2}{m}$$

where $\Delta := 2\log(2/\delta) + 9(R/\gamma)^2 \log m + 2\,\mathrm{KL}\left(Q^\Phi, P_0^\Phi\right)$.

(Sketch of proof). Use Lemma 3 in the proof of the second part of Theorem 3 to obtain 1-sub-Gaussianity, adding the extra KL contribution from the feature map. □

Such a bound (with L_1 norm restrictions) could be used to derandomise the final layer of neural networks with a bounded (e.g. tanh) activation functions on the penultimate layer and randomised weights on the rest of the structure. In the next section we will take this approach further and derandomise the final *two* layers through margins, which can be straightforwardly used to obtain a bound on one-hidden-layer networks. In conjunction with the above ideas it yields bounds for deep stochastic networks with the final two layers derandomised.

5 AVERAGING ONE-HIDDEN-LAYER NETWORKS

In this section we prove generalisation bounds for a one-hidden-layer neural network (possibly with a randomised input feature map) with a slightly unusual erf activation function that looks much like a tanh or other sigmoidal-type function as more commonly used. This is inspired by the work of Germain et al. (2009) and Letarte et al. (2019), which consider averaging over the predictions of functions like $f_w : \mathbb{R}^d \to \mathbb{R}, x \mapsto \mathrm{sign}(w \cdot x)$, where $w \sim \mathcal{N}(u, I)$, giving "aggregated" prediction functions of the form

$$F(x) = \mathbb{E}_{w\sim\mathcal{N}(u,I)} \mathrm{sign}(w \cdot x) = \mathrm{erf}(u \cdot x/\sqrt{2}\|x\|_2). \quad (2)$$

With a clever choice of weight distribution, we can combine the sub-Gaussian concentration of the bounded sign function with its tractable average to get bounds for one-hidden layer networks.

Definition 3. Single Hidden Erf Layer (SHEL) Network*: Given $V \in \mathbb{R}^{c\times K}$ and $U \in \mathbb{R}^{K\times d}$, this is the neural network $F : d \to c$ defined by*

$$F_{U,V}(x) = V\,\mathrm{erf}\left(\frac{Ux}{\sqrt{2}\|x\|_2}\right) \quad (3)$$

where the erf *activation function is applied elementwise. We also consider the "binary" case, where V is a vector, $v \in \mathbb{R}^K$.*

The generalisation bound for this depends on a set of prior parameters (or "random features"), U_0, chosen independently of the training data, for example the initialisation of the network (this choice has been extensively discussed in the literature, beginning with Dziugaite and Roy, 2017).

Theorem 5. *Fix prior parameters $U^0 \in \mathbb{R}^{K\times d}$ and $\delta \in (0,1)$. With probability $\geq 1-\delta$ over $S \sim \mathcal{D}^m$, $L(F_{U,V})$ is upper bounded by*

$$\hat{L}_\gamma(F_{U,V}) + \tilde{O}\left(\frac{\sqrt{K}}{\gamma\sqrt{m}}\left(V_\infty\|U - U^0\|_F + \|V\|_F\right)\right),$$

for any margin $\gamma > 0$ and any prediction function $F_{U,V}$ specified as in Equation (3) with parameters U, V, and $V_\infty := \max_{ij}|V_{ij}|$. A full (tighter) expression with constants is given in the proof in Appendix E.

Remark. At first glance this bound might appear to grow with width, since although the norm terms are usually seen to be roughly constant under increasing K, the $\sqrt{K}$ term is obviously not. However, this is not necessarily true: the range of the network (and thus maximum margin) is bounded by KV_∞, so provided the margin per-unit (γ/K for the γ used in the bound) remains constant, the bound would actually decrease with K.

To emphasise this, we note that the above bound is unchanged under two simple transformations, which ensures dimensional consistency (if it were not, we could perform these operations to obtain a possibly arbitrarily tight bound). (1) Scale V; the bound and norm term exactly cancel since we can scale γ by the

same amount and obtain the same empirical margin loss. (2) Double the width of network, with exact copies of weights in the copy: we can again double γ for a fixed margin loss, while the squared norms also double.

Proof outline. The central idea underlying the proof is the construction of a stochastic neural network with Equation (3) as its average. We replace the normal distribution of Equation (2) with a *mixture* of isotropic Gaussians: if the mixture weights are uniform and their means are given by the columns of U (notated as the set $\{U_{1;\cdot}, \ldots, U_{K;\cdot}\}$), we note that

$$\mathbb{E}_{i\sim\mathrm{Unif}(K),\, w\sim\mathcal{N}(U_{i;\cdot},I)} \operatorname{sign}(w\cdot x) = \sum_{k=1}^{K} \operatorname{erf}\left(\frac{U_{k;\cdot}\cdot x}{\sqrt{2}\|x\|_2}\right) \quad (4)$$

which is directly proportional to one of the output components of the SHEL network F. To obtain the final layer weights we multiply the sign by a random vector $\boldsymbol{r}$ supported on $\{+1,-1\}^c$ and re-scale everything to fit the scale of the SHEL network.

We note that the function $f(x) = \operatorname{sign}(w\cdot x)$ is also sub-Gaussian (for any distribution on w) as it is a bounded random variable for any fixed z. To obtain control over the constant and thus ϵ, we average over a number of copies of the network, an approach inspired by the approach of Schapire et al. (1998) or Langford and Seeger (2001), but for a hidden-layer network. Combination with Theorem 1, careful bounding of the KL divergence of such hierarchical distributions, and a union bound over margin values completes the proof.

Generalisation to bounded functions. We note that in the proof of Theorem 5 we can replace the sign activation functions used in the proxy function distribution by any bounded activations, for example sigmoid. Indeed, any feature map which is bounded and independent from the final layer is possible. The caveat is that the obtained networks have modified activation functions which may not be analytically tractable.

Partial derandomisation. A more straightforward extension to deep networks follows through the partial derandomisation framework discussed in Section 4; the boundedness of the activation then means the theorem and proof hold with only slight modification. A simple way to do this is to "stack" our SHEL network on top of a ReLU network with Gaussian weights, adding only a small KL contribution to the bound; this is discussed further in Appendix G.

This is interesting because it enables empirical comparisons with deeper networks on more complex datasets without severe overfitting, which we hope can form a stepping stone between totally-randomised PAC-Bayesian bounds and non-random margin bounds, while helping in the understanding of one-hidden-layer network generalisation. This provides a middle ground between a series of works obtaining bounds for stochastic neural networks such as Dziugaite and Roy (2017), and those providing margin bounds for non-stochastic DNNs, such as (in a PAC-Bayesian context) Neyshabur et al. (2018).

Empirical evaluation. Although the main contribution of this paper is in the refinement of methods for proving PAC-Bayes margin bounds, in Appendix G we also make some empirical evaluations of Theorem 5, and a partially derandomised generalisation of it. Since these bounds were in general vacuous, we adopt the procedures of Jiang et al. (2020) and Dziugaite et al. (2020) to compare such bounds; training to a fixed cross-entropy of 0.3 and setting margin loss $\hat{L}_\gamma(F) = 0.2$, we examine changes in the big-O complexity measure in Theorem 5 versus generalisation error under different hyperparameter changes. Our complexity measure is predictive under training set size changes and somewhat predictive under learning rate changes, but like most such measures (Dziugaite et al., 2020), it is not predictive under changes of width, implying the per-unit margin decreases significantly with width. We interpret this as follows: at initialisation $u_i\cdot x \sim d^{-\frac{1}{2}}$ is small, so if weights stay near their initialisation (as is usual for wider networks trained by SGD), units are less saturated and the per-unit margin decreases. This is avoided in lower dimensions or by scaling up the weight initialisations with d, but as this is further from the typical SGD training scenario we avoid this.

Optimisation of the prior. We have in the above empirical evaluation neglected to utilise optimised data-dependent priors (as initiated by Ambroladze et al., 2006; Parrado-Hernández et al., 2012), which has been demonstrated to vastly tighten bounds in the case of neural networks due to the stability of training. These ideas have been heavily used in recent papers for neural networks (Dziugaite et al., 2021, for example) and were found to significantly improve the actual bound values in preliminary experiments, in some cases leading to non-vacuous (although loose) results. As our focus is more on the theoretical side of providing a method to prove margin bounds, we decided to focus on the data-independent case for simplicity.

Related Work. Here we mention previous work (Letarte et al., 2019; Biggs and Guedj, 2021) on PAC-Bayesian neural networks with erf activations, as well as a wide range of results obtaining generalisation bounds for neural networks, in particular Neyshabur et al. (2019) which focuses specifically on one-hidden-layer networks. Banerjee et al. (2020) uses similar methods to ours by looking at Gaussian perturbations to the weights of a deep ReLU network, but their bound relies

on the strong assumption of bounds on the Hessian and gradients of the network across weight values, and as formulated is not evaluable nor decreases with m.

We also highlight an interesting connection to a strand of work (Kristiadi et al., 2020; Daxberger et al., 2021) in the Bayesian neural network literature, where networks involving only some randomised weights (effectively, partially-derandomised networks) were found to offer many of the benefits of more general networks while offering considerable computational saving.

6 BEYOND TWO LAYERS

Finally, we give a bound for deep feed-forward ReLU networks, similar in form and proof to that given by Neyshabur et al. (2018). Although the new result shares the same shortcomings (as discussed in, for example, Dziugaite et al., 2020), we hope our simplified proof and unifying perspective will help clear the way for future improvements.

The new bound also replaces a factor of d, the number of layers, with one of $\sqrt{\log m}$, while the proof is simplified by merely requiring $\mathrm{AV}_\gamma \in O(m^{-1})$ rather than $\mathrm{AV}_\gamma = 0$ as in the original. Bounding this term for simple Gaussian weights with the same perturbation bound as their proof, gives a simple form for KL divergence. Combination with Theorem 1 and a cover of different weight variances and margins completes the proof, given in Appendix F.

Theorem 6. *Let $F : \mathcal{X} \to \mathbb{R}^c$ on $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_2 \leq R\}$ be a fully-connected, feed-forward ReLU neural network with d layers and no more than h units per layer. For fixed $\delta \in (0,1), W_\star > 0$ and prior weight matrices $\{W_i^0\}_{i=1}^d$, with probability at least $1-\delta$ for all such networks F with weight spectral norms $\|W_i\|_2 \leq W_\star$ for all i, and $\theta > 0$, $L(F)$ is upper bounded by*

$$\hat{L}_\theta(F) + O\left(\sqrt{\frac{hr^2 \log(mdh)}{\theta^2 m} \cdot \sum_{i=1}^{d} \frac{\|W_i - W_i^0\|_F^2}{\|W_i\|_2^2} + \frac{A}{m}}\right)$$

where $r := R\prod_{i=1}^d \|W_i\|_2$ is an upper bound on the range of network, and $A := \log \frac{1}{\delta} + d \log\log W_\star$.

Remark. A second difference between Theorem 6 and the bound of Neyshabur et al. (2018) is the appearance of the prior matrices (to bring the bound into line with others which often set these to the initialisation) and the norm bound $W_\star$. This $W_\star$ term arises from these prior matrices and can be eliminated if the prior matrices are set to zero, since re-scaling the weights and margins will then not affect the bound (due to the positive homogeneity of the ReLU, $\|W_i\|_2/\theta = \|\tilde{W}_i\|_2/\tilde{\theta}$ and $\|W_i\|_F/\|W\|_2 = \|\tilde{W}_i\|_F/\|\tilde{W}\|_2$ for re-scaled $\tilde{W}_i$ and $\tilde{\theta}$).

7 CONCLUSION

In this work we have provided a unified framework for derandomising PAC-Bayes bounds using margins. In particular this leads to new bounds or greatly simplified proofs for a variety of settings. It also enables the novel idea of partial-derandomisation, which provides a halfway house for estimators which cannot be so easily derandomised.

Specifically: we provided in Theorem 3 bounds for L_2-regularised linear classification which improve upon classical results and match the state-of-the-art order given in Grønlund et al. (2020) while providing explicit (and small) constants as well as a considerably simplified proof. In Section 4 we extended this result to the novel situation where we are simultaneously learning a (randomised) feature map. We then gave further bounds in Theorem 5 for the novel setting of single-hidden-erf-layer (SHEL) networks, as well as a bound in Theorem 6 that improves slightly on Neyshabur et al. (2018, Theorem 1). We feel that SHEL networks have much potential as a setting to explore margin bounds for deterministic neural networks—matching contemporary practice—through improved concentration techniques and priors.

Although we recognise that our final results for linear prediction and deep ReLU networks are relatively small improvements on existing results, we believe our radically simplified proofs and explicit link of derandomisation to concentration (a link which has been occasionally used implicitly in proofs in the literature) are significant and novel contributions to a difficult and central problem in their own right. We show in Section 2.4 how reducing PAC-Bayes derandomisation to a covering approach leads to a sub-optimal dependence on the dimension, which is observed in some prior results such as the prior ReLU bound (Neyshabur et al., 2018). We believe that by highlighting this issue we point the way forward to further simplifications and improvements, and hope the machine learning and statistics community will leverage these tools going forward.

Similarly, we feel that a major implication of our work is to show that for non-vacuous neural network margin bounds, we need tighter bounds on the concentration properties of networks. Networks are observed to be quite robust to perturbation in practice, far better than the Lipschitz constant-dependent bounds of our Theorem 6 and Neyshabur et al. (2019) would suggest. Tighter concentration bounds would immediately lead to improved margin bounds through our framework and would represent a major contribution to contemporary statistical learning theory.

Acknowledgements

We warmly thank all reviewers and the area chair for their insightful feedback which improved the quality of our presentation. Felix Biggs gratefully acknowledges the support of the CDT for Foundational Artificial Intelligence through UKRI grant EP/S021566/1. Benjamin Guedj acknowledges partial support by the U.S. Army Research Laboratory and the U.S. Army Research Office, and by the U.K. Ministry of Defence and the U.K. Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/R013616/1; Benjamin Guedj also acknowledges partial support from the French National Agency for Research, grants ANR-18-CE40-0016-01 and ANR- 18-CE23-0015-02.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

Pierre Alquier and Gérard Biau. Sparse single-index model. *Journal of Machine Learning Research*, 14: 243–280, 2013.

Amiran Ambroladze, Emilio Parrado-Hernández, and John Shawe-Taylor. Tighter pac-bayes bounds. In Bernhard Schölkopf, John C. Platt, and Thomas Hofmann, editors, *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 9–16. MIT Press, 2006. URL https://proceedings.neurips.cc/paper/2006/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Arindam Banerjee, Tiancong Chen, and Yingxue Zhou. De-randomized pac-bayes margin bounds: Applications to non-convex and non-smooth predictors. *CoRR*, abs/2002.09956, 2020. URL https://arxiv.org/abs/2002.09956.

Peter Bartlett and John Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. In Bernard Schölkopf, Christopher J C Burges, and Alexander J Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, Cambridge, USA, 1998.

Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Felix Biggs and Benjamin Guedj. Differentiable pac–bayes objectives with partially aggregated neural networks. *Entropy*, 23(10), 2021. ISSN 1099-4300. doi: 10.3390/e23101280. URL https://www.mdpi.com/1099-4300/23/10/1280.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. ISBN 978-0-19-953525-5. doi: 10.1093/acprof:oso/9780199535255.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199535255.001.0001.

Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Institute of Mathematical Statistics lecture notes-monograph series. Institute of Mathematical Statistics, 2007. ISBN 9780940600720. URL https://books.google.fr/books?id=acnaAAAAMAAJ.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, 1995. doi: 10.1007/BF00994018. URL https://doi.org/10.1007/BF00994018.

Erik A. Daxberger, Eric T. Nalisnick, James Urquhart Allingham, Javier Antorán, and José Miguel Hernández-Lobato. Bayesian deep learning via subnetwork inference. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 2510–2521. PMLR, 2021. URL http://proceedings.mlr.press/v139/daxberger21a.html.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *Conference on Uncertainty in Artificial Intelligence 33.*, 2017.

Gintare Karolina Dziugaite and Daniel M. Roy. Data-dependent PAC-Bayes priors via differential privacy. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems*

2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, pages 8440–8450, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/9a0ee0a9e7a42d2d69b8f86b3a0756b1-Abstract.html.

Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, and Daniel M. Roy. In search of robust measures of generalization. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/86d7c8a08b4aaa1bc7c599473f5dddda-Abstract.html.

Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, Gabriel Arpino, and Daniel Roy. On the role of data in pac-bayes. In Arindam Banerjee and Kenji Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 604–612. PMLR, 2021. URL http://proceedings.mlr.press/v130/karolina-dziugaite21a.html.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

Wei Gao and Zhi-Hua Zhou. On the doubt about margin explanation of boosting. *Artif. Intell.*, 203:1–18, 2013. doi: 10.1016/j.artint.2013.07.002. URL https://doi.org/10.1016/j.artint.2013.07.002.

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pages 1–8, Montreal, Quebec, Canada, 2009. ACM Press. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553419.

Allan Grønlund, Lior Kamma, and Kasper Green Larsen. Near-tight margin-based generalization bounds for support vector machines. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3779–3788. PMLR, 2020. URL http://proceedings.mlr.press/v119/gronlund20a.html.

Benjamin Guedj. A Primer on PAC-Bayesian Learning. In *Proceedings of the second congress of the French Mathematical Society*, 2019. URL https://arxiv.org/abs/1901.05353.

Benjamin Guedj and Pierre Alquier. PAC-Bayesian estimation and prediction in sparse additive models. *Electron. J. Statist.*, 7:264–291, 2013. doi: 10.1214/13-EJS771. URL https://doi.org/10.1214/13-EJS771.

Steve Hanneke and Aryeh Kontorovich. Stable sample compression schemes: New applications and an optimal SVM margin bound. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato, editors, *Algorithmic Learning Theory, 16-19 March 2021, Virtual Conference, Worldwide*, volume 132 of *Proceedings of Machine Learning Research*, pages 697–721. PMLR, 2021. URL http://proceedings.mlr.press/v132/hanneke21a.html.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SJgIPJBFvH.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 5436–5446. PMLR, 2020. URL http://proceedings.mlr.press/v119/kristiadi20a.html.

John Langford and Matthias Seeger. Bounds for averaging classifiers. 2001. URL http://www.cs.cmu.edu/~jcl/papers/averaging/averaging_tech.pdf.

John Langford and John Shawe-Taylor. PAC-Bayes & margins. In *Advances in Neural Information Processing Systems*, pages 439–446, 2003.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Gaël Letarte, Pascal Germain, Benjamin Guedj, and Francois Laviolette. Dichotomize and generalize: PAC-Bayesian binary activated deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 6872–6882. Curran Associates, Inc., 2019.

Andreas Maurer. A note on the PAC-Bayesian theorem. *CoRR*, cs.LG/0411099, 2004. URL http://arxiv.org/abs/cs.LG/0411099.

David A McAllester. Some PAC-Bayesian theorems. In *Proceedings of the eleventh annual conference on Computational Learning Theory*, pages 230–234. ACM, 1998.

David A McAllester. PAC-Bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational Learning Theory*, pages 164–170. ACM, 1999.

David A. McAllester. Simplified PAC-Bayesian margin bounds. In Bernhard Schölkopf and Manfred K. Warmuth, editors, *Computational Learning Theory and Kernel Machines, 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003, Proceedings*, volume 2777 of *Lecture Notes in Computer Science*, pages 203–215. Springer, 2003. doi: 10.1007/978-3-540-45167-9_16.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11611–11622, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/05e97c207235d63ceb1db43c60db7bbb-Abstract.html.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=Skz_WfbCZ.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. The role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BygfghAcYX.

A. B. Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, volume 12, pages 615–622, New York, NY, USA, 1962. Polytechnic Institute of Brooklyn.

Emilio Parrado-Hernández, Amiran Ambroladze, John Shawe-Taylor, and Shiliang Sun. Pac-bayes bounds with data dependent priors. *J. Mach. Learn. Res.*, 13:3507–3531, 2012. URL http://dl.acm.org/citation.cfm?id=2503353.

Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, October 1998. doi: 10.1214/aos/1024691352.

Matthias Seeger, John Langford, and Nimrod Megiddo. An improved predictive accuracy bound for averaging classifiers. In *Proceedings of the 18th International Conference on Machine Learning*, number CONF, pages 290–297, 2001.

J. Shawe-Taylor and R. C. Williamson. A PAC analysis of a Bayes estimator. In *Proceedings of the 10th annual conference on Computational Learning Theory*, pages 2–9. ACM, 1997.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012. doi: 10.1007/s10208-011-9099-z. URL https://doi.org/10.1007/s10208-011-9099-z.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, 2021. doi: 10.1145/3446776. URL https://doi.org/10.1145/3446776.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. Non-vacuous generalization bounds at the ImageNet scale: a PAC-Bayesian compression approach. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=BJgqqsAct7.

Supplementary Material: On Margins and Derandomisation in PAC-Bayes

A PAC-BAYES BOUNDS

Here we give three different PAC-Bayesian bounds for losses in $[0,1]$, as used in the proof of Theorem 1. We also define the convenience function (for $C > 0, p \in [0,1]$)

$$\Phi_C(p) = -\frac{1}{C}\log(1 - p + pe^{-C})$$

which has inverse

$$\Phi_C^{-1}(t) = \frac{1 - e^{-Ct}}{1 - e^{-C}}.$$

Theorem 7. *Given data distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, prior $P_0 \in \mathcal{M}^1(\mathcal{H})$, $C > 0$ and $\delta \in (0,1)$, the following hold each with probability $\geq 1 - \delta$ over $S \sim D^m$, for all $Q \in \mathcal{M}^1(\mathcal{H})$*

"small-kl" (Langford and Seeger (2001), with improvement by Maurer (2004))

$$\mathrm{kl}(\hat{L}(Q) : L(Q)) \leq \frac{1}{m}\left(\mathrm{KL}\left(Q, P_0\right) + \log\frac{2\sqrt{m}}{\delta}\right)$$

"Catoni" (Catoni, 2007)

$$L(Q) \leq \Phi_C^{-1}\left(\hat{L}(Q) + \frac{\mathrm{KL}\left(Q, P_0\right) + \log\frac{1}{\delta}}{Cm}\right)$$

For completeness, we also include here Proposition 2.1 from Germain et al. (2009).

Lemma 4. *For any $0 \leq q \leq p < 1$,*

$$\sup_{C>0}\left[C\Phi_C(p) - Cq\right] = \mathrm{kl}(q : p).$$

Lemma 5. *For all $\epsilon \in [0, \frac{1}{2}], p \in [0,1], p > \epsilon$ (with the final condition ensuring the left hand side is well-defined),*

$$\mathrm{kl}(\epsilon : p - \epsilon) \geq p + 4\epsilon\log\epsilon.$$

Proof. Note that $-\log(p - \epsilon) \geq 0$ if $p \leq 1$ and thus $\epsilon\log\frac{\epsilon}{R-\epsilon} \geq \epsilon\log\epsilon$. Using the bound $\log x \leq x - 1$ we also have that $(1-\epsilon)\log\frac{1-\epsilon}{1+\epsilon-R} \geq R - 2\epsilon$. Combining these results, $\mathrm{kl}(\epsilon : p - \epsilon) \geq p + \epsilon(\log\epsilon - 2)$; combination with the bound $\epsilon(\log\epsilon - 2) \geq 4\epsilon\log\epsilon$ in the specified range to completes the proof. □

B PROOF OF THEOREM 2

First we prove the following lemma.

Lemma 6. *Let $\tilde{\mu}$ be a probability distribution supported only on A, then for any other probability distirbution ν*

$$\mathrm{KL}\left(\tilde{\mu}, \nu\right) \geq -\log\nu(A).$$

Proof. For the case $\nu(A) = 0$ or where $\tilde{\mu}$ is not absolutely continuous w.r.t. ν the above holds trivially as the KL is infinite.

Thus assume $\nu(A) > 0$ and $\tilde{\mu} \ll \nu$. Define the restriction (or conditional distribution) of ν to A as

$$\tilde{\nu} = \begin{cases} \nu/\nu(A) & \text{on A} \\ 0 & \text{else.} \end{cases}$$

Given the above assumptions, we have

$$\frac{d\tilde{\mu}}{d\nu} = \frac{d\tilde{\mu}}{d\tilde{\nu}} \frac{1}{\nu(A)}$$

so from the definition of and non-negativity of KL divergence,

$$\mathrm{KL}\left(\tilde{\mu}, \nu\right) = \log \frac{1}{\nu(A)} + \mathrm{KL}\left(\tilde{\mu}, \tilde{\nu}\right) \geq -\log \nu(A).$$

□

Proof of Theorem 2. Let Q_w be a deterministic distribution on as-yet-unspecified vector w.

$\kappa(Q_w, P_0; \gamma, 0) = \min_{P \in \mathcal{P}} \mathrm{KL}\left(P, P_0\right)$ where

$$\begin{aligned}
\mathcal{P} &= \{P : \mathrm{AV}(P, Q_w) = 0\} \\
&= \{P : \sup_{(x,y) \in \mathcal{Z}} \mathbb{P}_{u \sim P}(y\langle w - u, x\rangle > \gamma) = 0\} \\
&= \{P : \forall (x, y) \in \mathcal{Z}, \forall u \in \mathrm{supp}(P), y\langle w - u, x\rangle \leq \gamma\} \\
&= \{P : \forall u \in \mathrm{supp}(P), \|w - u\|_2 \leq \gamma\} \\
&= \{P : \mathrm{supp}(P) \subset \mathrm{Ball}(w, \gamma)\}.
\end{aligned}$$

Combining with the Lemma 6 we find that $\kappa(Q_w, P_0; \gamma, 0) \geq -\log P_0[\mathrm{Ball}(w, \gamma)]$.

Since w can be chosen in an adversarial manner based on P_0, P_0 must not over-weight any such ball. The P_0 which minimises κ over all choices of Q_w is thus uniform over the set of possible balls $\mathrm{Ball}(w, \gamma)$, which is the set $\mathrm{Ball}(0, 1 + \gamma)$ (since $\|w\|_2 \leq 1$).

Basic calculation then shows that

$$\kappa \geq -\log \frac{\mathrm{vol}[\mathrm{Ball}(w, \gamma)]}{\mathrm{vol}[\mathrm{Ball}(0, 1 + \gamma)]} = d \log \frac{1 + \gamma}{\gamma} = \Omega(d).$$

□

C COMPARISON OF THEOREM 3 TO EXISTING BOUNDS

Here we discuss how our results in Theorem 3 compare to existing results. All of the following will be in the setting of this theorem with $R = 1$ (since all bounds only depend on R through the scaling R/γ).

C.1 Hard Margin Case

The best existing result for this case is the following, from Bartlett and Shawe-Taylor (1998) (In theorem 1.6, they show that $\mathrm{fat}(\gamma) \leq (R/\gamma)^2$, then they use theorem 1.5; here we give the explicit constants from this theorem):

$$L \leq \frac{1}{m}\left(\frac{256}{\gamma_\star^2} \log \frac{em\gamma_\star^2}{16} \log 32m + 2 \log \frac{8m}{\delta}\right) \in O\left(\frac{1}{m}\left(\gamma_\star^{-2} \log^2 m + \log \frac{1}{\delta}\right)\right).$$

From this we see that not only does Theorem 3 improve in order by removing a factor of $\log m$, but also considerably improves the constant factors.

Grønlund et al. (2020) show that there exists a dataset, and an estimator with $\|w\|_2 \leq 1$, such that:

$$L \geq \Omega\left(\frac{\log m}{m\gamma_\star^2}\right)$$

which is matched by our Theorem 3 and confirms it cannot be improved in order without additional assumptions.

C.2 Soft Margin

Several somewhat different results appear in this case. Using Rademacher complexity, Theorem 21 of Bartlett and Mendelson (2002) implies the following (where the big-O follows by combination with the trivial bound $L \leq 1$):

$$L \leq \hat{L}_\gamma + \frac{4}{\sqrt{m}\gamma} + \left(\frac{8}{\gamma}+2\right)\sqrt{\frac{\log\frac{4}{\delta}}{m}} = \hat{L}_\gamma + O\left(\sqrt{\frac{\gamma^{-2}+\log(1/\delta)}{m}}\right).$$

Based on a more complex bound in Langford and Shawe-Taylor (2003), McAllester (2003) gives the bound (for $m \geq 4$):

$$\begin{aligned} L &\leq \hat{L}_\gamma + \frac{8}{m\gamma^2} + 2\sqrt{2\left(\frac{\hat{L}_\gamma}{m\gamma^2} + \frac{4}{m^2\gamma^4}\right)\log\frac{m\gamma^2}{4}} + O\left(\sqrt{\frac{\log m + \log(1/\delta)}{m}}\right) \\ &= \hat{L}_\gamma + O\left(\frac{\log m}{m\gamma^2} + \sqrt{\frac{\log m}{m\gamma^2}\hat{L}_\gamma} + \sqrt{\frac{\log m + \log(1/\delta)}{m}}\right). \end{aligned}$$

The above also leads to a hard-margin formulation which is however weaker than the Bartlett and Shawe-Taylor (1998) bound for all but very tiny margins, as pointed out by Grønlund et al. (2020).

The state-of-the-art, nearly tight bound given by Grønlund et al. (2020) is the following, not given with any constants,

$$L \leq \hat{L}_\gamma + O\left(\frac{\gamma^{-2}\log m + \log(1/\delta)}{m} + \sqrt{\frac{\gamma^{-2}\log m + \log(1/\delta)}{m}\cdot\hat{L}_\gamma}\right)$$

which is shown in the same paper to be nearly-tight, in the existential sense that there exist data distributions for which it cannot be improved.

This matches exactly the bound given in Theorem 3 in order, but we emphasise both the simplicity of our proof and that we give constants, making it actually evaluable in practice.

D LINEAR CLASSIFICATION WITH L_1/L_∞ NORMS

Here we provide a bound for L_1-normed linear predictors, that essentially replicates the results of Gao and Zhou (2013) or Langford and Seeger (2001).

Theorem 8. *In the binary classification setting with $\mathcal{X} \subset \mathbb{R}^K$ such that $\|x\|_\infty \leq R$, for any $\delta \in (0,1)$, $m \geq 8$, and $\gamma > 0$, with probability $\geq 1-\delta$ over $S \sim D^m$, for all linear predictors $F_w(x) = \sum_{k=1}^K w_k x_k$ with $\|w\|_1 \leq 1$*

$$L(F_{\boldsymbol{w}}) \leq \hat{L}_\gamma(F_{\boldsymbol{w}}) + \sqrt{\frac{\hat{L}_\gamma(F_{\boldsymbol{w}})\cdot\Delta}{m}} + \frac{\Delta + \sqrt{\Delta} + 2}{m},$$

where we define $\Delta := 2\log(2/\delta) + 19(R/\gamma)^2\log(2K)\log m$.

Proof of Theorem 8. Without loss of generality (since we can always simultaneously re-scale the margin and these) we consider $R = 1$. For simplicity we will also assume initially that the weights are non-negative; negative weights can later be included through the standard method of doubling the dimension.

Our prediction function then has the form $F(x) = \sum_{k=1}^K w_k x_k$. For a fixed margin $\theta > 0$, we approximate this by unweighted averages of the form

$$f(x) = \frac{1}{T}\sum_{t=1}^T x_{d(t)}$$

where the indices $d(t) \sim P$ for some distribution P over $[K]$. When T such indices are drawn, we denote this distribution over functions by P^T. As an average of T independent bounded variables, P^T is $(1/T)$-sub-Gaussian with mean F, and thus by Lemma 3

$$\mathrm{AV}_\theta(P^T, \delta(F)) \leq e^{T\theta^2/8} = \epsilon.$$

Choosing P_0 as a uniform distribution on $[K]$ and P_0^T as T independent copies of this, we see that

$$\mathrm{KL}\left(P^T, P_0^T\right) = T\,\mathrm{KL}\left(P, P_0\right) = T\left(\log K - H[w]\right) \leq T \log K$$

where $H[w]$ is the entropy of a categorical variable with (normalised) weights w. This expression using $H[w]$ could be explicitly used (or with a non-uniform prior) to improve the bound, as in Seeger et al. (2001); we will ignore this here and just use the upper bound.

Setting $T = \lceil 8\theta^{-2} \log m \rceil$ in the small-kl formulation of Theorem 1, we obtain that

$$\mathrm{kl}(\hat{L}_\gamma + m^{-1} : L - m^{-1}) \leq \frac{1}{m}\left(\lceil 8\theta^{-2}\log m\rceil \log K + \frac{1}{2}\log m + \log\frac{2}{\delta}\right) \leq \frac{\Delta}{2m} \tag{5}$$

with probability at least $1-\delta$. $\Delta := 19\theta^{-2}\log K \log m + 2\log(2/\delta)$, since $\theta^{-2} \geq 1$ and $m \geq 2$ for a non-vacuous bound.

Relaxing using the lower bound $\mathrm{kl}(q:p) \geq (p-q)^2/(2p)$ for $p > q$ as in the proof of Theorem 3, we obtain $L \leq \hat{L}_\gamma + 2m^{-1} + \sqrt{(\hat{L}_\gamma + m^{-1})\cdot\Delta/m} + \Delta/m$. To complete the proof, we allow negative weights by doubling the dimensions. □

E PROOF OF THEOREM 5

We begin by stating the following useful lemma.

Lemma 7. *Let $X \in \mathcal{M}^1(\{+1,-1\})$ be a random variable with $\mathbb{E}[X] = x$, and*

$$h(x) := \mathrm{KL}\left(X, \mathrm{Uniform}(\{+1,-1\})\right)$$

the KL divergence from a uniform prior. Then

$$h(x) = \frac{1}{2}\left[(1+x)\log(1+x) + (1-x)\log(1-x)\right] \leq x^2 \log 2.$$

Proof. The second equation is simply an explicit statement of the KL divergence. It is easy to see from convexity that $h(x) \leq x^2$; the improved (and optimal) constant of $\log 2$ requires a more complex argument, as follows.

Calculation gives the Maclaurin series

$$(1+x)\log(1+x) + (1-x)\log(1-x) = x^2 + \sum_{n=2}^{\infty} \frac{x^{2n}}{n(2n-1)}$$

which has a radius of convergence of 1. Therefore

$$h(x)/x^2 = \frac{1}{2} + \frac{1}{2}\sum_{n=2}^{\infty}\frac{x^{2n}}{(n+1)(2n+1)}$$

which is an increasing function on $(0,1)$ with supremum $\log 2$. From the definition, $x \in [-1,1]$. A similar argument applies for $(-1,0)$ and equality is achieved at $x = 0$, so the inequality holds (and is the tightest possible). □

Proof of Theorem 5. Let P be a probability measure on $\mathbb{R}^d \times \{+1,-1\}^c$ defined by the following hierarchical procedure: draw a mixture component $k \sim \mathrm{Uniform}(K)$; then $\boldsymbol{w} \in \mathbb{R}^d$ from a Gaussian $\mathcal{N}(U_{k,\cdot}, I)$ (with mean vector as a row of U) and for $i \in [c]$ draw a component of $\boldsymbol{r} \in \{+1,-1\}^c$ such that $\mathbb{E}\boldsymbol{r}[i] = V_{ik}/V_\infty$. A sample from P is a tuple $(\boldsymbol{w}, \boldsymbol{r})$.

P^T is then defined for $T \in \mathbb{N}$ as a distribution on functions $f : \mathbb{R}^d \to \mathbb{R}^c$, of the following form:

$$f(\boldsymbol{x}) = \frac{1}{T}\sum_{t=1}^{T} \mathrm{sign}(\boldsymbol{w}^t \cdot \boldsymbol{x})\boldsymbol{r}^t$$

for T independently drawn samples $(\boldsymbol{w}^t, \boldsymbol{r}^t) \sim P$. It is straightforward to see that $F(\boldsymbol{x}) = V_\infty K \cdot \mathbb{E}_{P^T}[f(\boldsymbol{x})]$ and therefore $L_{\theta'}(F) = L_\theta(\mathbb{E}[f])$ where $\theta' = \theta V_\infty K$. Further, since P^T is the average of T independent bounded variables for any fixed $\boldsymbol{x}$, it is $(1/T)$-sub-Gaussian.

Thus, for any fixed T, $\theta > 0$ and prior P_0^T, we have by Theorem 1 and Lemma 2 that

$$\mathrm{kl}(\hat{L}_{\theta'}(F) + e^{-T\theta^2/16} : L(F) - e^{-T\theta^2/16}) \leq \frac{1}{m}\left(\mathrm{KL}\left(P^T, P_0^T\right) + \log\frac{2\sqrt{m}}{\delta}\right).$$

We now define a prior distribution on individual parameters P_0, and functions P_0^T, in a similar way, but with the distribution over each component of $\boldsymbol{r}$ uniform on $\{+1, -1\}$, and the Gaussian mixture means as rows of the data-free matrix U^0. Since the samples are independently drawn and the distributions P, P_0 over parameters imply those over functions, $\mathrm{KL}\left(P^T, P_0^T\right) \leq T\,\mathrm{KL}\left(P, P_0\right)$.

P_0 and P can be seen as distributions on $([K] \times \mathbb{R}^d \times \{+1, -1\}^c)$ with the index $k \in [K]$ marginalised out. From the chain rule for conditional entropy and Lemma 7, (in a slight abuse of notation since P_0, P are not necessarily densities)

$$\begin{aligned}
\mathrm{KL}\left(P(\boldsymbol{w}, \boldsymbol{r}), P_0(\boldsymbol{w}, \boldsymbol{r})\right) &\leq \mathrm{KL}\left(P(k, \boldsymbol{w}, \boldsymbol{r}), P_0(k, \boldsymbol{w}, \boldsymbol{r})\right) \\
&= \mathrm{KL}\left(P(k), P_0(k)\right) + \mathrm{KL}\left(P(\boldsymbol{w}, \boldsymbol{r}|k), P_0(\boldsymbol{w}, \boldsymbol{r}|k)\right) \\
&= \mathrm{KL}\left(P(\boldsymbol{w}, \boldsymbol{r}|k), P_0(\boldsymbol{w}, \boldsymbol{r}|k)\right) \\
&= \mathrm{KL}\left(P(\boldsymbol{w}|k), P_0(\boldsymbol{w}|k)\right) + \mathrm{KL}\left(P(\boldsymbol{r}|k), P_0(\boldsymbol{r}|k)\right) \\
&= \frac{1}{K}\sum_{k=1}^{K}\frac{\|U_{k,\cdot} - U^0_{k,\cdot}\|_2^2}{2} + \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{c} h(V_{ik}/V_\infty) \\
&\leq \frac{\|U - U^0\|_F^2}{2K} + \frac{\|V\|_F^2}{V_\infty^2 K}\log 2.
\end{aligned}$$

For any fixed $\theta = \gamma/(V_\infty K) > 0$ and $m' > 2$, we set $T = \lceil 16\theta^{-2}\log m' \rceil$. The following then holds with probability at least $1 - \delta$:

$$m \cdot \mathrm{kl}\left(\hat{L}_{\theta'}(F) + \frac{1}{m'} : L - \frac{1}{m'}\right) \leq \lceil 16\theta^{-2}\log m' \rceil \left(\frac{\|U - U^0\|_F^2}{2K} + \frac{\|V\|_F^2}{V_\infty^2 K}\log 2\right) + \log\frac{2\sqrt{m}}{\delta}.$$

It remains to cover possible values of θ. Firstly we note that for $\theta \geq 1$ the bound is trivially true by the boundedness of $f(\boldsymbol{x})$, and thus we need only consider $\theta^{-2} > 1$.

For $\alpha > 1$ and $i = 0, 1, \ldots$, set $\theta_i = \alpha^{-i}$ and $\delta_i = \delta/2(i+1)^2$. Applying the union bound over the above equation with these parameters we get that with probability at least $1 - (\pi^2/6)\delta \geq 1 - 2\delta$ that the above is true for each pair of θ_i and δ_i. We choose the largest θ_i such that $\theta_i \leq \theta < \theta_{i-1}$, so that $i \leq 1 - \log_a(\theta)$. Since $\hat{L}_\theta \leq \hat{L}_{\theta_i}$ is increasing, $1/\theta_i \leq a/\theta$, and $\log(1/\delta_i) \leq \log(1/\delta) + 2\log(2 + \log_a(1/\theta)) = \log(1/\delta) + 2\log(\log(\alpha^2/\theta)/\log(\alpha))$, we finally obtain with probability $1 - \delta$

$$\begin{aligned}
m \cdot \mathrm{kl}\left(\hat{L}_\gamma(F) + \frac{1}{m'} : L - \frac{1}{m'}\right) \leq& 17\left(\frac{\alpha V_\infty K}{\gamma}\right)^2\left(\frac{\|U - U^0\|_F^2}{2K} + \frac{\|V\|_F^2}{V_\infty^2 K}\log 2\right)\log m' \\
&+ \log\frac{4\sqrt{m}}{\delta} + 2\log\left(\frac{\log(\alpha^2 V_\infty K/\gamma)}{\log\alpha}\right)
\end{aligned}$$

for all weight matrices and every $\gamma > 0$, and fixed $K > 0, \alpha > 1$.

Relaxing the bound with Pinsker's inequality $\mathrm{kl}(a : b) \geq (a - b)^2$ and setting $m' = m$ and $\alpha = 2$ completes the proof. □

F PROOF OF THEOREM 6

Here we give the proof of Theorem 6, beginning with two lemmas used.

Lemma 8 (Neyshabur et al., 2018; Lemma 2, Perturbation Bound.)**.** *In the setting of Theorem 6, for any layer weights W_i, $x \in \mathcal{X}$ and weight perturbations U_i such that $\|U_i\|_2 \leq d^{-1}\|W_i\|_2$,*

$$\|f(x) - F(x)\|_2 \leq eR\left(\prod_{i=1}^{d}\|W_i\|_2\right)\sum_{i=1}^{d}\frac{\|U_i\|_2}{\|W_i\|_2}$$

where F is the unperturbed and f the perturbed network (with weights W_i and $W_i + U_i$ respectively).

Lemma 9. *Let $Q = \delta(F)$ for such a feed-forward ReLU network with weights W_i, and P be the same network with Gaussian weights, with per-layer means W_i and variances σ_i^2. Then for all $0 < \theta < 2\sup_{x\in\mathcal{X}, y\in[K]}|F(x)[y]|$,*

$$\mathrm{AV}_\theta(P,Q) \leq 2h\sum_{i=1}^{d}\exp\left(-\frac{1}{32h}\left(\frac{\theta\|W_i\|_2}{\sigma_i eR(\prod_i\|W_i\|_2)}\right)^2\right).$$

Proof. From Lemma 8, we see immediately that if for all i, the perturbations have $\|U_i\|_2 \leq c\theta\|W_i\|_2$ for $c^{-1} = 4edR\left(\prod_i\|W_i\|_2\right)$, then $\|f(x) - g(x)\|_2 \leq \theta/4$. The perturbation condition of Lemma 8 is satisfied if $\theta < 2eR\prod_i\|W_i\|_2$, which is true for any θ in the range of the function margins (as in the lemma assumption, since $R\prod_i\|W_i\|_2$ is an upper bound on the range). If the perturbations are randomised, we see that (letting $y' \neq y$ achieve the maximum margin)

$$\begin{aligned}
\mathrm{AV}_\theta(P,Q) &\leq \mathbb{P}\{|M(f,z) - M(g,z)| > \theta/2\} \\
&\leq \mathbb{P}\{|f(x)[y] - f(x)[y'] - g(x)[y] + g(x)[y']| > \theta/2\} \\
&\leq \mathbb{P}\{2\|f(x) - g(x)\|_\infty > \theta/2\} \\
&\leq \mathbb{P}\{\|f(x) - g(x)\|_2 > \theta/4\} \\
&\leq \mathbb{P}\{\exists i : \|U_i\|_2 > c\theta\|W_i\|_2\} \\
&\leq \sum_{i=1}^{d}\mathbb{P}\{\|U_i\|_2 > c\theta\|W_i\|_2\}.
\end{aligned}$$

We set the weights of g to be Gaussian with diagonal covariance, and per-layer variances of σ_i^2. To complete the proof we use a result of Tropp (2012) for Gaussian random matrices, that

$$\mathbb{P}\{\|U_i\|_2 > t\} \leq 2he^{-t^2/2h\sigma_i^2}.$$

□

Proof of Theorem 6. We choose P and P^0 to have Gaussian weight matrices with means W_i and W_i^0, and identical per-layer variances σ_i. From Lemma 9 we have for any fixed θ and set of σ_i, and for all F with weights W_i, such that the inverse variances

$$\sigma_i^{-2} \geq 32h\left(\frac{eR(\prod_i\|W_i\|_2)}{\theta\|W_i\|_2}\right)^2\log(mhd) \tag{6}$$

we have $\mathrm{AV}_\theta \leq 2/m$. Therefore from Theorem 1 and Pinsker's inequality we have the generalisation bound (for the weight matrices and θ satisfying the condition on the set of σ_i)

$$L(F) \leq \hat{L}_\theta(F) + \frac{2}{m} + \sqrt{\frac{1}{2m}\left(\sum_{i=1}^{d}\frac{\|W_i - W_i^0\|_F^2}{\sigma_i^2} + \log\frac{2\sqrt{m}}{\delta}\right)}.$$

We complete the proof by constructing covers for σ_i and θ. We only need to consider $\theta < R\prod_i \|W_i\|_2 =: C_\theta$ (an upper bound on the range of the function) as otherwise the $\hat{L}_\theta$ term is 1 and the bound is vacuous. Since $\|W_i\|_2 \leq W_\star$ for all i we have that $\sigma_i^{-2} \geq 32e^2 h \|W_i\|_2^{-2} \geq 32e^2 h W_\star^{-2}$ and $\sigma_i \leq 15h^{-1/2}W_\star^2 =: C_\sigma$.

For $t = 0, 1, 2, \ldots$ choose margins $\theta^{(t)} = C_\theta/2^t$ and let the bound for this margin hold with probability $\delta^{(t)} = \delta/(t+1)^2$, so that taking a union bound the above holds simultaneously for every $\theta^{(t)}$ with probability at least $1 - \pi^2\delta/6 \geq 1 - 2\delta$. To find a bound holding simultaneously for all θ, we choose the t such that $\theta^{(t)} \leq \theta < \theta^{(t-1)}$, and then replace this term with θ by using the facts that $\hat{L}_{\theta^{(t)}} \leq \hat{L}_\theta$, $1/\theta^{(t)} \leq 2/\theta$, and $\log(1/\delta^{(t)}) \leq \log(1/\delta) + 2\log\log_2(4C_\theta/\theta)$.

Repeating this same covering process for every choice of σ_i, we obtain with probability at least $1-2\delta$ simultaneously for all θ, σ_i (and thus also for the tightest σ_i satisfying Equation (6)) that $L(F) - \hat{L}_\theta(F)$ is upper bounded by

$$\frac{2}{m} + \sqrt{\frac{1}{2m}\left(4\sum_{i=1}^d \frac{\|W_i - W_i^0\|_F^2}{\sigma_i^2} + \log\frac{2(d+1)\sqrt{m}}{\delta} + 2\log\log_2(4C_\theta/\theta) + \sum_{i=1}^d 2\log\log_2(4C_\sigma/\sigma_i)\right)}$$
$$\in O\left(\sqrt{\frac{hR^2\left(\prod_{i=1}^d \|W_i\|_2^2\right)\log(mdh)}{\theta^2 m} \cdot \sum_{i=1}^d \frac{\|W_i - W_i^0\|_F^2}{\|W_i\|_2^2} + \frac{\log\frac{1}{\delta} + d\log\log W_\star}{m}}\right)$$

□

G EMPIRICAL EVALUATION OF Theorem 5

G.1 Experimental setup

All experiments were performed using the Tensorflow 2 library (Abadi et al., 2015) in Python, on a single workstation with a Nvidia RTX 2080 Ti GPU. Code for the results is licensed under an MIT license and available in the supplementary material.

We train SHEL networks and a partially-aggregated variation thereof under different hyperparameter configurations. We use this to compare changes in the generalisation error (the difference between test and train misclassification errors) with the complexity term from Theorem 5 given by

$$\frac{\sqrt{K}}{\gamma\sqrt{m}}\left(V_\infty\|U - U^0\|_F + \|V\|_F\right). \tag{7}$$

Following previous empirical evaluations of such complexity terms, we train to a fixed value of cross-entropy; see Jiang et al. (2020) for further discussion. The margin γ is set as that giving a fixed $L_\gamma(F) = 0.2$, or $E_{f\sim Q}L_\gamma(f) = 0.2$ for the partially aggregated version.

For the partially-aggregated version, we include a feature map of three additional dense ReLU layers with Gaussian weight matrices with independent components, means $\{W_i\}_{i=1}^3$ and variances of σ. Again using the initialisation as a prior, this adds a term of

$$\sqrt{\sum_{i=1}^3 \|W_i - W_i^0\|_F^2/4m\sigma^2}$$

to the right hand side of the bound. To enable comparison, we set σ to make this term constant and equal to a half when calculating $E_{f\sim Q}L_\gamma(f)$. This is done during the evaluation phase, and training is performed on the non-stochastic version (weights as means) as in Dziugaite and Roy (2017).

These experiments aim to evaluate the predictive ability of this complexity measure under changes of procedures. To this end we provide plots of the generalisation, $G(\omega)$, and complexity measure, $C(\omega)$, for trained parameters ω versus some change in hyperparameter value.

We also provide evaluations using the sign-error, a measure of predictive power defined in Dziugaite et al. (2020) as

$$\frac{1}{2}\mathbb{E}_{\omega,\omega'}[1 - \text{sign}(C(\omega') - C(\omega)) \cdot \text{sign}(G(\omega') - G(\omega))]$$

where ω and ω' are parameters obtained through training with one changed hyperparameter between them. The maximum over such pairs of hyperparameter settings is a measure of the robustness of predictions made about the generalisation based on the complexity measure; if this value is low, the complexity measure makes robust predictions. We provide this maximum, the median, and the mean of the above (as in Jiang et al., 2020) for different setups and allowing different hyperparameters to vary.

G.2 SHEL Network

On the MNIST (LeCun et al., 2010) dataset, we examine the following hyperparameter settings, finding through the sign error (Table 1) that predictions under changes of training size are quite robust, while those under changes of learning rate or width are poor. We additionally provide plots (Figures 1 to 3) for some selected hyperparameter values to verify the above. This poor prediction under such changes is unfortunately a feature of many such complexity measures (Dziugaite et al., 2020).

- Learning rate $\in \{10^{-3}, 3\times10^{-3}, 10^{-2}, 3\times10^{-2}, 10^{-1}\}$.
- Train set sizes $\in \{60\,000, 30\,000, 15\,000, 7\,500\}$.
- Width $\in \{50, 100, 200, 400, 800\}$.
- Batch size $= 200$.
- Learning algorithm SGD with momentum parameter $= 0.9$.

Variable Hyperparameter	Max SE	Median SE	Mean SE
Learning Rate	1.0	0.60	0.56
Width	1.0	1.0	0.90
Train Size	0.2	0.0	0.00
All	1.0	0.60	0.53

Table 1: Statistics of the sign error, SE, under different varying hyperparameters for a SHEL network trained on MNIST.

G.3 Partially-Derandomised SHEL

Again on the MNIST dataset, we evaluate the partially-derandomised version of the above under the same hyperparameter values, excluding learning rates of 0.1 and 0.03 which sometimes led to numerical instability. Figures 4 to 6 provide sample results and the sign-error results are reported in Table 2.

These sign error results show that predictions under changes of training size are completely robust, while those under changes of learning rate or width are still poor. The predictions for width are somewhat improved, though we note that our estimate of this quantity may be somewhat noisy as the generalisation error appears largely independent of width.

Variable Hyperparameter	Max SE	Median SE	Mean SE
Learning Rate	1.0	0.60	0.49
Width	1.0	0.40	0.46
Train Size	0.0	0.0	0.0
All	1.0	0.20	0.31

Table 2: Statistics of the sign error, SE, under different varying hyperparameters for a partially-derandomised SHEL network trained on MNIST.

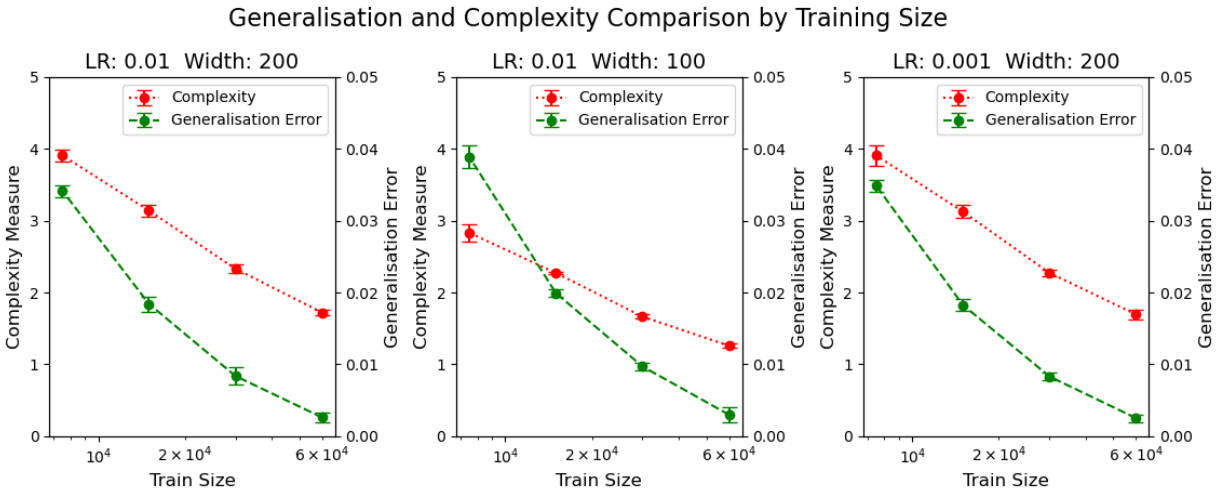

Figure 1: Changes in complexity measure and generalisation error versus training set size under fixed other hyperparameters, for a SHEL network trained on MNIST.

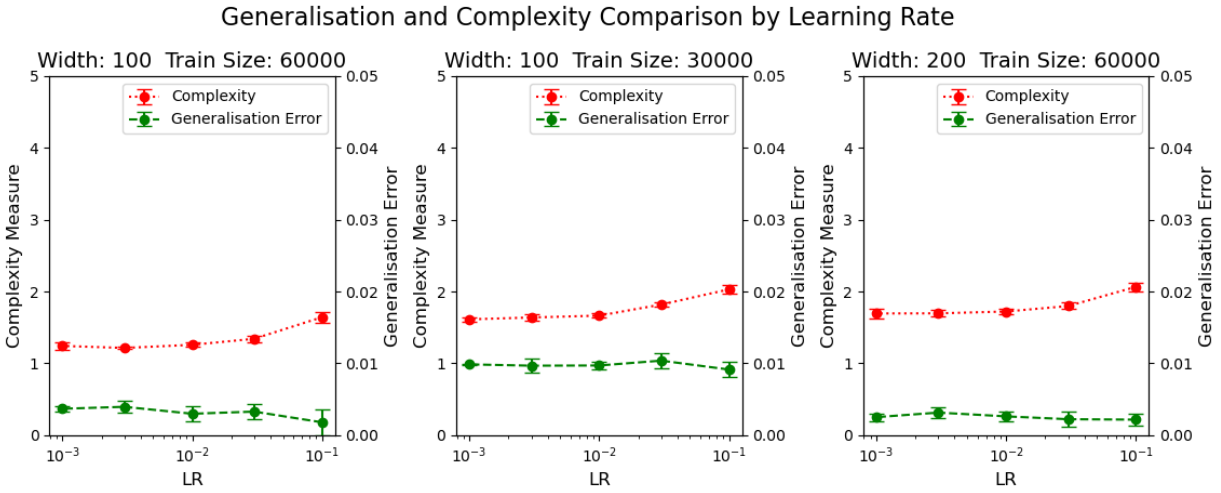

Figure 2: Changes in complexity measure and generalisation error versus learning rate under fixed other hyperparameters, for a SHEL network trained on MNIST.

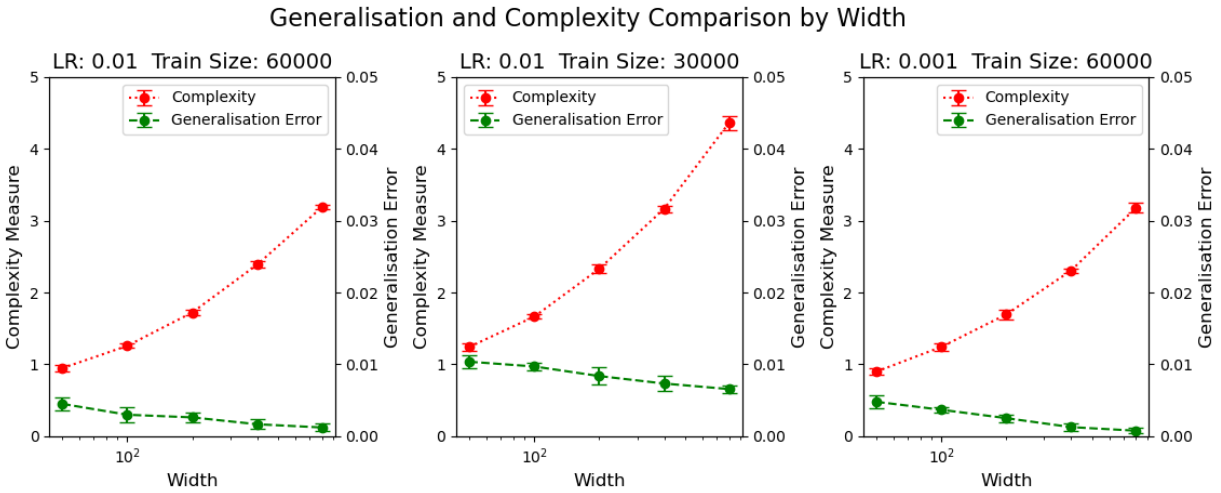

Figure 3: Changes in complexity measure and generalisation error versus width under fixed other hyperparameters, for a SHEL network trained on MNIST.

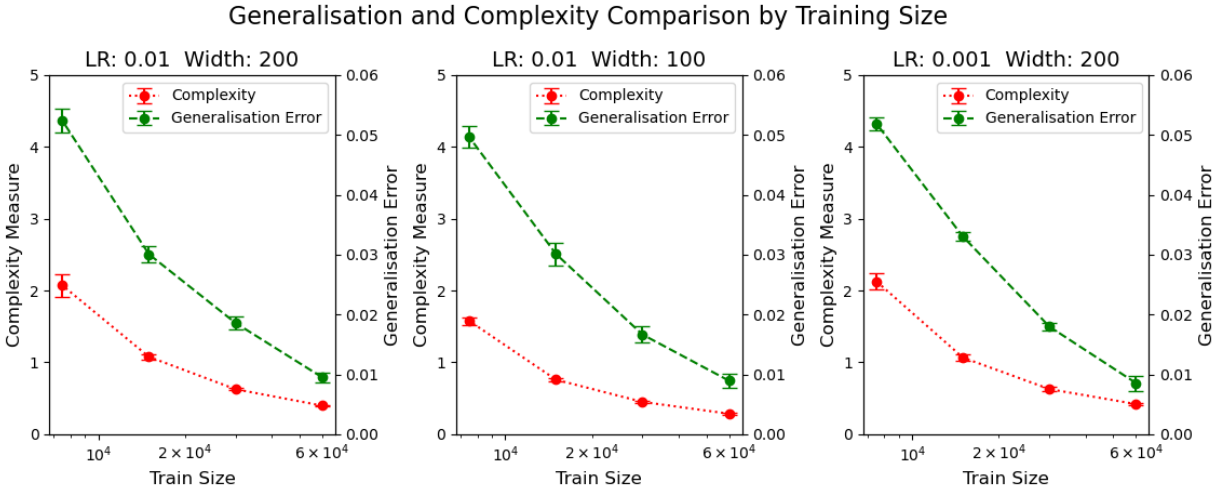

Figure 4: Changes in complexity measure and generalisation error versus training set size under fixed other hyperparameters, for a partially-derandomised SHEL network trained on MNIST.

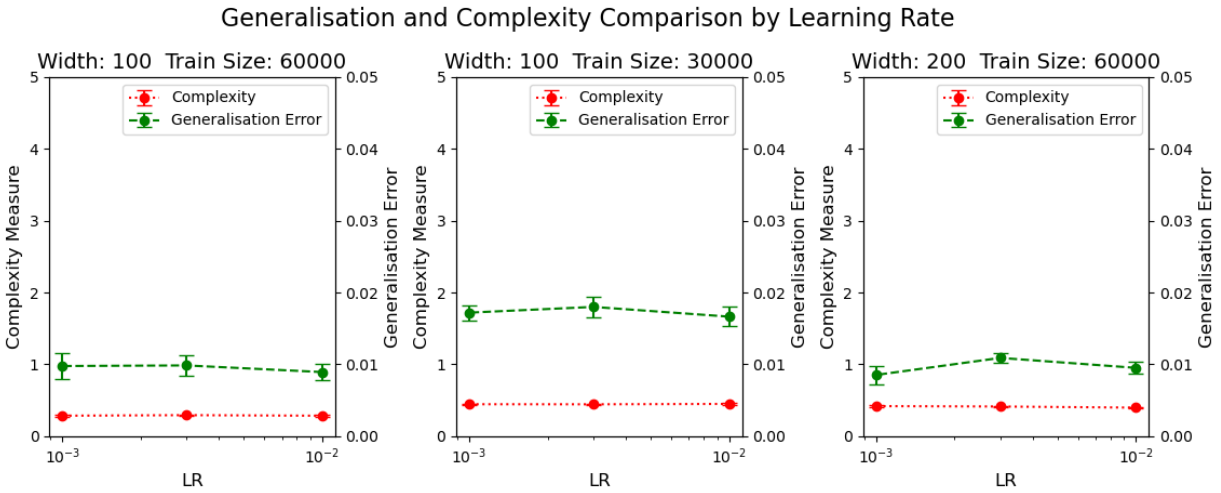

Figure 5: Changes in complexity measure and generalisation error versus learning rate under fixed other hyperparameters, for a partially-derandomised SHEL network trained on MNIST.

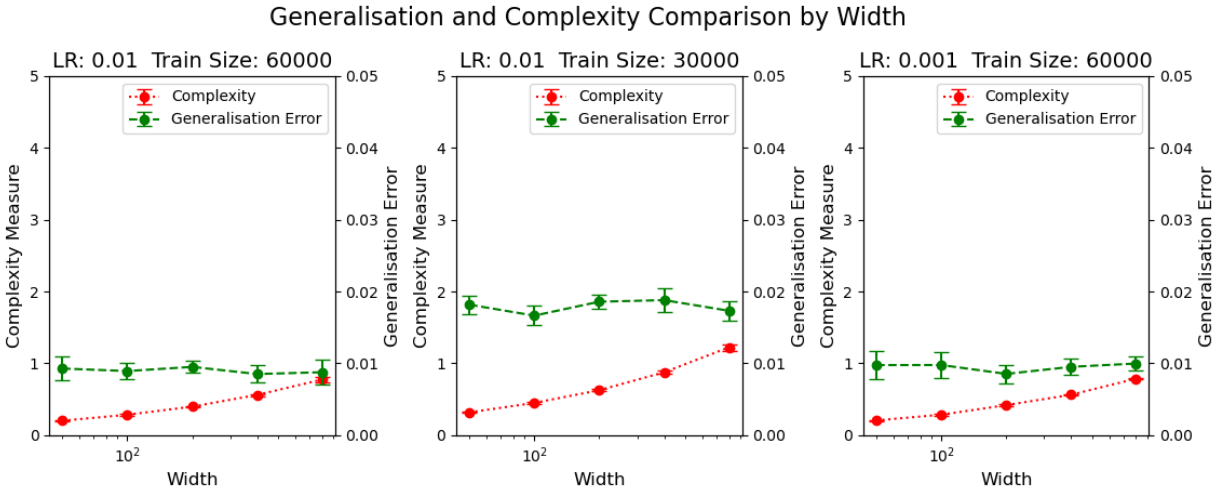

Figure 6: Changes in complexity measure and generalisation error versus width under fixed other hyperparameters, for a partially-derandomised SHEL network trained on MNIST.