Smoothness-Based Derandomization of PAC-Bayes Bounds

Anonymous authors
Paper under double-blind review

Abstract

We study PAC-Bayes derandomization for smooth loss functions. Our goal is to obtain generalization bounds that hold with high probability for deterministic predictors by exploiting smoothness properties of both the loss and the predictor class. We show that passing from the Gibbs predictor to the deterministic predictor at the posterior mean has a precise cost, given by the generalization gap of the Jensen gap class. We control this class through its Rademacher complexity, leading to bounds for deterministic predictors that involve flatness quantities expressed in terms of parameter Jacobians and Hessians of the score map. The framework applies to both bounded and unbounded smooth loss functions, and we specialize the results to linear predictors and smooth neural networks. Finally, the Jacobian and Hessian quantities appearing in the theory motivate a practical regularizer. For BatchNorm networks, we compute this regularizer with respect to effective BatchNorm weights obtained by folding the BatchNorm transformation into the adjacent affine weights. Experiments on CIFAR-10 illustrate the behavior of this regularizer under different batch sizes.

1 Introduction

Flatness measures defined through parameter perturbations are among the more promising predictors of the generalization gap in large-scale empirical studies on neural networks (Jiang et al., 2020; Dziugaite et al., 2020). PAC-Bayes theory provides a natural framework for deriving high probability generalization bounds involving such measures (McAllester, 1998; 1999; Seeger, 2002; Catoni, 2007; Haddouche et al., 2025). However, standard PAC-Bayes bounds apply to stochastic Gibbs predictors, whereas a single deterministic predictor is often used in practice. Derandomization of PAC-Bayes bounds provides a rigorous bridge between guarantees for stochastic classifiers and guarantees for deterministic predictors (Langford & Shawe-Taylor, 2002; Viallard et al., 2021; Neyshabur et al., 2018; Banerjee et al., 2020; Biggs & Guedj, 2022). Existing results connect ideas involving classification margins and flatness through an analysis of parameter perturbations. While this provides an important connection, it ties the resulting guarantees to margin losses. As discussed in Section 2, these margin losses are defined using indicator functions and therefore cannot be directly optimized with gradient based methods. In this work, we consider a new general strategy for PAC-Bayes derandomization that relies on smoothness assumptions. This strategy yields bounds involving flatness measures that apply directly to deterministic predictors, for both classification and regression with bounded or unbounded smooth loss functions.

We show that one can pass from a PAC-Bayes bound for the stochastic Gibbs predictor to a bound for the deterministic predictor at the posterior mean by paying an additional penalty term. This penalty is the generalization gap of the Jensen gap of the loss class; see Theorem 3.3 and Corollary 3.5. The Jensen gap of a function f with respect to a posterior Q is the difference $f(\mathbb{E}_{w\sim Q}[w]) - \mathbb{E}_{w\sim Q} f(w)$; see Definition 3.6. The remaining task is to control this class uniformly with high probability. Section 6 establishes Rademacher complexity bounds using a Talagrand comparison inequality for processes with sub-Gaussian increments (Proposition C.2). The resulting derandomized bounds explicitly depend on smoothness and flatness properties of both the loss and the hypothesis class. In particular, Theorem 3.7 gives the general result in the unbounded case, while Theorem 4.1 gives its specialization to linear predictors.

The main contributions of this work are as follows.

1. We obtain uniform high probability control of the Jensen gap class by bounding its Rademacher complexity (Corollary 6.4, Proposition 6.6, and Proposition 6.7).

2. We derive general derandomized PAC-Bayes bounds for unbounded and bounded smooth loss functions. The unbounded case is treated in Section 3, with the main general result stated in Theorem 3.7. The bounded case is treated separately in Section 5, leading in particular to Theorem 5.1 and to the inverse-kl bound of Theorem 5.3.

3. We specialize the theory to linear predictors. In the unbounded case, Theorem 4.1 gives the corresponding bound, while Theorem 5.2 gives the analogue for bounded losses. A key step in the unbounded analysis is the evaluation of the PAC-Bayes complexity term under Gaussian priors. Lemma 3.8 gives a square root dependence on the dimension for this term. This improves on the linear dependence on the dimension appearing in the bounds of Haddouche et al. (2020). Specific results for classification with the multi-class unhinged loss (Shoham et al., 2021; Zhou et al., 2023; Authors, 2026) and for regression are given in Proposition 4.4, Corollary 4.5, and Corollary 4.6. A comparison with a classical Rademacher complexity bound for a bounded smooth loss is presented in Section 7 and Figures 2–3.

4. We specialize our bounds to SHEL networks and compare them with the margin bound of Biggs & Guedj (2022) in Section 8 and Figures 4–5.

5. We propose a practical regularizer inspired by the Jacobian and Hessian quantities appearing in the theory. For BatchNorm networks, we investigate a parametrization obtained by combining the BatchNorm transformation with the preceding affine weights, leading to effective BatchNorm weights $w_{\mathrm{eff},B'}$. The resulting experiments are presented in Section 9 and Figures 6–8.

Structure of the document. In Section 2, we contextualize our general method. The presentation in Section 3 follows the conceptual development of the bounds. Some of the technical ingredients required for Theorem 3.7 are established later in Section 6 and Appendix A. For clarity of exposition, we state the main results first and invoke the corresponding technical estimates when they are needed. Proofs omitted from the main text are collected in the appendices.

2 Motivation and relation to existing derandomization approaches

We first motivate the role of the Jensen gap through the simple case where the loss is convex in the parameters. If a loss function $L(h_w(x), y)$ is convex with respect to w, then Jensen's inequality implies

$$L(h_{\overline{w}}(x), y) \leq \mathbb{E}_{w\sim Q}[L(h_w(x), y)],$$

where $\overline{w} = \mathbb{E}_{w\sim Q}[w]$. Starting from a PAC-Bayesian bound for L, one can then obtain in a straightforward manner a generalization bound for the deterministic classifier $\overline{w}$ of the form

$$L_D(\overline{w}) \leq L_S(\overline{w}) + \frac{1}{n}\sum_{i=1}^{n}\Big[\mathbb{E}_{w\sim Q}L(h_w(x_i), y_i) - L(h_{\overline{w}}(x_i), y_i)\Big] + \widetilde{O}\left(\frac{1}{\sqrt{n}}\right).$$

The quantity

$$\frac{1}{n}\sum_{i=1}^{n}\Big[\mathbb{E}_{w\sim Q}L(h_w(x_i), y_i) - L(h_{\overline{w}}(x_i), y_i)\Big]$$

can be interpreted as an empirical measure of flatness of the loss L around the mean $\overline{w}$. This direct argument constitutes perhaps the simplest form of derandomization of a PAC-Bayes bound. However, it relies on convexity and therefore cannot be applied to typical neural network hypothesis classes. Moreover, even for linear predictors, it does not apply for example to non-linear symmetric loss functions $L(z, y)$, which are non-convex (Ghosh et al., 2015). This motivates the development of a more general approach.

The framework developed in this work replaces this convexity argument by an exact decomposition. The additional term required for derandomization is the generalization gap of the Jensen gap of the loss class, and the resulting bounds take the schematic form

$$L_D(\overline{w}) \leq L_S(\overline{w}) + \frac{\text{smoothness/flatness terms}}{\sqrt{n}} + \widetilde{O}\left(\frac{1}{\sqrt{n}}\right).$$

This perspective can be advantageous even in convex settings, since the empirical flatness term appearing in the direct convexity argument need not vanish with the sample size. We now position our approach within the literature that seeks guarantees on deterministic classifiers derived from PAC-Bayesian bounds.

Viallard et al. (2021) develop a general framework for deriving pointwise PAC-Bayesian generalization bounds that apply to an individual hypothesis sampled from the posterior, rather than to the Gibbs predictor. This form of derandomization is referred to as the *disintegration* of PAC-Bayesian bounds. Related deterministic PAC-Bayesian perspectives were later developed by Clerico et al. (2025). In the disintegrated setting, for instance, if one assumes an explicit posterior family (e.g., a Gaussian centered at the algorithm's output), the resulting guarantee applies to a specific hypothesis drawn from that posterior and is probabilistic with respect to this sampling step (in addition to the randomness of the sample S). In contrast, our bounds apply directly to the deterministic classifier at the posterior mean in the classical sense and do not introduce additional hypothesis level randomness.

Among direct derandomization approaches, a prominent line of work relies on margin-based arguments. Early such PAC-Bayesian bounds appear in Langford & Shawe-Taylor (2002). Subsequent developments derived results for modern neural network architectures (Neyshabur et al., 2018; Biggs & Guedj, 2022; Banerjee et al., 2020). Define the margin of an example (x, y) as

$$m_{\overline{w}}(x, y) = h_{\overline{w}}(x)_y - \max_{k \neq y} h_{\overline{w}}(x)_k.$$

For a margin parameter $\gamma > 0$, the corresponding margin loss is defined by

$$L_\gamma(h_{\overline{w}}(x), y) = \mathbf{1}\{m_{\overline{w}}(x, y) \leq \gamma\},$$

which upper bounds the 0-1 loss. Margin-based PAC-Bayesian approaches exploit inequalities involving the 0-1 loss, the margin loss and the margin loss under a posterior Q. Consider as an illustration the strategy of Neyshabur et al. (2018). They define a subset $S_{\overline{w}}$ of weights w such that the outputs of h_w remain uniformly close to those of $h_{\overline{w}}$, namely

$$S_{\overline{w}} \subseteq \Big\{ w : \sup_{x \in \mathscr{X}} \|h_w(x) - h_{\overline{w}}(x)\|_\infty < \gamma/4 \Big\}.$$

Conditioning the posterior on this set yields a truncated posterior $\widetilde{Q}$ under which the inequalities

$$L_D^{0\text{-}1}(h_{\overline{w}}) \leq \mathbb{E}_{w \sim \widetilde{Q}}\big[L_{D,\gamma/2}(h_w)\big] \qquad \text{and} \qquad \mathbb{E}_{w \sim \widetilde{Q}}\big[L_{S,\gamma/2}(h_w)\big] \leq L_{S,\gamma}(h_{\overline{w}})$$

hold. This enables the conversion of a stochastic PAC-Bayes guarantee into a deterministic margin-based bound. A related derandomization strategy, developed in Banerjee et al. (2020), also relies on inequalities that directly relate the margin loss of the deterministic predictor to the margin loss of the stochastic predictor with parameters sampled from a posterior Q. These inequalities are obtained by controlling how Gaussian perturbations of the parameters affect the margin, and they introduce a residual term of the form

$$c_0 \exp\big(- \min(c_2\gamma^2,\ c_1\gamma)\big).$$

For fixed margin parameter γ, this term is independent of the sample size n and therefore does not vanish as $n \to \infty$. In contrast, the additional derandomization terms in our bounds vanish with the sample size.

Fundamentally, the Jensen gap framework adopted in this work does not rely on such "derandomization inequalities" for the margin loss. Instead, it starts from an exact decomposition relating deterministic and stochastic risks through the Jensen gap. This gives a conceptually distinct route to derandomization

that is naturally suited to smooth loss functions for regression and classification. Moreover, the empirical loss appearing in our bounds can coincide with the loss used for training, so gradient-based optimization directly acts on a term that enters the bound. By contrast, the margin loss L_γ cannot be directly optimized using gradient-based methods and therefore requires a surrogate training loss that is different from the loss appearing in the bound.

Another line of work obtaining deterministic guarantees from PAC-Bayes bounds is the literature on weighted majority-vote classifiers (McAllester, 1999; Langford & Shawe-Taylor, 2002). Explicit bounds for the risk of the majority vote were derived in terms of the mean and variance of the error of the Gibbs classifier in (Lacasse et al., 2007), leading to the C-bound and the MinCq learning algorithm (Germain et al., 2015). Second-order refinements of the C-bound were later established in (Masegosa et al., 2020). More recently, Leblanc & Germain (2025) proposed a general framework for extracting deterministic risk guarantees from stochastic PAC-Bayesian bounds, with applications to majority-vote classifiers. While powerful, this line of work is inherently tied, in its main applications, to vote-based predictors and does not directly address deterministic predictors defined through the posterior mean of parameters.

3 Generalization bounds for unbounded loss functions from the PAC-Bayes framework

Unbounded loss functions arise naturally in several learning problems. Standard examples include regression losses such as the squared loss, the absolute loss, and the Huber loss, as well as classification losses such as the cross-entropy loss. They also arise in classification through symmetric losses such as the multi-class unhinged loss (Shoham et al., 2021; Zhou et al., 2023) and through related symmetrized losses studied in Authors (2026). In this section, we rely on a PAC-Bayes result applicable to unbounded loss functions from Alquier et al. (2016); see also (Rivasplata et al., 2020; Casado et al., 2024) for related PAC-Bayesian treatments beyond the bounded loss setting. We then exploit the assumption that, for each individual hypothesis h, the loss function $L(h(x), y)$ is bounded on X (typically a bounded domain) in order to apply Hoeffding's Lemma. This strategy is in the same spirit as Haddouche et al. (2020), where hypothesis-dependent boundedness is used to extend PAC-Bayes bounds to learning problems involving unbounded regression loss functions.

Theorem 3.1 (Alquier et al. (2016))**.** *Given a distribution D over $X \times Y$, a hypothesis set $\mathscr{H}$, a loss function $L : \mathscr{H} \times X \times Y \to \mathbb{R}$, a prior distribution P over $\mathscr{H}$, a $\delta \in (0, 1]$, and a real number $\lambda > 0$, with probability at least $1 - \delta$ over the choice of $S \sim D^n$, we have:*

$$\forall Q \text{ on } \mathscr{H} : \quad \mathbb{E}_{h\sim Q} L_D(h) \leq \mathbb{E}_{h\sim Q} L_S(h) + \frac{1}{\lambda}\left[D_{\mathrm{KL}}(Q\|P) + \ln\frac{1}{\delta} + \Psi_{L,P,D}(\lambda, n)\right],$$

where

$$\Psi_{L,P,D}(\lambda, n) := \ln \mathbb{E}_{h\sim P}\mathbb{E}_{S\sim D^n} \exp\left[\lambda\left(L_D(h) - L_S(h)\right)\right].$$

Proof: see Alquier et al. (2016) and Germain et al. (2016). ■

Our next step is to upper bound $\Psi_{L,P,D}(\lambda, n)$ under a hypothesis-dependent boundedness condition.

Proposition 3.2. *Assume that $\|h(x)\| \leq a_h$ on X for every $h \in \mathscr{H}$. Furthermore, assume that $L(z, y)$ is ℓ-Lipschitz in z for every y and that $L(0, y)$ is independent of y. Then, for every $\lambda > 0$,*

$$\Psi_{L,P,D}(\lambda, n) \leq \ln \mathbb{E}_{h\sim P} \exp\left(\frac{\lambda^2\ell^2 a_h^2}{2n}\right).$$

Proof: Without loss of generality, assume that $L(0, y) = 0$, since adding a constant to the loss does not affect $L_D(h) - L_S(h)$. Then, for every fixed h,

$$|L(h(x), y)| \leq \ell\|h(x)\| \leq \ell a_h.$$

Thus, by Hoeffding's Lemma,

$$\mathbb{E}_{S\sim D^n}\exp(\lambda(L_D(h)-L_S(h)))\le\exp\left(\frac{\lambda^2\ell^2a_h^2}{2n}\right).$$

Taking expectation over $h\sim P$ and then the logarithm gives

$$\Psi_{L,P,D}(\lambda,n)=\ln\mathbb{E}_{h\sim P}\mathbb{E}_{S\sim D^n}\exp(\lambda(L_D(h)-L_S(h)))\le\ln\mathbb{E}_{h\sim P}\exp\left(\frac{\lambda^2\ell^2a_h^2}{2n}\right).$$

■

We now convert the bound on the Gibbs risk from Theorem 3.1 into a bound at the mean of the distribution.

Theorem 3.3. *Let D be a distribution on $X\times Y$, let $S=\{(x_i,y_i)\}_{i=1}^n\sim D^n$, let $\mathscr{H}=\{h_w:w\in\mathbb{R}^m\}$, and let P be a prior on $\mathbb{R}^m$. Fix $\delta\in(0,1]$ and $\lambda>0$. For any distribution Q on $\mathbb{R}^m$ and any sample S, define*

$$\Omega_{L,Q,D,S}(n):=\Big(L_D(h_{\overline{w}})-\mathbb{E}_{w\sim Q}L_D(h_w)\Big)-\Big(L_S(h_{\overline{w}})-\mathbb{E}_{w\sim Q}L_S(h_w)\Big),\qquad \overline{w}:=\mathbb{E}_{w\sim Q}[w].$$

Then, with probability at least $1-\delta$ over the draw of S, for all distributions Q on $\mathbb{R}^m$,

$$L_D(h_{\overline{w}})\le L_S(h_{\overline{w}})+\Omega_{L,Q,D,S}(n)+\frac{1}{\lambda}\Big[D_{\mathrm{KL}}(Q\|P)+\ln\tfrac{1}{\delta}+\Psi_{L,P,D}(\lambda,n)\Big],$$

where $\Psi_{L,P,D}(\lambda,n)$ is as in Theorem 3.1.

Proof: Start from the identity

$$L_D(h_{\overline{w}})=\mathbb{E}_{w\sim Q}L_D(h_w)+\Big(L_D(h_{\overline{w}})-\mathbb{E}_{w\sim Q}L_D(h_w)\Big).$$

On the event of probability at least $1-\delta$ given by Theorem 3.1, we have simultaneously for all Q:

$$\mathbb{E}_{w\sim Q}L_D(h_w)\le\mathbb{E}_{w\sim Q}L_S(h_w)+\frac{1}{\lambda}\Big[D_{\mathrm{KL}}(Q\|P)+\ln\tfrac{1}{\delta}+\Psi_{L,P,D}(\lambda,n)\Big].$$

Substituting this bound into the previous identity yields

$$L_D(h_{\overline{w}})\le\mathbb{E}_{w\sim Q}L_S(h_w)+\Big(L_D(h_{\overline{w}})-\mathbb{E}_{w\sim Q}L_D(h_w)\Big)+\frac{1}{\lambda}\Big[D_{\mathrm{KL}}(Q\|P)+\ln\tfrac{1}{\delta}+\Psi_{L,P,D}(\lambda,n)\Big].$$

Next, decompose the Gibbs empirical risk by adding and subtracting $L_S(h_{\overline{w}})$:

$$\mathbb{E}_{w\sim Q}L_S(h_w)=L_S(h_{\overline{w}})+\Big(\mathbb{E}_{w\sim Q}L_S(h_w)-L_S(h_{\overline{w}})\Big).$$

Plugging this decomposition into the last inequality gives

$$\begin{aligned}L_D(h_{\overline{w}})\le L_S(h_{\overline{w}})&+\Big[\Big(L_D(h_{\overline{w}})-\mathbb{E}_{w\sim Q}L_D(h_w)\Big)-\Big(L_S(h_{\overline{w}})-\mathbb{E}_{w\sim Q}L_S(h_w)\Big)\Big]\\&+\frac{1}{\lambda}\Big[D_{\mathrm{KL}}(Q\|P)+\ln\tfrac{1}{\delta}+\Psi_{L,P,D}(\lambda,n)\Big],\end{aligned}$$

which concludes the proof. ■

Remark 3.4. The term $\Omega_{L,Q,D,S}(n)$ is the exact penalty term needed to obtain a bound on the deterministic classifier at the mean of the distribution Q from a bound on the stochastic classifier.

Corollary 3.5. *Let D be a distribution on $X\times Y$, let $S\sim D^n$, and let L be a loss function. Let Θ be a nonempty subset of probability distributions on $\mathbb{R}^m$. Assume that there exists a function $g(\Theta,L,S,\delta,n)$ such that, for any $\delta\in(0,1]$, with probability at least $1-\delta$ over $S\sim D^n$,*

$$\sup_{Q\in\Theta}\Omega_{L,Q,D,S}(n)\ \le\ g(\Theta,L,S,\delta,n).$$

Fix $\delta\in(0,1]$, $\lambda>0$, and a prior P on $\mathbb{R}^m$. Then, with probability at least $1-\delta$ over $S\sim D^n$, we have simultaneously for all $Q\in\Theta$,

$$L_D(h_{\overline{w}})\le L_S(h_{\overline{w}})+g(\Theta,L,S,\delta/2,n)+\frac{1}{\lambda}\Big[D_{\mathrm{KL}}(Q\|P)+\ln\tfrac{2}{\delta}+\Psi_{L,P,D}(\lambda,n)\Big].$$

Proof: This is a direct union bound. ■

Corollary 3.5 reduces the problem of controlling deterministic predictors to obtaining a uniform upper bound on $\Omega_{L,Q,D,S}(n)$ over a family of posteriors. Such bounds will be established in Section 6. In the next definition, we formalize the key object that will control this penalty term: the Jensen gap.

Definition 3.6. Let $f : \mathbb{R}^m \to \mathbb{R}^n$ be a measurable function and let Q be a probability distribution on $\mathbb{R}^m$. We will refer to

$$\mathrm{Jen}_Q[f(w)] := f\left(\mathbb{E}_{w\sim Q}[w]\right) - \mathbb{E}_{w\sim Q}[f(w)]$$

as the *Jensen gap* of f with respect to Q. The quantity $\|\mathrm{Jen}_Q[f(w)]\|$ (or $|\mathrm{Jen}_Q[f(w)]|$) is a measure of non-linearity of the function f. Note that f is not required to be convex.

Combining Corollary 3.5 with Proposition 3.2 and Proposition 6.6 from Section 6 leads to the following result:

Theorem 3.7. *Assume that $L(z, y)$ is a real-valued loss function such that, for every y, the map $z \mapsto L(z, y)$ is ℓ-Lipschitz and β-smooth, and that $L(0, y)$ is independent of y. Let D be a distribution over $X \times Y$ and let $S = \{(x_i, y_i)\}_{i=1}^n \sim D^n$.*

Let $\mathscr{H} = \{h_w : \ w \in \mathbb{R}^m\}$ be a hypothesis class. Assume that for every $h_w \in \mathscr{H}$ there exists $a_w > 0$ such that

$$\|h_w(x)\| \le a_w \qquad \text{for all } x \in X.$$

Let

$$\Theta := \{\overline{w} \in \mathbb{R}^m : \ \|\overline{w}\|_2 \le R\}$$

and fix $\sigma > 0$. For each $\overline{w} \in \Theta$, define

$$Q := \mathscr{N}(\overline{w}, \sigma^2 I_m), \qquad \text{so that} \qquad \overline{w} = \mathbb{E}_{w\sim Q}[w].$$

Assume that the hypotheses of Corollary 6.4 (Section 6) hold, and assume moreover that there exists $B > 0$ such that, for all (x, y) and all $\overline{w} \in \Theta$,

$$|\mathrm{Jen}_Q\big[L(h_w(x), y)\big]| \le \sigma^2 B.$$

In Appendix A, we provide such bounds for smooth neural networks. Let P be any prior distribution on $\mathbb{R}^m$, let $\delta \in (0, 1]$, and let $\lambda > 0$. Then, with probability at least $1 - \delta$ over $S \sim D^n$, we have uniformly for all $\overline{w} \in \Theta$,

$$\begin{aligned} L_D(h_{\overline{w}}) \ \le \ L_S(h_{\overline{w}}) \ + \ 2c\,\frac{R\sigma}{\sqrt{n}}\sqrt{\beta^2 J_{S,\sigma} + \ell^2 H_{S,\sigma}} \ + \ 8\sigma^2 B\sqrt{\frac{2\ln(8/\delta)}{n}} \\ + \ \frac{1}{\lambda}\left[D_{\mathrm{KL}}(Q\|P) + \ln\left(\frac{2}{\delta}\mathbb{E}_{w\sim P}\exp\left(\frac{\lambda^2\ell^2 a_w^2}{2n}\right)\right)\right]. \end{aligned}$$

where $c > 0$ is an absolute constant, and $J_{S,\sigma}$ and $H_{S,\sigma}$ are defined as in Corollary 6.4 and are empirical measures of flatness depending on the hypothesis class.

Lemma 3.8 (Evaluation of the PAC-Bayes complexity term for Gaussian P)**.** *Let P be a prior distribution on the full parameter space of a T-layer model, denoted $W_{1:T}$. Let $W_T \in \mathbb{R}^{m_T}$ be the weights of the last linear layer, and assume that the marginal distribution of W_T under P is*

$$W_T \sim \mathscr{N}(0, \sigma^2 I_{m_T}), \qquad \sigma > 0.$$

Assume that there exists $M > 0$ such that, for all $x \in X$,

$$\|h_{W_{1:T}}(x)\| \le M\|W_T\|.$$

This holds both in the linear case $h_{W_T}(x) = W_T x$ with $\|x\| \le M$, and in the neural-network case when the representation vector entering the last layer has norm upper bounded by M. Let $\ell > 0$, $n \ge 1$, and choose

$$\lambda = \frac{\sqrt{n}}{M\ell\sigma\sqrt{m_T + 1}}.$$

Then, with $a_{W_{1:T}} := M\|W_T\|$*, we have*

$$\frac{1}{\lambda}\ln \mathbb{E}_{W_{1:T}\sim P}\exp\left(\frac{\lambda^2\ell^2 a_{W_{1:T}}^2}{2n}\right) \le \frac{M\ell\sigma\sqrt{m_T+1}}{2\sqrt{n}}.$$

Lemma 3.8 turns the PAC-Bayes logarithmic term into an explicit expression when the prior is Gaussian on the last layer. We now illustrate how the full derandomized bound simplifies in the linear case.

4 Specialization of the results to linear hypothesis classes

Theorem 4.1 (Linear case of Theorem 3.7)**.** *Assume that each hypothesis* $h_W \in \mathscr{H}$ *is given by* $h_W(x) = Wx$ *with* $W \in \mathbb{R}^{C\times d}$ *and let* $m := Cd$*. Assume that* $\|x\| \le M$ *for all* $x \in X$*. Let* $L(z, y)$ *be a real-valued loss function such that, for every* y*, the map* $z \mapsto L(z, y)$ *is* ℓ*-Lipschitz and* β*-smooth, and assume that* $L(0, y)$ *is independent of* y*. Let*

$$\Theta := \{\, W \in \mathbb{R}^{C\times d} : \ \|W\|_F \le R \,\}.$$

Fix $\sigma > 0$ *and let* $\delta \in (0, 1]$*. Then, with probability at least* $1-\delta$ *over* $S \sim D^n$*, we have uniformly for all* $W \in \Theta$*,*

$$\begin{aligned} L_D(h_W) \ &\le L_S(h_W) \ + \ 2\sqrt{2}\,\frac{\beta R\sigma C M^2}{\sqrt{n}} \ + \ 4\beta C M^2\sigma^2\sqrt{\frac{2\ln(8/\delta)}{n}} \\ &+ \ \frac{M\ell\sqrt{m+1}}{2\sigma\sqrt{n}}\,\|W\|_F^2 \ + \ \frac{M\ell\sigma\sqrt{m+1}}{\sqrt{n}}\left(\ln\frac{2}{\delta}+\frac{1}{2}\right). \end{aligned}$$

Remark 4.2. In the scalar-output case $C = 1$, Theorem 4.1 is also true but can be sharpened by replacing Maurer's vector-contraction inequality with the scalar contraction lemma in the analysis. This removes the $\sqrt{2}$ constant in the Rademacher complexity term. Consequently, the first complexity term in Theorem 4.1 becomes

$$2\,\frac{\beta R\sigma M^2}{\sqrt{n}}$$

instead of

$$2\sqrt{2}\,\frac{\beta R\sigma C M^2}{\sqrt{n}}.$$

We now want to apply this result to the multi-class unhinged loss function. We first define a proper surrogate version of the loss that can be used within our generalization bounds.

Lemma 4.3. *Let* $C > 2$ *and* $r > 0$*. Assume that* $\|z\| \le r$*. Define the multi-class unhinged surrogate as*

$$L_r^{\text{unh}}(z, y) := a\left(-z_y + \frac{1}{C}\sum_{k=1}^{C} z_k\right) + b$$

with

$$a = \frac{\sqrt{C}}{r\big(\sqrt{C-1}-\sqrt{\frac{C-2}{2}}\big)}, \qquad b = ar\sqrt{\frac{C-1}{C}} = \frac{1}{1-\sqrt{\frac{C-2}{2(C-1)}}}.$$

Then, we have

1. $\inf_{\|z\|\le r} L_r^{\text{unh}}(z, y) = 0$*,*
2. $L^{0\text{-}1}(z, y) \le L_r^{\text{unh}}(z, y)$*,*
3. *The Lipschitz constant of* $L_r^{\text{unh}}(z, y)$ *is*

$$\frac{1}{r\left(1-\sqrt{\frac{C-2}{2(C-1)}}\right)}.$$

The multi-class unhinged loss is typically studied in the context of learning with uniform label noise; see, for example, Authors (2026). Given a distribution D on $X \times Y$ and a corruption probability $p \in [0, 1)$, let $\overline{D}$ denote the distribution obtained by retaining each label with probability $1 - p$ and, with probability p, drawing its observed value uniformly from the C classes. We state below a result for this setting.

Proposition 4.4. *Let $C > 2$ be the number of classes. Define*

$$b_C := \frac{1}{1 - \sqrt{\frac{C-2}{2(C-1)}}}.$$

Assume that $\|x\| \leq M$. Furthermore, assume that each hypothesis $h_W \in \mathscr{H}$ is given by $h_W(x) = Wx$, where $W \in \mathbb{R}^{C \times d}$. We will simplify notations by replacing h_W with W. Let μ_S^{unh} be the unhinged multi-class data centroid (see Authors (2026)), defined by:

$$\mu_S^{unh} := \tfrac{1}{N} \textstyle\sum_{i=1}^{N} c_{y_i} x_i^\top,$$

where c_y is a column vector having its yth entry given by $\frac{C-1}{C}$ and every other entry given by $\frac{-1}{C}$. Given a distribution D over $X \times Y$, a probability of corruption $p \in [0, 1)$ and a confidence parameter $\delta \in (0, 1]$, with probability at least $1 - \delta$ over the choice of $\overline{S} \sim \overline{D}^n$, we have:

$$L_D^{0\text{-}1}(\mu_{\overline{S}}^{unh}) \leq b_C \left[1 - \sqrt{\frac{C}{C-1}} \frac{\|\mu_{\overline{S}}^{\text{unh}}\|}{(1-p)M}\right] + \tilde{O}\left(\frac{b_C \sqrt{dC}}{(1-p)\sqrt{n}}\right).$$

We can also apply our results to the problem of linear regression.

Corollary 4.5 (Linear regression with absolute loss)**.** *Let $h_w(x) = \langle w, x \rangle$, $\|x\| \leq M$ and*

$$L(z, y) = |z - y|.$$

Assume that $|y| \leq Y_{\max}$ almost surely. Let $P = \mathscr{N}(0, \sigma^2 I_d)$ with $\sigma > 0$. Then, with probability at least $1 - \delta$ over $S \sim D^n$, for all probability distributions Q on $\mathbb{R}^d$,

$$\mathbb{E}_{w \sim Q} L_D(h_w) \leq \mathbb{E}_{w \sim Q} L_S(h_w) + \frac{M\sigma\sqrt{d+1}}{\sqrt{n}} \left[D_{\text{KL}}(Q\|P) + \ln \frac{2}{\delta} + \frac{1}{2}\right] + Y_{\max} \sqrt{\frac{\ln(2/\delta)}{2n}}.$$

In particular, the complexity term has a square root dependence on the dimension d.

Since the absolute loss is not smooth, our derandomized framework does not apply to it directly. Nevertheless, it can be approached by the following smooth family of Huber losses, which converges uniformly to the absolute loss as the smoothing parameter tends to zero.

Corollary 4.6 (Linear regression with Huber losses)**.** *Let $h_w(x) = \langle w, x \rangle$, assume that $\|x\| \leq M$, and fix $\rho > 0$. Define the Huber loss by*

$$L_\rho(z, y) := \begin{cases} \dfrac{(z-y)^2}{2\rho}, & \text{if } |z - y| \leq \rho, \\ |z - y| - \dfrac{\rho}{2}, & \text{if } |z - y| > \rho. \end{cases}$$

Assume that $|y| \leq Y_{\max}$ almost surely and define

$$Y_{\rho,\max} := \sup_{|y| \leq Y_{\max}} L_\rho(0, y).$$

Let

$$\Theta := \{w \in \mathbb{R}^d : \|w\| \leq R\}$$

and fix $\sigma > 0$. Then, with probability at least $1 - \delta$ over $S \sim D^n$, we have uniformly for all $w \in \Theta$,

$$L_{\rho,D}(h_w) \leq L_{\rho,S}(h_w) + \frac{2\,R\sigma M^2}{\rho\sqrt{n}} + \frac{4M^2\sigma^2}{\rho}\sqrt{\frac{2\ln(16/\delta)}{n}} + Y_{\rho,\max}\sqrt{\frac{\ln(2/\delta)}{2n}}$$
$$+ \frac{M\sqrt{d+1}}{2\sigma\sqrt{n}}\|w\|^2 + \frac{M\sigma\sqrt{d+1}}{\sqrt{n}}\left(\ln\frac{4}{\delta} + \frac{1}{2}\right).$$

5 Generalization bounds for bounded loss functions from the PAC-Bayes framework

When the loss is uniformly bounded, one can instead invoke standard PAC-Bayes bounds for bounded loss functions. Using the same proof as in the case of unbounded loss functions, but replacing Theorem 3.1 with Theorem 31.1 of Shalev-Shwartz & Ben-David (2014) applied to the rescaled loss $(L - A')/(A - A')$, we obtain the following result for bounded loss functions:

Theorem 5.1. *Assume that the loss $L(z, y)$ takes values in an interval $[A', A] \subset \mathbb{R}$ on the domain $\mathbb{R}^C$, and that, for every y, the map $z \mapsto L(z, y)$ is ℓ-Lipschitz and β-smooth. Let D be a distribution over $X \times Y$ and let $S = \{(x_i, y_i)\}_{i=1}^n \sim D^n$. Let $\mathscr{H} = \{h_w : \ w \in \mathbb{R}^m\}$ be a hypothesis class. Let*

$$\Theta := \{\overline{w} \in \mathbb{R}^m : \ \|\overline{w}\|_2 \leq R\},$$

and fix $\sigma > 0$. For any $\overline{w} \in \Theta$, define a posterior $Q := \mathscr{N}(\overline{w}, \sigma^2 I_m)$. Assume that the hypotheses of Corollary 6.4 hold, and assume moreover that there exists $B > 0$ such that, for all (x, y) and all $\overline{w} \in \Theta$,

$$|\mathrm{Jen}_Q\big[L(h_w(x), y)\big]| \leq \sigma^2 B.$$

Let P be any prior distribution on $\mathbb{R}^m$, and let $\delta \in (0, 1)$. Then, with probability at least $1 - \delta$ over $S \sim D^n$, we have uniformly for all $\overline{w} \in \Theta$,

$$\begin{aligned} L_D(h_{\overline{w}}) \ \leq L_S(h_{\overline{w}}) \ + \ 2c\,\frac{R\sigma}{\sqrt{n}}\sqrt{\beta^2 J_{S,\sigma} + \ell^2 H_{S,\sigma}} \ + \ 8\sigma^2 B\sqrt{\frac{2\ln(8/\delta)}{n}} \\ + \ (A - A')\sqrt{\frac{D_{\mathrm{KL}}(Q\|P) + \ln\big(\frac{2n}{\delta}\big)}{2(n-1)}}\ , \end{aligned}$$

where $c > 0$ is an absolute constant, and $J_{S,\sigma}$ and $H_{S,\sigma}$ are defined as in Corollary 6.4.

Similarly to Theorem 4.1, we can obtain a straightforward specialization to the linear case:

Theorem 5.2 (Linear case of Theorem 5.1)**.** *Assume that each hypothesis $h_W \in \mathscr{H}$ is given by $h_W(x) = Wx$ with $W \in \mathbb{R}^{C \times d}$ and let $m := Cd$. Assume that $\|x\| \leq M$ for all $x \in X$. Let $L(z, y)$ be a loss function taking values in an interval $[A', A] \subset \mathbb{R}$ on the domain $\mathbb{R}^C$, and assume that, for every y, the map $z \mapsto L(z, y)$ is ℓ-Lipschitz and β-smooth. Let*

$$\Theta := \{\, W \in \mathbb{R}^{C \times d} : \ \|W\|_F \leq R\,\}.$$

Fix $\sigma > 0$ and let $\delta \in (0, 1)$. Then, with probability at least $1 - \delta$ over $S \sim D^n$, we have uniformly for all $W \in \Theta$,

$$\begin{aligned} L_D(h_W) \ \leq \ L_S(h_W) \ + \ 2\sqrt{2}\,\frac{\beta R\sigma C M^2}{\sqrt{n}} \ + \ 4\beta C M^2 \sigma^2 \sqrt{\frac{2\ln(8/\delta)}{n}} \\ + \ (A - A')\sqrt{\frac{\frac{\|W\|_F^2}{2\sigma^2} + \ln\big(\frac{2n}{\delta}\big)}{2(n-1)}}\ . \end{aligned}$$

One can alternatively invoke Theorem 5 of Maurer (2004), which provides an inverse-kl PAC-Bayesian bound for bounded losses. This may lead to a sharper complexity term, but at the cost of an implicit final expression that is less transparent. We obtain the following variant of Theorem 5.1.

Theorem 5.3 (Bounded-loss PAC-Bayes bound in kl^{-1} form, uniform over $\{\sigma_i\}_{i \geq 1}$)**.** *Assume that the loss $L(z, y)$ takes values in an interval $[A', A] \subset \mathbb{R}$ on the domain $\mathbb{R}^C$, and that, for every y, the map $z \mapsto L(z, y)$ is ℓ-Lipschitz and β-smooth. Let D be a distribution over $X \times Y$, let $S = \{(x_j, y_j)\}_{j=1}^n \sim D^n$, with $n \geq 8$, and let*

$$\Theta := \{\overline{w} \in \mathbb{R}^m : \ \|\overline{w}\|_2 \leq R\}.$$

Fix $\sigma_{\max} > 0$ and for $i \geq 1$ define

$$\sigma_i := 2^{\,1-i}\sigma_{\max}.$$

For $\overline{w} \in \Theta$, let

$$Q_{\overline{w},\sigma_i} := \mathscr{N}(\overline{w}, \sigma_i^2 I_m).$$

Assume the same conditions as in Corollary 6.4. Assume moreover that there exists $B > 0$ such that, for all (x, y), all $\overline{w} \in \Theta$ and all $i \geq 1$,

$$|\mathrm{Jen}_{Q_{\overline{w},\sigma_i}}\big[L(h_w(x), y)\big]| \leq \sigma_i^2 B.$$

Let P be any prior distribution on $\mathbb{R}^m$ and let $\delta \in (0, 1)$. Then, with probability at least $1 - \delta$ over $S \sim D^n$, we have uniformly for all $i \geq 1$ and all $\overline{w} \in \Theta$,

$$L_D(h_{\overline{w}}) \leq 2c\,\frac{R\sigma_i}{\sqrt{n}}\sqrt{\beta^2 J_{S,\sigma_i} + \ell^2 H_{S,\sigma_i}} + 8\sigma_i^2 B\sqrt{\frac{4\ln(i) + 2\ln(16/\delta)}{n}} + A' + \mathrm{Jen}_{Q_{\overline{w},\sigma_i}}\big[L_S(h_w)\big]$$
$$+\ (A - A')\ \mathrm{kl}^{-1}\left(\frac{\mathbb{E}_{w\sim Q_{\overline{w},\sigma_i}} L_S(h_w) - A'}{A - A'}\ \middle|\ \frac{D_{\mathrm{KL}}(Q_{\overline{w},\sigma_i}\|P) + \ln\big(\frac{4\sqrt{n}}{\delta}\big)}{n}\right),$$

where

$$\mathrm{kl}^{-1}(q \mid \varepsilon) := \sup\Big\{p \in [q, 1] :\ \mathrm{kl}(q\|p) \leq \varepsilon\Big\}$$

with

$$\mathrm{kl}(q\|p) := q\ln\frac{q}{p} + (1 - q)\ln\frac{1 - q}{1 - p}.$$

Proof: Apply Theorem 5 of Maurer (2004) to the rescaled loss $(L-A')/(A-A')$ and then use Proposition 6.7 with a union bound after rearranging terms. ■

The results of Sections 3–5 are stated in terms of quantities measuring the complexity of the Jensen gap class. The Rademacher complexity bounds used to control these quantities are established next.

6 Generalization bounds for the Jensen gap from Rademacher complexity

The remaining technical task is to control the Rademacher complexity of the Jensen gap class. In the linear setting, this is achieved using a multi-class vector contraction inequality (Lemma C.1 in Appendix C). In the nonlinear setting, we instead use a Talagrand comparison inequality (Proposition C.2) to compare the Rademacher process with a Gaussian process whose metric is controlled by Jacobian and Hessian quantities. The proofs of the results in this section are deferred to Appendix C.

We first apply the vector contraction inequality to obtain the following bound for linear multiclass predictors.

Theorem 6.1. *Let $S = \{(x_i, y_i)\}_{i=1}^n$ with $x_i \in \mathbb{R}^d$ and $y_i \in \{1, \cdots, C\}$. Consider the linear multiclass score function $h_W(x) = Wx$ with $W \in \mathbb{R}^{C\times d}$ and $\Theta := \{W \in \mathbb{R}^{C\times d} :\ \|W\|_F \leq R\}$. Assume that for any y, the map $z \mapsto L(z, y)$ is differentiable and β-smooth. For $\sigma > 0$, let Q be an isotropic Gaussian distribution with mean $\overline{W}$ and covariance matrix $\sigma^2 I_{Cd\times Cd}$. Define*

$$F_\sigma(\overline{W}; x, y) := L(h_{\overline{W}}(x), y) - \mathbb{E}_{W\sim Q}\big[L(h_W(x), y)\big]$$

and let $\mathscr{F}_\sigma := \{(x, y) \mapsto F_\sigma(\overline{W}; x, y) :\ \overline{W} \in \Theta\}$. Then, the empirical Rademacher complexity of $\mathscr{F}_\sigma$ satisfies

$$\mathfrak{R}_S(\mathscr{F}_\sigma)\ \leq\ \frac{\sqrt{2}\beta R\sigma C}{n}\sqrt{\sum_{i=1}^n \|x_i\|^4}.$$

In particular, if $\|x_i\| \leq M$ for all i, then

$$\mathfrak{R}_S(\mathscr{F}_\sigma) \leq \frac{\sqrt{2}\beta R\sigma C\, M^2}{\sqrt{n}}.$$

For bounding the Rademacher complexity of the Jensen gap class for non-linear neural networks, we will need more advanced techniques than a contraction argument. The next lemma is the key technical result. It bounds the pseudo-metric induced from the Jensen gap (defined in the lemma) with a pseudo-metric involving the properties of the loss function and measures of sensitivity to noise (first order and second order) of the neural networks in the class.

Lemma 6.2 (Bounding the metric induced by the Jensen gap)**.** *Let $S = \{(x_i, y_i)\}_{i=1}^n$ and let $h_w : \mathscr{X} \to \mathbb{R}^C$ be a neural network parameterized by $w \in \Theta \subset \mathbb{R}^m$. Assume that Θ is convex. Fix $\sigma > 0$ and let $u \sim \mathscr{N}(0, I_m)$. For each $i \in \{1, \dots, n\}$ define*

$$F_i(w) := L(h_w(x_i), y_i) - \mathbb{E}_u[L(h_{w+\sigma u}(x_i), y_i)]\,,$$

and the sample-dependent pseudo-metric

$$d_S(w_1, w_2) := \left(\sum_{i=1}^n (F_i(w_1) - F_i(w_2))^2 \right)^{1/2}.$$

Assume that for each y, the map $z \mapsto L(z, y)$ is differentiable, ℓ-Lipschitz and β-smooth. Assume moreover that, for each i, the map $w \mapsto h_w(x_i)$ is twice continuously differentiable on $\mathbb{R}^m$, and that the Gaussian expectations below are finite and may be differentiated under the expectation sign. Write

$$J_i(w) := \nabla_w h_w(x_i) \in \mathbb{R}^{C \times m}, \qquad H_{i,k}(w) := \nabla_w^2 \big(h_w(x_i)\big)_k \in \mathbb{R}^{m \times m}.$$

Assume that there exist sample-dependent positive semi-definite matrices $G_i \succeq 0$ such that, for all $w \in \Theta$,

$$J_i(w)^\top J_i(w) \preceq G_i.$$

Define

$$A_{i,\sigma} := \sup_{w \in \Theta} \int_0^1 \mathbb{E}_u \big[u^\top J_i(w + t\sigma u)^\top J_i(w + t\sigma u) u \big]\, dt,$$

and assume that $A_{i,\sigma} < \infty$ for all i. For each $w \in \Theta$, define the Gaussian averaged Hessian matrix

$$B_{i,\sigma}(w) := \int_0^1 \mathbb{E}_u \left[\sum_{k=1}^C H_{i,k}(w + t\sigma u)^\top u u^\top H_{i,k}(w + t\sigma u) \right] dt.$$

Assume that there exist sample-dependent positive semi-definite matrices $B_{i,\sigma} \succeq 0$ such that, for all $w \in \Theta$,

$$B_{i,\sigma}(w) \preceq B_{i,\sigma}.$$

Define the sample-dependent matrix

$$M_{S,\sigma} := 2\sigma^2 \sum_{i=1}^n \Big(\beta^2 A_{i,\sigma} G_i + \ell^2 B_{i,\sigma} \Big).$$

Then, for all $w_1, w_2 \in \Theta$,

$$d_S(w_1, w_2)^2 \le (w_1 - w_2)^\top M_{S,\sigma} (w_1 - w_2).$$

It follows from Talagrand-type comparison inequalities (Talagrand, 2005) that we can bound the Rademacher complexity of the Jensen gap class from the trace of the matrix $M_{S,\sigma}$.

Proposition 6.3. *Let $S = \{(x_i, y_i)\}_{i=1}^n$ be a sample and let*

$$\Theta := \{\, w \in \mathbb{R}^m : \|w\|_2 \le R \,\}.$$

Let $f : \Theta \times (\mathscr{X} \times \mathscr{Y}) \to \mathbb{R}$ and define

$$\mathscr{F} := \{\, (x, y) \mapsto f(w, (x, y)) : w \in \Theta \,\}.$$

Define the sample-dependent pseudo-metric

$$d_S(w_1, w_2) := \left(\sum_{i=1}^n \big(f(w_1, (x_i, y_i)) - f(w_2, (x_i, y_i)) \big)^2 \right)^{1/2}.$$

Assume there exists a sample-dependent positive semi-definite matrix $M_S \in \mathbb{R}^{m\times m}$ such that

$$d_S(w_1, w_2) \;\leq\; \sqrt{(w_1 - w_2)^\top M_S (w_1 - w_2)} \qquad \forall w_1, w_2 \in \Theta.$$

Then, the empirical Rademacher complexity of $\mathscr{F}$ satisfies

$$\mathfrak{R}_S(\mathscr{F}) \;\leq\; c\,\frac{R}{n}\sqrt{\mathrm{Trace}(M_S)},$$

where $c > 0$ is a constant.

We can now provide an upper bound on the Rademacher complexity of the Jensen gap class for neural networks.

Corollary 6.4 (Rademacher complexity of the Jensen gap class)**.** *Let $S = \{(x_i, y_i)\}_{i=1}^n$ and let*

$$\Theta := \{\, w \in \mathbb{R}^m : \|w\|_2 \leq R \,\}.$$

Fix $\sigma > 0$ and let $\mathscr{F}_\sigma$ be the Jensen gap class for the loss function $L(h_w(x), y)$. Assume that for each y, the map $z \mapsto L(z, y)$ is differentiable, ℓ-Lipschitz and β-smooth. Assume moreover that the hypotheses of Lemma 6.2 hold. Define

$$J_{S,\sigma} := \frac{1}{n}\sum_{i=1}^n A_{i,\sigma}\mathrm{Trace}(G_i), \qquad H_{S,\sigma} := \frac{1}{n}\sum_{i=1}^n \mathrm{Trace}(B_{i,\sigma}).$$

Then, the empirical Rademacher complexity of $\mathscr{F}_\sigma$ satisfies

$$\mathfrak{R}_S(\mathscr{F}_\sigma) \leq c\,\frac{R\sigma}{\sqrt{n}}\sqrt{\beta^2 J_{S,\sigma} + \ell^2 H_{S,\sigma}},$$

where $c > 0$ is a universal constant.

Remark 6.5. As a sanity check of the result, consider the case of linear hypothesis classes ($h_W(x) = Wx$). Then, the Hessian term $H_{S,\sigma}$ is zero and the Jacobian term $J_{S,\sigma}$ is easily seen to be bounded by C^2M^4 if $||x|| \leq M$. Therefore, Corollary 6.4 implies

$$\mathfrak{R}_S(\mathscr{F}_\sigma) \;\leq\; c\,\frac{R\sigma\beta C M^2}{\sqrt{n}}.$$

This is exactly the same result as the one obtained in Theorem 6.1 up to a constant.

We are now ready to state the main result of this section.

Proposition 6.6 (Uniform generalization bound for the Jensen gap class)**.** *Let $S = \{(x_i, y_i)\}_{i=1}^n \sim D^n$ and let*

$$\Theta := \{\, \overline{w} \in \mathbb{R}^m : \|\overline{w}\|_2 \leq R \,\}.$$

Fix $\sigma > 0$ and let Q denote an isotropic Gaussian distribution

$$Q := \mathscr{N}(\overline{w}, \sigma^2 I_m), \qquad \overline{w} \in \Theta.$$

Consider the same assumptions as in Corollary 6.4. Assume moreover that there exists $B > 0$ such that for all (x, y) and all $\overline{w} \in \Theta$,

$$|\mathrm{Jen}_Q\big[L(h_w(x), y)\big]| \leq \sigma^2 B.$$

Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over $S \sim D^n$, we have uniformly for all $\overline{w} \in \Theta$,

$$\begin{aligned}\mathbb{E}_{(x,y)\sim D}\Big[\mathrm{Jen}_Q(L(h_w(x), y))\Big] \leq\; & \frac{1}{n}\sum_{i=1}^n \mathrm{Jen}_Q(L(h_w(x_i), y_i)) \\ & +\; 2c\,\frac{R\sigma}{\sqrt{n}}\sqrt{\beta^2 J_{S,\sigma} + \ell^2 H_{S,\sigma}} \;+\; 8\sigma^2 B\sqrt{\frac{2\ln(4/\delta)}{n}},\end{aligned}$$

where $c > 0$ is an absolute constant.

If needed, a standard argument can be used to obtain a result holding uniformly over a set of standard deviations $\{\sigma_i\}$.

Proposition 6.7 (Uniform generalization bound on $\{\sigma_i\}_{i\geq 1}$)**.** *Let $S = \{(x_j, y_j)\}_{j=1}^n \sim D^n$ and let*

$$\Theta := \{\, \overline{w} \in \mathbb{R}^m : \|\overline{w}\|_2 \leq R \,\}.$$

Fix $\sigma_{\max} > 0$ and for $i \geq 1$ define

$$\sigma_i := 2^{\,1-i}\sigma_{\max}.$$

For $\overline{w} \in \Theta$ and $i \geq 1$, let

$$Q_{\overline{w},\sigma_i} := \mathscr{N}(\overline{w}, \sigma_i^2 I_m).$$

Assume the same conditions as in Corollary 6.4. Assume moreover that there exists $B > 0$ such that for all (x, y), all $\overline{w} \in \Theta$ and all $i \geq 1$,

$$|\mathrm{Jen}_{Q_{\overline{w},\sigma_i}}\big[L(h_w(x), y)\big]| \leq \sigma_i^2 B.$$

Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over $S \sim D^n$, we have uniformly for all $i \geq 1$ and all $\overline{w} \in \Theta$,

$$\mathbb{E}_{(x,y)\sim D}\Big[\mathrm{Jen}_{Q_{\overline{w},\sigma_i}}\big(L(h_w(x), y)\big)\Big] \leq \frac{1}{n}\sum_{j=1}^{n} \mathrm{Jen}_{Q_{\overline{w},\sigma_i}}\big(L(h_w(x_j), y_j)\big) + 2c\,\frac{R\sigma_i}{\sqrt{n}}\sqrt{\beta^2 J_{S,\sigma_i} + \ell^2 H_{S,\sigma_i}}$$
$$+ 8\sigma_i^2 B\sqrt{\frac{4\ln(i) + 2\ln(8/\delta)}{n}}.$$

Remark 6.8 (From Gaussian averaged controls to simpler global controls). The quantities $A_{i,\sigma}$ and $B_{i,\sigma}$ appearing above are averaged along the perturbation path $w + t\sigma u$. This is the appropriate form for ordinary Gaussian posteriors, since the perturbation is not confined to the original parameter set Θ. However, in settings where the posterior perturbation has bounded support, we can obtain a simpler expression. Suppose that the perturbed parameter is supported in a larger parameter set $\widetilde{\Theta}$, in the sense that

$$w + t\sigma u \in \widetilde{\Theta} \qquad \text{for all } w \in \Theta,\ t \in [0, 1],$$

almost surely. Assume also that $\mathbb{E}[uu^\top] \preceq I_m$, and that, for each i, there exists a positive semidefinite matrix $G_i \succeq 0$ such that

$$J_i(v)^\top J_i(v) \preceq G_i \qquad \text{for all } v \in \widetilde{\Theta}.$$

Then, the Gaussian averaged Jacobian quantity satisfies

$$A_{i,\sigma} = \sup_{w\in\Theta} \int_0^1 \mathbb{E}_u\big[u^\top J_i(w + t\sigma u)^\top J_i(w + t\sigma u)u\big]\, dt$$
$$\leq \mathbb{E}_u[u^\top G_i u] = \mathrm{Tr}\big(G_i\, \mathbb{E}[uu^\top]\big) \leq \mathrm{Tr}(G_i).$$

Consequently,

$$J_{S,\sigma} = \frac{1}{n}\sum_{i=1}^{n} A_{i,\sigma}\,\mathrm{Tr}(G_i) \leq \frac{1}{n}\sum_{i=1}^{n} \mathrm{Tr}(G_i)^2.$$

Similarly, suppose that there exist positive semidefinite matrices $M_i \succeq 0$ such that

$$\sum_{k=1}^{C} H_{i,k}(v)^\top H_{i,k}(v) \preceq M_i \qquad \text{for all } v \in \widetilde{\Theta}.$$

Then, the proof of Lemma 6.2 can be modified by using this global Hessian control directly in the estimate of the second order term. In that setting, the Hessian contribution in the final Rademacher bound becomes

$$H_S := \frac{1}{n}\sum_{i=1}^{n} \mathrm{Tr}(M_i).$$

Therefore, under bounded support perturbations, Corollary 6.4 implies the simpler bound

$$\mathfrak{R}_S(\mathscr{F}_\sigma) \le c\,\frac{R\sigma}{\sqrt{n}}\sqrt{\beta^2 J_S + \ell^2 H_S},$$

where

$$J_S := \frac{1}{n}\sum_{i=1}^{n}\mathrm{Tr}(G_i)^2, \qquad H_S := \frac{1}{n}\sum_{i=1}^{n}\mathrm{Tr}(M_i).$$

For ordinary Gaussian posteriors, the perturbation has unbounded support, so the bounded support simplification does not apply directly. A possible extension would be to use a localization argument: control the Jacobian and Hessian on a larger set $\widetilde{\Theta}$ that contains the perturbation path with high probability and then bound the contribution of the complementary event separately. This would lead to a global bound as above together with a residual term.

The practical regularizer introduced in Section 9 takes a much simpler route. Instead of keeping the supremum over parameters and the Gaussian average along the perturbation path, it evaluates empirical Jacobian and Hessian quantities at the current effective parameters. Thus, it should be viewed as a convenient proxy inspired by the structure of the bound, rather than as a direct evaluation of $J_{S,\sigma}$ and $H_{S,\sigma}$. In the SHEL analysis of Section 8, we instead retain the quantities averaged over Gaussian perturbations, while replacing each supremum by the largest realization observed during training.

Before turning to nonlinear SHEL networks, we first compare the resulting linear bounds with a classical Rademacher complexity bound.

7 Bound comparisons for linear classifiers

When the loss function is bounded, both the bounded and unbounded variants of the derandomized PAC-Bayes bounds can be applied. We first introduce a convenient bounded smooth multi-class loss obtained by extending the sigmoid loss.

Lemma 7.1 (Surrogate multiclass sigmoid loss function)**.** *Fix $R > 0$ and $C > 2$, and let $\sigma(z) := (1+e^{-z})^{-1}$. For $z \in \mathbb{R}^C$ and $y \in \{1,\dots,C\}$, define*

$$L(z,y) := \sum_{k\neq y}\sigma(z_k - z_y).$$

Let

$$m_R := (C-1)\sigma\left(-R\sqrt{\frac{C}{C-1}}\right), \qquad q_R := \frac{1}{2} + (C-2)\sigma\left(-R\sqrt{\frac{C}{2(C-2)}}\right),$$

and

$$a(R) := \frac{1}{q_R - m_R}, \qquad b(R) := -a(R)\,m_R.$$

Define the surrogate loss

$$\Phi_R(z,y) := a(R)L(z,y) + b(R).$$

Then the following hold:

(i) On the domain $\|z\| \le R$:

$$L^{0\text{-}1}(z,y) \le \Phi_R(z,y).$$

(ii) The infimum of Φ_R on the domain $\|z\| \le R$ equals 0, whereas on the domain $\mathbb{R}^C$ it equals $b(R)$.

(iii) The supremum of Φ_R on the domain $\|z\| \le R$ equals

$$a(R)(C-1)\big(2\sigma\left(R\sqrt{\frac{C}{C-1}}\right) - 1\big),$$

whereas on the domain $\mathbb{R}^C$ *it equals*

$$a(R)(C-1)\sigma\left(R\sqrt{\frac{C}{C-1}}\right).$$

(iv) Φ_R *is* ℓ_R*-Lipschitz on the domain* $\mathbb{R}^C$ *with*

$$\ell_R = \frac{a(R)}{4}\sqrt{C(C-1)}.$$

(v) Φ_R *is* β_R*-smooth on the domain* $\mathbb{R}^C$ *with*

$$\beta_R \leq \frac{a(R)\,C}{6\sqrt{3}}.$$

In Figure 1, we compare the generalization bound terms obtained from the unbounded derandomized PAC-Bayes bound of Theorem 4.1 with those obtained from the bounded derandomized PAC-Bayes bound of Theorem 5.2, both applied to the multi-class sigmoid surrogate of Lemma 7.1. We set $n = 100000$, $d = 50$ and $\delta = 0.05$, and vary R over the interval $[0.1, 1]$. The bounded version gives a looser bound for very small values of R, but becomes substantially tighter than the unbounded version as R increases. This behavior reflects, in part, the less favorable dimensional dependence of the unbounded bound. The unbounded bound is tighter only in a restricted regime of very small parameter norms and low dimension. Outside this regime, when both results are applicable, the bounded version should be preferred.

We next compare our bounded derandomized PAC-Bayes bounds with a Rademacher complexity bound for the same class of linear predictors. The Rademacher complexity bound is given by the following proposition.

Proposition 7.2. *Let* $S = \{(x_i, y_i)\}_{i=1}^n$*. Assume that* $\|x\| \leq M$ *and consider the hypothesis class*

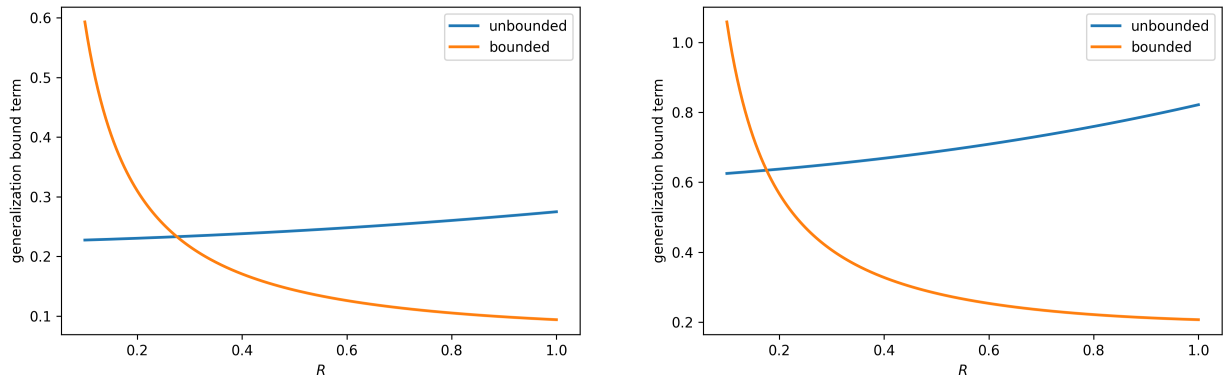
$$\mathscr{F} := \left\{\, x \mapsto Wx \in \mathbb{R}^C : \ \|W\|_F \leq R \,\right\}.$$

Let $L(z, y)$ *take values in* $[0, A]$ *and assume that, for every* y*, the map* $z \mapsto L(z, y)$ *is* ℓ*-Lipschitz. Then, with probability at least* $1 - \delta$ *over* $S \sim D^n$*, we have uniformly for all* W *with* $\|W\|_F \leq R$*,*

$$L_D(h_W) \ \leq \ L_S(h_W) + 2\sqrt{2}\,\ell\,\frac{RM\sqrt{C}}{\sqrt{n}} \ + \ 4A\sqrt{\frac{2\ln(4/\delta)}{n}}.$$

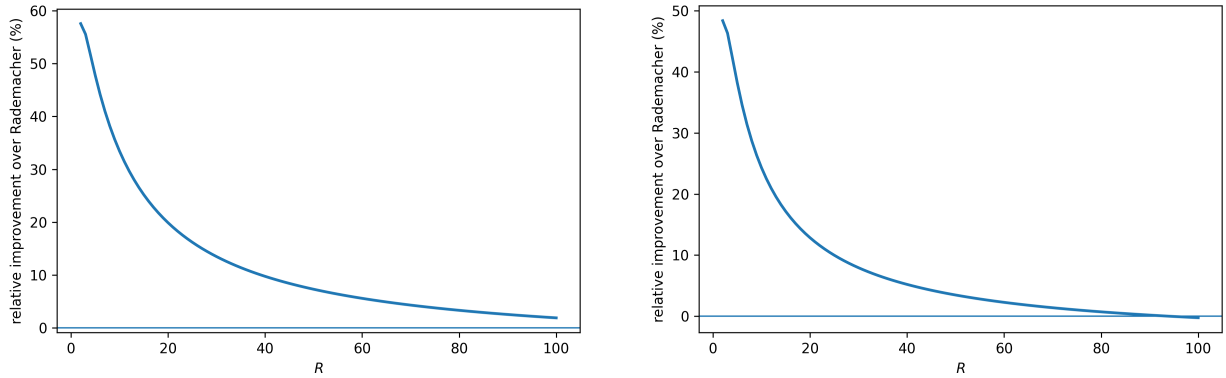
Proof: Immediate from Maurer's vector-contraction inequality applied to $L(\cdot, y)$, Theorem 26.5 (item 2) of Shalev-Shwartz & Ben-David (2014), and the bound $\mathfrak{R}_S(\mathscr{F}) \leq RM\sqrt{C/n}$ for the linear class with $\|W\|_F \leq R$. ■

The relative improvement of our bound in Theorem 5.2 over the Rademacher complexity bound is shown in Figure 2. Our bound is tighter for moderate values of R and becomes comparable to the Rademacher bound for larger values of R. Figure 3 reports a complementary experiment on MNIST. In this case, the kl^{-1} version of Theorem 5.3 remains tighter over the displayed range.



(a) $C = 3$ (b) $C = 10$

Figure 1: Comparison between the bounded and unbounded derandomized PAC-Bayes generalization bound terms for the multi-class sigmoid loss.



(a) $C = 3$ (b) $C = 10$

Figure 2: Relative improvement of our derandomized PAC-Bayes bound over the Rademacher complexity bound as a function of $R \in [2, 100]$ for the multi-class sigmoid loss.

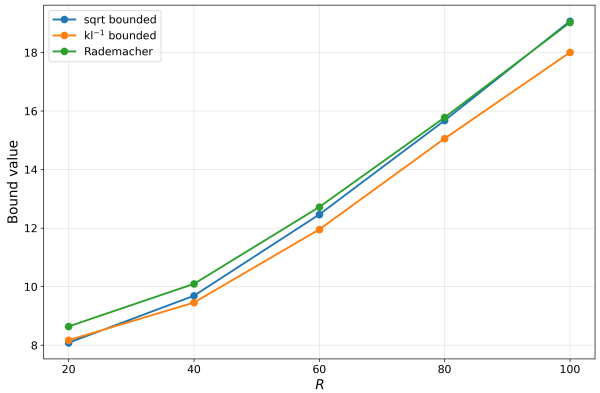


Figure 3: Comparison of the square-root and kl^{-1} derandomized PAC-Bayes bounds with the Rademacher complexity bound for linear classifiers trained with the cross-entropy loss on MNIST, as functions of the parameter radius R. For each R, the classifier is trained for 40 epochs using projected SGD with momentum, with a linear warm-up followed by cosine annealing.

8 Single hidden layer erf (SHEL) networks

We consider a single hidden layer neural network with the error function activation, which may be seen as a Gaussian smoothing of the sign function (Germain et al., 2009; Letarte et al., 2019). Our goal is to compare the resulting bounds with Theorem 5 of Biggs & Guedj (2022). We begin by collecting in the next lemma some quantities needed to evaluate our bounds for the SHEL network.

Lemma 8.1 (Specialization to the SHEL network)**.** *For $x \neq 0$, consider*

$$F_{U,V}(x) = V \ \mathrm{erf}\left(\frac{Ux}{\sqrt{2}\,\|x\|}\right), \qquad U \in \mathbb{R}^{K\times d}, \quad V \in \mathbb{R}^{C\times K},$$

where erf *is applied element-wise and is defined by*

$$\mathrm{erf}(x) := \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt.$$

The Lipschitz constant and β-smoothness of the activation function erf *are given respectively by*

$$\ell_{\mathrm{erf}} := \frac{2}{\sqrt{\pi}}, \qquad \beta_{\mathrm{erf}} := 2\sqrt{\frac{2}{\pi e}}.$$

Assume that $\|U\|_F \leq R_U$, $\|V\|_F \leq R_V$ and that, for every y, the map $z \mapsto L(z,y)$ is ℓ_L-Lipschitz and β_L-smooth. Then,

(i) (Jensen gap upper bound for bounded activation functions) From Theorem A.5 with

$$T = 2, \quad M_0^2 = \frac{1}{2}, \quad M_1^2 = K, \quad d_1 = K, \quad d_2 = C,$$
$$\ell_1 = \ell_{\mathrm{erf}}, \quad \beta_1 = \beta_{\mathrm{erf}}, \quad \ell_2 = \ell_L, \quad \beta_2 = \beta_L.$$

we obtain

$$|\mathrm{Jen}_Q[L(F_{U,V}(x), y)]| \leq \sigma^2 K \left(\frac{\ell_L}{\sqrt{2\pi e}} R_V + \frac{\beta_L}{\pi} R_V^2 + \frac{\beta_L C}{2}\right).$$

(ii) (Jensen gap upper bound for unbounded activation functions) From Theorem A.6 with

$$T = 2, \quad M_0^2 = \frac{1}{2}, \quad d_1 = K, \quad d_2 = C, \quad \ell_1 = \ell_{\mathrm{erf}}, \quad \beta_1 = \beta_{\mathrm{erf}}, \quad \ell_2 = \ell_L, \quad \beta_2 = \beta_L,$$

we obtain

$$|\mathrm{Jen}_Q[L(F_{U,V}(x), y)]| \leq \sigma^2 \left[K\left(\frac{\ell_L}{\sqrt{2\pi e}} R_V + \frac{\beta_L}{\pi} R_V^2\right) + \frac{\beta_L C}{\pi} R_U^2 + \frac{\beta_L C}{\pi} K\sigma^2\right].$$

(iii) (KL upper bound for isotropic Gaussian prior and posterior) If

$$P = \mathscr{N}(0, \sigma^2 I_m), \qquad Q = \mathscr{N}\big((\mathrm{vec}(U), \mathrm{vec}(V)), \sigma^2 I_m\big), \qquad m = K(d + C),$$

then

$$D_{\mathrm{KL}}(Q\|P) = \frac{\|U\|_F^2 + \|V\|_F^2}{2\sigma^2} \leq \frac{R_U^2 + R_V^2}{2\sigma^2}.$$

Since both parts (i) and (ii) apply to the SHEL network, the Jensen gap may be bounded by the minimum of the two corresponding upper bounds. We will compare our results to the following margin based bound.

Theorem 8.2 (Explicit version of Theorem 5 of Biggs & Guedj (2022))**.** *Consider the SHEL network*

$$F_{U,V}(x) = V \operatorname{erf}\left(\frac{Ux}{\sqrt{2}\,\|x\|}\right), \qquad U \in \mathbb{R}^{K\times d}, \quad V \in \mathbb{R}^{C\times K},$$

and let

$$V_\infty := \max_{i,j} |V_{ij}|.$$

Then, with probability at least $1-\delta$ *over* $S = \{(x_j, y_j)\}_{j=1}^n \sim D^n$*, we have uniformly for all* $\gamma \in (0, V_\infty K)$*,*

$$L_D^{0\text{-}1}(F_{U,V}) \le \frac{1}{n} + \mathrm{kl}^{-1}\left(L_{S,\gamma}(F_{U,V}) + \frac{1}{n} \,\middle|\, \varepsilon_{U,V,\gamma,\delta,n}\right),$$

where

$$\varepsilon_{U,V,\gamma,\delta,n} := \frac{17\ln n}{n}\left(\frac{2V_\infty K}{\gamma}\right)^2 \left(\frac{\|U-U^0\|_F^2}{2K} + \frac{\ln 2}{V_\infty^2 K}\|V\|_F^2\right) + \frac{1}{n}\ln\frac{4\sqrt{n}}{\delta} + \frac{2}{n}\ln\left(\frac{\ln(4V_\infty K/\gamma)}{\ln 2}\right).$$

Proof: This follows directly from the proof of Theorem 5 in Biggs & Guedj (2022), but applying kl^{-1} instead of Pinsker's inequality at the end of the proof in order to keep a tighter result. ■

In our experiments, the SHEL network is trained for 100 epochs on MNIST with Adam and with a cosine annealing learning rate schedule preceded by a linear warm-up phase. The training objective is the cross-entropy loss. We use the following temperature-scaled multi-class sigmoid loss for evaluating our bounds:

$$L_\tau(z, y) = 2\sum_{j\neq y} s(\tau(z_j - z_y)), \qquad s(t) := \frac{1}{1+e^{-t}},$$

where $\tau > 0$. Our reported bounds use the bounded derandomized PAC-Bayes bound in kl^{-1} form from Theorem 5.3. The temperature parameter is optimized over a dyadic grid of the form

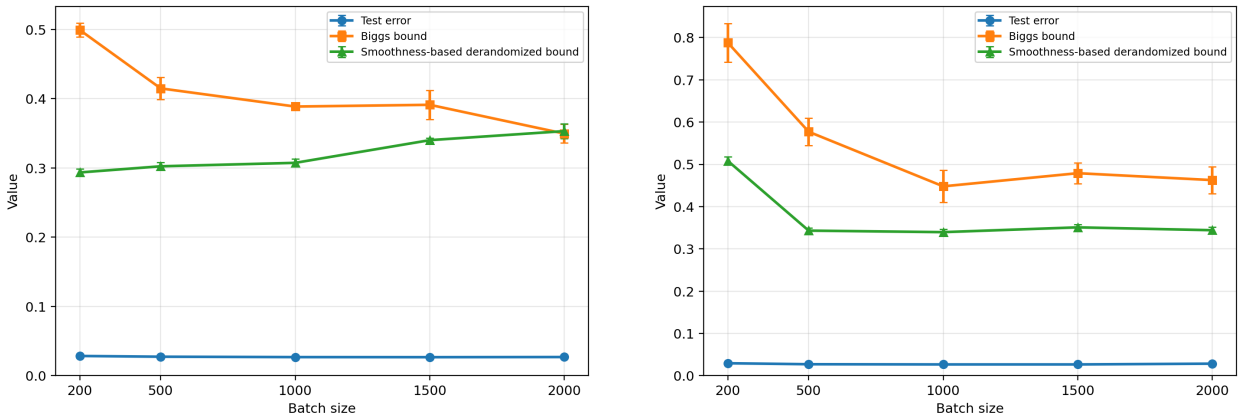
$$\tau_i = 2^{1-i}\tau_{\max}, \qquad \tau_{\max} = 2,$$

and a union bound is used over the τ-grid. The posterior scale parameter σ is similarly optimized over a dyadic grid with $\sigma_{\max} = 1$. Finally, in order to compare the relative behavior of the bounds outside the vacuous regime, we evaluate the bounds with an effective sample size n^* much larger than the true sample size n. This allows the optimization over the margin parameter γ in the Biggs bound and over τ in our bounds to avoid the trivial value 1. Indeed, at the true sample size n, the optimized bounds are equal to 1 and therefore do not allow a meaningful comparison between the different methods. The evaluations based on n^* should thus be interpreted only as comparisons of the relative behavior of the bounds, and not as PAC guarantees for the true sample size.

Figures 4 –5 report the resulting total bounds and test error as functions of either the hidden width K or the batch size, for learning rates 0.01 and 0.05. The reported results are averaged over three random seeds. Whenever the quantities J and H involve a supremum over the parameter class, we replace it by the largest value observed during training. The Gaussian averaged quantities are estimated by Monte Carlo sampling. These empirical approximations are not certified upper bounds on the corresponding theoretical quantities and therefore lead to an optimistic evaluation. Consequently, our main objective is not to compare the absolute tightness of the bounds, but rather to study their relative behavior as functions of the width K and the batch size.

As can be seen in Figure 4, the test error appears essentially constant as a function of the batch size in these experiments. To quantify how close each quantity is to being constant with respect to the batch size B, we fit

$$\log q = a + \alpha \log B,$$



(a) Learning rate 0.01.

(b) Learning rate 0.05.

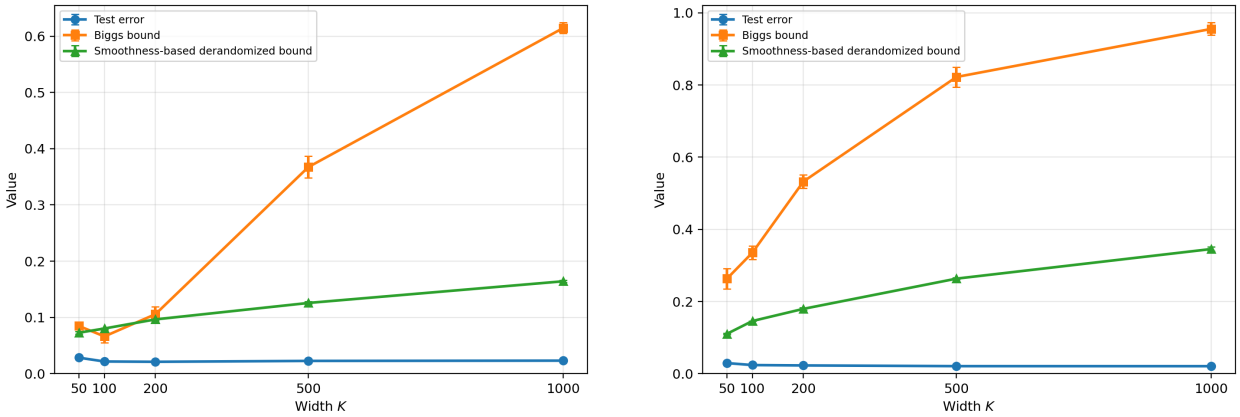
Figure 4: Generalization bounds after training on the cross-entropy loss as a function of the batch size (with $n^* = 6 \times 10^9$).

where q denotes the quantity under consideration. The fitted exponents for the test error are -0.026 and -0.024 for learning rates 0.01 and 0.05, respectively. The bounds exhibit a more noticeable dependence on the batch size. For learning rate 0.01, the Biggs bound decreases with exponent -0.136, whereas the smoothness based derandomized bound increases mildly with exponent 0.078. This opposite behavior may reflect the greater sensitivity of our bound to the empirical flatness quantities, which can increase with the batch size. For learning rate 0.05, both bounds decrease, with exponents -0.237 for the Biggs bound and -0.152 for our bound. The decrease of our bound is, however, mainly concentrated between batch sizes 200 and 500. Indeed, over the range from 500 to 2000, its fitted exponent is 0.009, indicating essentially constant behavior.

As shown in Figure 5, the test error also appears essentially constant as the width K increases. This indicates that the larger number of parameters does not lead to increased overfitting over the range considered. This observation is confirmed by the slightly negative fitted exponents reported in Table 1. By contrast, both bounds increase with K, although the smoothness based derandomized bound grows more slowly than the Biggs bound for both learning rates. Its estimated exponents are 0.274 and 0.378, compared with 0.759 and 0.457 for the Biggs bound.

It is difficult to decompose the optimized bound cleanly and isolate the contribution of each term, since the infimum over the bound parameters ties the different terms together. To obtain some insight into the source of the observed width dependence, we therefore fix σ and τ across the different values of K. We then report in Table 2 the global exponents of different terms in our bound. The Jacobian–Hessian term grows more slowly than the remainder of the bound. Moreover, removing the factor $\|w\|$ reduces its exponent from 0.232 to 0.110. These results suggest that the empirical Jacobian and Hessian quantities contribute to the weaker dependence of our bound on K.

The results of this section suggest that, in the derandomization process, we should seek to retain as many data dependent quantities as possible beyond the empirical surrogate used to upper-bound the 0-1 loss. This appears to be beneficial even for arguably small neural networks such as the SHEL networks. In the next section, we move from this correlational evidence toward a more causal investigation of the role played by these quantities.



(a) Learning rate 0.01.

(b) Learning rate 0.05.

Figure 5: Generalization bounds after training on the cross-entropy loss as a function of the width K (with $n^* = 10^{11}$).

Table 1: Estimated global exponents of the bounds and test error with respect to the width K.

Quantity	lr = 0.01	lr = 0.05
Biggs margin bound	0.759	0.457
Smoothness-based derandomized bound	0.274	0.378
Test error	−0.045	−0.107

Table 2: Estimated global exponents with respect to the width K for lr = 0.01 with $\sigma = 0.03125$ and $\tau = 0.5$ fixed across all widths. The exponent α is obtained by fitting $\log y = a + \alpha \log K$ by ordinary least squares, and R^2 denotes the coefficient of determination of this linear fit in log-log space.

Quantity	Exponent α	R^2
Full bound	0.293	0.973
$\|w\|\sqrt{\beta_\tau^2 J + \ell_\tau^2 H}$	0.232	0.972
$\sqrt{\beta_\tau^2 J + \ell_\tau^2 H}$	0.110	0.871
Remainder after subtracting the full Jacobian-Hessian term	0.378	0.868

9 A regularizer inspired by our theory for batch-normalized networks

The Rademacher complexity bound for the Jensen gap class obtained in our theory (see Corollary 6.4 and Remark 6.8) motivates data dependent quantities that can be used as regularizers for neural networks. Rather than attempting to compute a fully faithful version of the quantity appearing in the bound, we evaluate a local and computationally tractable proxy at the current network parameters during training. Let S_B denote the current training minibatch of size B, and let $S_{B'} \subseteq S_B$ denote a possibly smaller minibatch of size B' used to estimate the regularizer for computational efficiency. We define

$$\mathrm{JH}(w; S_{B'}) := \sqrt{\beta^2 J(w; S_{B'})^2 + \ell^2 H(w; S_{B'})},$$

where

$$J(w; S_{B'}) := \frac{1}{B'} \sum_{i=1}^{B'} \|J_i(w)\|_F^2$$

and

$$H(w; S_{B'}) := \frac{1}{B'} \sum_{i=1}^{B'} \sum_{k=1}^{C} \|H_{i,k}(w)\|_F^2.$$

The constants ℓ and β denote the Lipschitz and smoothness constants of the loss with respect to the network output.

For batch-normalized networks, we assume that each batch-normalization layer is applied before the nonlinearity. To remain close to the setting of the theoretical results, we work with an effective parametrization in which each batch-normalization layer is folded into the preceding affine layer. Indeed, for fixed batch-normalization statistics, batch normalization acts as an affine transformation of the pre-activation. Its composition with the preceding affine map can therefore be represented as a single effective affine map. More precisely, consider an affine layer followed by batch normalization,

$$z_t = W_t u_{t-1} + b_t, \qquad u_t = g_t(\mathrm{BN}_t(z_t)).$$

Equivalently, we absorb the bias into an augmented weight matrix by defining

$$\bar{u}_{t-1} := \begin{pmatrix} u_{t-1} \\ 1 \end{pmatrix}, \qquad \bar{W}_t := \begin{pmatrix} W_t & b_t \end{pmatrix},$$

so that

$$z_t = W_t u_{t-1} + b_t = \bar{W}_t \bar{u}_{t-1}.$$

Let $\mu_{t,B'}$ and $v_{t,B'}$ denote the empirical mean and variance of z_t. For convolutional layers, these statistics are computed channelwise over both the batch and spatial dimensions. If the batch-normalization layer has affine parameters γ_t and δ_t, then

$$\mathrm{BN}_t(z) = \gamma_t \odot \frac{z - \mu_{t,B'}}{\sqrt{v_{t,B'} + \varepsilon_t}} + \delta_t,$$

where $\odot$ denotes componentwise multiplication, and the division and square root are also taken componentwise. Define

$$s_{t,B'} := \frac{\gamma_t}{\sqrt{v_{t,B'} + \varepsilon_t}}.$$

Then, the affine map followed by batch normalization can be written as a single effective affine map

$$\mathrm{BN}_t(z_t) = \bar{W}_{t,\mathrm{eff},B'} \bar{u}_{t-1},$$

where

$$\bar{W}_{t,\mathrm{eff},B'} := \begin{pmatrix} \mathrm{Diag}(s_{t,B'}) W_t & s_{t,B'} \odot (b_t - \mu_{t,B'}) + \delta_t \end{pmatrix}.$$

For convolutional layers, $\mathrm{Diag}(s_{t,B'}) W_t$ means that each output channel of the convolutional kernel is multiplied by the corresponding batch-normalization scale. In the augmented notation, the final column of $\bar{W}_{t,\mathrm{eff},B'}$ represents the corresponding effective bias, which is shared across spatial locations. We denote by

$$w_{\mathrm{eff},B'} = \left\{ \bar{W}_{t,\mathrm{eff},B'} \right\}_t$$

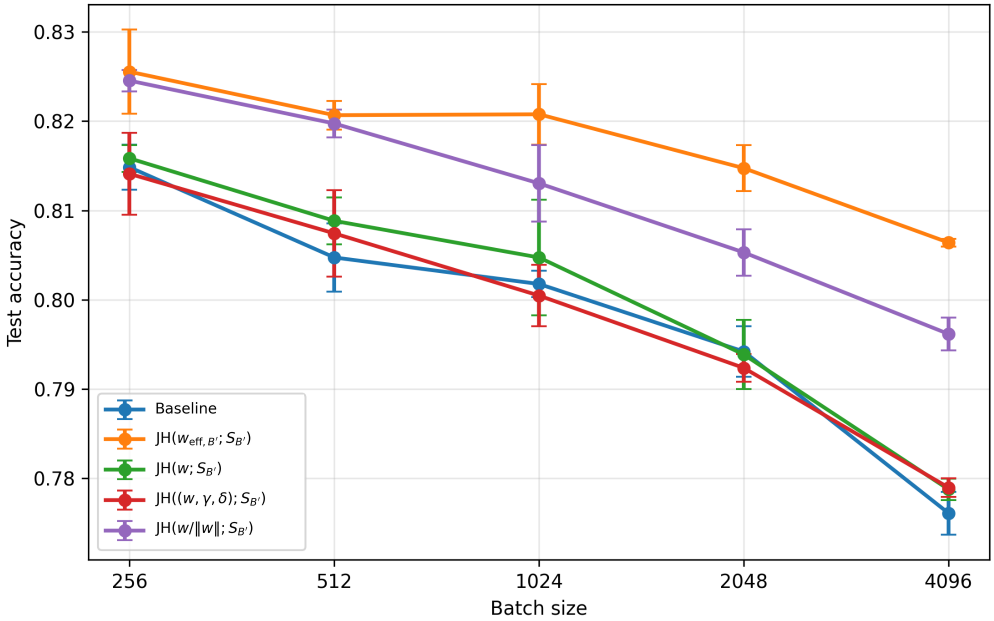


Figure 6: CIFAR-10 test accuracy as a function of the training batch size B for the baseline and four variants of the JH regularizer, estimated using a regularizer minibatch of size $B' = 50$. Among the tested variants, only when the regularizer is computed with respect to the effective weights or with respect to the normalized weights does the performance consistently and meaningfully improve over the baseline. Error bars indicate the standard deviation over three seeds.

the collection of all such effective affine parameters. Layers that are not followed by batch normalization are included unchanged in this collection. The regularizer is computed with respect to the effective parametrization $w_{\text{eff},B'}$. In other words, the Jacobian and Hessian terms are formed in the effective coordinates and evaluated at the effective parameters induced by the current network weights. The parameters being optimized remain the original weights w. The gradient of the regularized objective is backpropagated through the map $w \mapsto w_{\text{eff},B'}(w)$. We emphasize that the batch statistics $\mu_{t,B'}$ and $v_{t,B'}$ are not treated as constants: they depend also on w and gradients are propagated through this dependence as well.

The stochastic estimator used to compute the regularizer and the training details for the experiments below are provided in Appendix D.

In the first experiment, we compare four parametrizations of the JH regularizer with the unregularized cross-entropy baseline. The first version computes JH with respect to the effective weights $w_{\text{eff},B'}$, as described in this section. The second computes JH with respect to the original weights w, while the third also includes the affine parameters of the BatchNorm layers. The fourth computes JH with respect to layerwise normalized weights, where the weights of every layer of the network are replaced by $W_t/\|W_t\|_F$. We denote this method by $\text{JH}(w/\|w\|; S_{B'})$, where $w/\|w\|$ is used as shorthand for the collection $\{W_t/\|W_t\|_F\}_t$. Figure 6 shows the results across batch sizes. Only the effective weight version, $\text{JH}(w_{\text{eff},B'}; S_{B'})$, and the layerwise normalized-weight version, $\text{JH}(w/\|w\|; S_{B'})$, meaningfully improve over the baseline. This highlights the importance of choosing a parametrization that is compatible with the effective function implemented by networks with BatchNorm layers.

Remark 9.1. Jacobian and Hessian regularizers have been studied as training objectives, including regularization in input space (Sokolić et al., 2017; Varga et al., 2017; Cui et al., 2022) and Hessian trace regularization with respect to the model parameters (Liu et al., 2022). To our knowledge, previous approaches based on curvature or sharpness for BatchNorm networks define these quantities with respect to the original weights or to normalized or scale invariant parameterizations of these weights (Rangamani et al., 2021; Tsuzuku et al., 2020; Yi et al., 2019; Lyu et al., 2022). In contrast, we compute the regularizer with respect to the

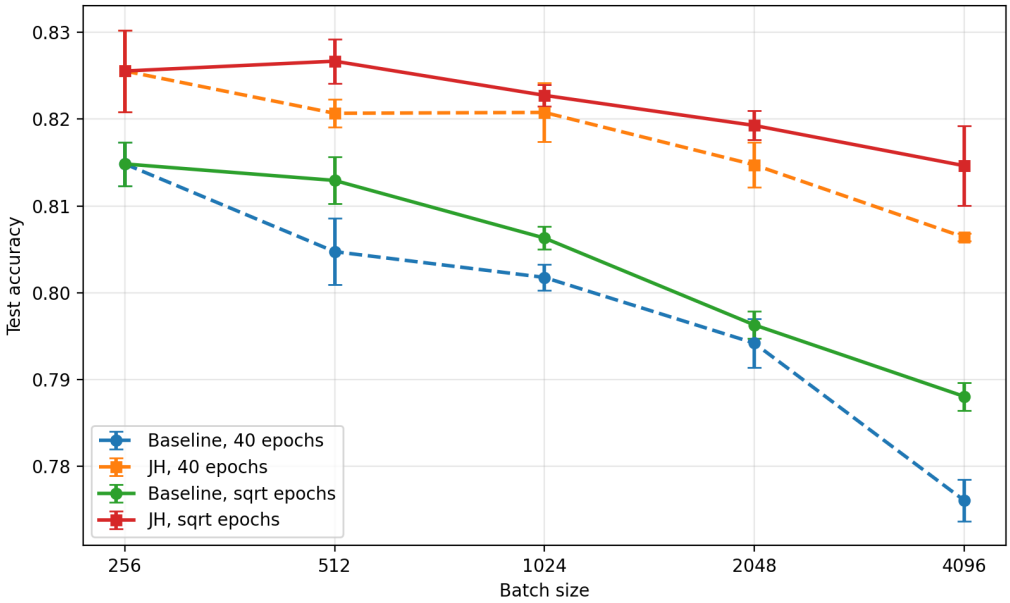


Figure 7: CIFAR-10 test accuracy as a function of the batch size for the baseline and the JH-regularized model. The fixed 40-epoch protocol is compared with the variable-epoch protocol where the number of epochs is scaled by $\sqrt{B/256}$. Error bars indicate the standard deviation over three seeds.

BatchNorm effective weights $w_{\mathrm{eff},B'}$. Such folding of the BatchNorm layer into the adjacent affine weights is used in the quantization and inference literature (Jacob et al., 2018; Nagel et al., 2021), where the goal is to deploy the same trained function more efficiently. In our case, the folding is used for a different purpose: it defines the coordinate system in which the regularizer is computed during training. Moreover, the effective weights are batch dependent because they are constructed using the current regularizer minibatch statistics rather than fixed test time BatchNorm statistics.

In the second experiment, we further investigate whether $\mathrm{JH}(w_{\mathrm{eff},B'}; S_{B'})$ can mitigate the degradation in performance sometimes observed when increasing the batch size. We compare the baseline and the JH-regularized model under two training-time protocols: the fixed 40-epoch protocol and the variable-epoch protocol, in which the number of epochs is scaled by $\sqrt{B/256}$. We perform this comparison both on clean CIFAR-10 and under 20% uniform label noise (with probability 0.2, each label is replaced by a uniformly sampled incorrect class). The results are shown in Figure 7 for the clean-label setting and in Figure 8 for the label-noise setting.

We first observe that the regularizer leads to larger absolute gains over the baseline in the label-noise setting than in the clean-label setting. This is expected since label noise exacerbates overfitting, thereby creating more room for the regularizer to improve generalization. In most cases, increasing the number of epochs benefits both the baseline and the JH-regularized model. The main exception occurs under label noise at batch size 4096, where the fixed-epoch JH run performs unusually well. We conjecture that this corresponds to an early-stopping-like regime, which may prevent excessive adaptation to the corrupted labels. Under label noise, when the number of epochs is increased according to the variable-epoch protocol, the larger batch sizes even outperform batch size 256. On clean data, the baseline performance deteriorates more rapidly as the batch size increases compared to the JH-regularized model. Overall, these results suggest that $\mathrm{JH}(w_{\mathrm{eff},B'}; S_{B'})$ can be beneficial in regimes where large batch training or noisy labels deteriorate performance.

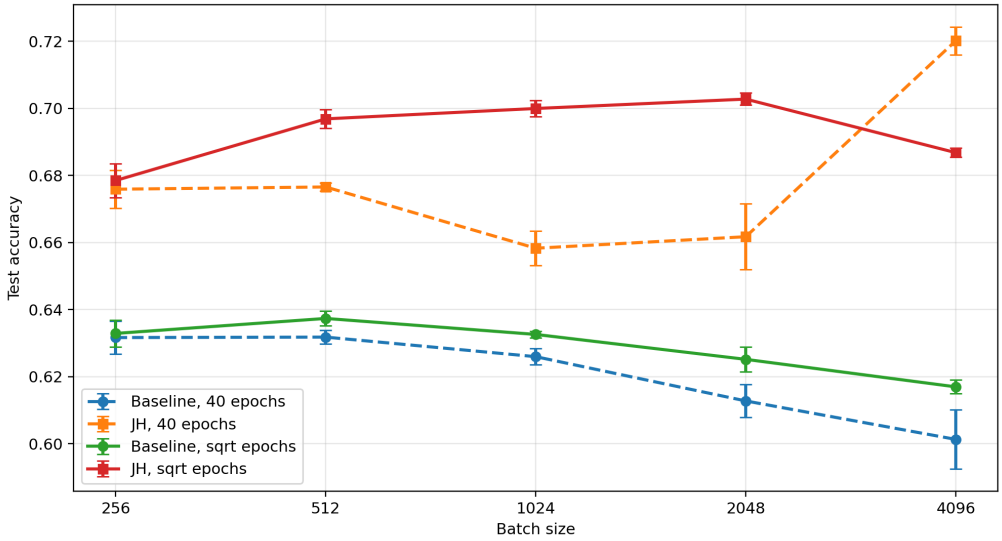


Figure 8: Clean CIFAR-10 test accuracy under 20% uniform label noise in the training and validation labels. The JH-regularized model is compared with the baseline under both the fixed 40-epoch protocol and the variable-epoch protocol where the number of epochs is scaled by $\sqrt{B/256}$. Error bars indicate the standard deviation over three seeds.

Remark 9.2. Unlike a regularizer built from the Hessian of the labeled loss, the JH regularizer is label-free: it is computed from the Jacobian and Hessian of the score map with respect to the (effective) parameters. Consequently, JH can also be evaluated on unlabeled inputs, making it potentially suitable as an auxiliary regularizer in settings where unlabeled data are available, such as semi-supervised or unsupervised learning. Future research could investigate this direction.

10 Conclusion

We introduced a smoothness based approach to PAC-Bayes derandomization. The main message is that the passage from a randomized Gibbs predictor to the deterministic predictor at the posterior mean can be isolated as a separate generalization problem, namely bounding the generalization gap of the Jensen gap class. This viewpoint makes it possible to combine PAC-Bayes control of the randomized predictor with data dependent control of the derandomization penalty. Controlling this derandomization term through Rademacher complexity naturally brings in Jacobian and Hessian quantities of the score map.

An important direction for future work is to obtain sharper high probability bounds for the Jensen gap class. In particular, local Rademacher complexity bounds could lead to tighter bounds than those obtained here. Another promising direction is an algorithmic stability analysis of the Jensen gap class, which could provide a different route to controlling the derandomization penalty and clarify its behavior under a given optimization algorithm.

References

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of gibbs posteriors. *Journal of Machine Learning Research*, 17(239):1–41, 2016.

Anonymous Authors. Anonymized prior work, 2026. Anonymized for double-blind review.

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`.

Fergus Immanuel Biggs and Benjamin Guedj. On margins and derandomization in pac-bayes. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Ioar Casado, Luis A. Ortega, Aritz Pérez, and Andrés R. Masegosa. PAC-Bayes-Chernoff bounds for unbounded losses. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://openreview.net/forum?id=CyzZeND3LB.

Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes–Monograph Series*. Institute of Mathematical Statistics, Beachwood, Ohio, 2007. doi: 10.1214/074921707000000391.

Eugenio Clerico, Tyler Farghly, George Deligiannidis, Benjamin Guedj, and Arnaud Doucet. Generalisation under gradient descent via deterministic PAC-Bayes. In Gautam Kamath and Po-Ling Loh (eds.), *Proceedings of The 36th International Conference on Algorithmic Learning Theory*, volume 272 of *Proceedings of Machine Learning Research*, pp. 349–389. PMLR, 24–27 Feb 2025. URL https://proceedings.mlr.press/v272/clerico25a.html.

Chenwei Cui, Zehao Yan, Guangshen Liu, and Liangfu Lu. Generalizing and improving Jacobian and Hessian regularization. *arXiv preprint arXiv:2212.00311*, 2022. URL https://arxiv.org/abs/2212.00311.

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 *Advances in Neural Information Processing Systems*, volume 33, pp. 11723–11733. Curran Associates, Inc., 2020.

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, pp. 353–360, Montreal, Quebec, Canada, 2009. ACM. doi: 10.1145/1553374.1553419.

Pascal Germain, Alexandre Lacasse, François Laviolette, Mario Marchand, and Jean-Francis Roy. Risk bounds for the majority vote: From a pac-bayesian analysis to a learning algorithm. *Journal of Machine Learning Research*, 16:787–860, 2015.

Pascal Germain, Francis Bach, Alexandre Lacoste, and Simon Lacoste-Julien. Pac-bayesian theory meets bayesian inference. In *Advances in Neural Information Processing Systems 29 (NeurIPS 2016)*, pp. 1–9, 2016. URL https://papers.nips.cc/paper/6569-pac-bayesian-theory-meets-bayesian-inference.

Aritra Ghosh, Naresh Manwani, and P.S. Sastry. Making risk minimization tolerant to label noise. *Neurocomput.*, 160(C):93–107, jul 2015. ISSN 0925-2312. doi: 10.1016/j.neucom.2014.09.081. URL https://doi.org/10.1016/j.neucom.2014.09.081.

Maxime Haddouche, Benjamin Guedj, Omar Rivasplata, and John Shawe-Taylor. Pac-bayes unleashed: Generalisation bounds with unbounded losses. *arXiv preprint arXiv:2006.07279*, 2020.

Maxime Haddouche, Paul Viallard, Umut Simsekli, and Benjamin Guedj. A PAC-Bayesian link between generalisation and flat minima. In Gautam Kamath and Po-Ling Loh (eds.), *Proceedings of The 36th International Conference on Algorithmic Learning Theory*, volume 272 of *Proceedings of Machine Learning Research*, pp. 481–511. PMLR, 24–27 Feb 2025. URL https://proceedings.mlr.press/v272/haddouche25a.html.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2704–2713, 2018.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJgIPJBFvH.

Alexandre Lacasse, François Laviolette, Mario Marchand, Pascal Germain, and Jean-Francis Roy. Pac-bayes bounds for the risk of the majority vote and the variance of the gibbs classifier. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 19, 2007.

John Langford and John Shawe-Taylor. Pac-bayes & margins. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 15, 2002.

Benjamin Leblanc and Pascal Germain. A framework for bounding deterministic risk with PAC-Bayes: Applications to majority votes. *arXiv preprint arXiv:2510.25569*, 2025. URL `https://arxiv.org/abs/2510.25569`.

Gaël Letarte, Pascal Germain, Benjamin Guedj, and François Laviolette. Dichotomize and generalize: PAC-Bayesian binary activated deep neural networks. In *Advances in Neural Information Processing Systems 32*, pp. 6869–6879. Curran Associates, Inc., 2019.

Yucong Liu, Shixing Yu, and Tong Lin. Regularizing deep neural networks with stochastic estimators of Hessian trace. *arXiv preprint arXiv:2208.05924*, 2022. URL `https://arxiv.org/abs/2208.05924`.

Kaifeng Lyu, Zhiyuan Li, and Sanjeev Arora. Understanding the generalization benefit of normalization layers: Sharpness reduction. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Andrés R. Masegosa, Serafín Moral, and Manuel Cebrián. Second-order pac-bayesian bounds for weighted majority votes. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Andreas Maurer. A note on the PAC-Bayesian theorem, 2004. URL `https://arxiv.org/abs/cs/0411099`.

Andreas Maurer. A vector-contraction inequality for rademacher complexities. *CoRR*, abs/1605.00251, 2016. URL `http://arxiv.org/abs/1605.00251`.

David A. McAllester. Some PAC-bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pp. 230–234. Association for Computing Machinery, 1998. doi: 10.1145/279943.279989.

David A. McAllester. Pac-bayesian model averaging. In *Proceedings of the 12th Annual Conference on Computational Learning Theory (COLT)*, pp. 164–170. ACM, 1999.

Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*, 2021.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Akshay Rangamani, Nam H. Nguyen, Abhishek Kumar, Dzung Phan, Sang H. Chin, and Trac D. Tran. A scale invariant measure of flatness for deep network minima. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1680–1684. IEEE, 2021. doi: 10.1109/ICASSP39728.2021.9413771.

Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvári, and John Shawe-Taylor. PAC-Bayes analysis beyond the usual bounds. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL `https://arxiv.org/abs/2006.13057`.

Matthias Seeger. PAC-bayesian generalisation error bounds for gaussian process classification. *Journal of Machine Learning Research*, 3:233–269, October 2002. URL `https://www.jmlr.org/papers/v3/seeger02a.html`.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, Cambridge, UK, 2014. ISBN 9781107057135.

Neta Shoham, Tomer Avidor, and Nadav Israel. An exploration into why output regularization mitigates label noise. *arXiv preprint arXiv:2104.12477*, 2021.

Jure Sokolić, Raja Giryes, Guillermo Sapiro, and Miguel R. D. Rodrigues. Robust large margin deep neural networks. *IEEE Transactions on Signal Processing*, 65(16):4265–4280, 2017. doi: 10.1109/TSP.2017.2708039.

Michel Talagrand. *The Generic Chaining: Upper and Lower Bounds of Stochastic Processes*. Springer Monographs in Mathematics. Springer, Berlin, Heidelberg, 2005.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using PAC-bayesian analysis. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9636–9647. PMLR, 2020. URL https://proceedings.mlr.press/v119/tsuzuku20a.html.

Dániel Varga, Adrián Csiszárik, and Zsolt Zombori. Gradient regularization improves accuracy of discriminative models. *arXiv preprint arXiv:1712.09936*, 2017.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, 2018. ISBN 9781108415194. URL https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-2.pdf. See Corollary 8.5.6 (Talagrand comparison inequality).

Paul Viallard, Pascal Germain, Amaury Habrard, and Emilie Morvant. A general framework for the practical disintegration of pac-bayesian bounds. *CoRR*, abs/2102.08649, 2021. URL https://arxiv.org/abs/2102.08649.

Mingyang Yi, Huishuai Zhang, Wei Chen, Zhi-Ming Ma, and Tie-Yan Liu. BN-invariant sharpness regularizes the training model to better generalization. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 4164–4170. International Joint Conferences on Artificial Intelligence Organization, 2019.

Xiong Zhou, Xianming Liu, Hanzhang Wang, Deming Zhai, Jiangjunjun, and Xiangyang Ji. On the dynamics under the unhinged loss and beyond. *Journal of Machine Learning Research*, 24:1–62, 2023. URL https://jmlr.org/papers/v24/23-0771.html.

A Uniform upper bounds for the Jensen gap for smooth neural networks

We start with a general result for bounding the Jensen gap of the composition of a function f with a neural network $h_w(x)$ exploiting the Lipschitz constant and the β-smoothness of the function f.

Lemma A.1. *Assume that the hypothesis class is parameterized by $w \in \mathbb{R}^m$ and Q is a probability distribution on $\mathbb{R}^m$. Denote by h_w a neural network mapping inputs $x \in \mathbb{R}^d$ to vectors in $\mathbb{R}^H$. Let $f : \mathbb{R}^H \to \mathbb{R}^K$ (or $f : \mathbb{R}^H \times Y \to \mathbb{R}^K$ if f is a loss function) be differentiable. Furthermore, assume that f is ℓ-Lipschitz and β-smooth (for all $y \in Y$ if there is a dependency on y). We will omit the variable y in order to simplify notations in what follows. The property of β-smoothness means that the Jacobian $\nabla f \in \mathbb{R}^{K\times H}$ is β-Lipschitz with respect to the spectral norm $\|\cdot\|_2$, that is,*

$$\|\nabla f(z) - \nabla f(z')\|_2 \leq \beta\|z - z'\|, \quad \text{for all } z, z' \in \mathbb{R}^H.$$

For any fixed input $x \in \mathbb{R}^d$, $f(h_w(x))$ is a function of w. Then, we have that the Jensen gap of $f(h_w(x))$ with respect to Q satisfies

$$\|\mathrm{Jen}_Q[f(h_w(x))]\| \leq \ell\|\mathrm{Jen}_Q[h_w(x)]\| + \frac{\beta}{2}\mathrm{Var}_Q[h_w(x)],$$

where $\mathrm{Var}_Q[h_w(x)]$ denotes the total variance of the vector $h_w(x)$ under distribution Q (trace of the covariance matrix).

Proof: Let $\overline{w} = \mathbb{E}_{w\sim Q}[w]$ and $\overline{z} = \mathbb{E}_{w\sim Q}[h_w(x)]$. Then,

$$\begin{aligned}\|\mathrm{Jen}_Q[f(h_w(x))]\| &= \|f(h_{\overline{w}}(x)) - \mathbb{E}_{w\sim Q}[f(h_w(x))]\| \\ &\leq \|f(h_{\overline{w}}(x)) - f(\overline{z})\| + \|f(\overline{z}) - \mathbb{E}_{w\sim Q}[f(h_w(x))]\|.\end{aligned}$$

For the first term, the ℓ-Lipschitz property of f gives

$$\|f(h_{\overline{w}}(x)) - f(\overline{z})\| \leq \ell\|h_{\overline{w}}(x) - \overline{z}\| = \ell\|\mathrm{Jen}_Q[h_w(x)]\|.$$

For the second term, we apply the integral form of first-order Taylor's theorem to $f(z)$ around $\overline{z}$:

$$f(z) = f(\overline{z}) + \nabla f(\overline{z})(z - \overline{z}) + \int_0^1 \left(\nabla f(\overline{z} + t(z - \overline{z})) - \nabla f(\overline{z})\right)(z - \overline{z})\, dt.$$

Taking expectation over w on both sides with $z = h_w(x)$, we obtain

$$\mathbb{E}_Q[f(h_w(x))] = f(\overline{z}) + \mathbb{E}_{w\sim Q}\left[\int_0^1 \left(\nabla f(\overline{z} + t(h_w(x) - \overline{z})) - \nabla f(\overline{z})\right)(h_w(x) - \overline{z})\, dt\right],$$

Using the β-smoothness of f, we get

$$\|\nabla f(\overline{z} + t(h_w(x) - \overline{z})) - \nabla f(\overline{z})\|_2 \leq \beta t\|h_w(x) - \overline{z}\|.$$

Thus,

$$\|f(\overline{z}) - \mathbb{E}_Q[f(h_w(x))]\| \leq \mathbb{E}_{w\sim Q}\int_0^1 \beta t\|h_w(x) - \overline{z}\|^2\, dt = \frac{\beta}{2}\mathrm{Var}_Q[h_w(x)].$$

■

Our proof for bounding the variance term for neural networks will be in the form of an inductive argument on the number of layers and will exploit the following lemma.

Lemma A.2. *Assume that $f : \mathbb{R}^m \to \mathbb{R}^k$ is ℓ-Lipschitz. Then,*

$$\mathrm{Var}[f(x)] \leq \ell^2\mathrm{Var}[x].$$

Proof: On one hand,

$$\begin{aligned}\mathrm{Var}[f(x)] = \mathbb{E}_x[\|f(x) - \mathbb{E}_x f(x)\|^2] &= \mathbb{E}_x[\|f(x)\|^2 - 2(f(x)\cdot\mathbb{E}_x f(x)) + \|\mathbb{E}_x f(x)\|^2] \\ &= \mathbb{E}_x[\|f(x)\|^2] - \|\mathbb{E}_x f(x)\|^2.\end{aligned}$$

On the other hand, if x' follows the same distribution as x but is independent from x,

$$\begin{aligned}\mathbb{E}_x\mathbb{E}_{x'}[\|f(x) - f(x')\|^2] &= \mathbb{E}_{x'}\mathbb{E}_x[\|f(x)\|^2 - 2\mathbb{E}_{x'}\mathbb{E}_x(f(x)\cdot f(x')) + \mathbb{E}_{x'}\mathbb{E}_x\|f(x')\|^2] \\ &= 2\mathbb{E}_x[\|f(x)\|^2] - 2\|\mathbb{E}_x f(x)\|^2.\end{aligned}$$

Therefore,

$$\mathrm{Var}[f(x)] = \frac{1}{2}\mathbb{E}_x\mathbb{E}_{x'}[\|f(x) - f(x')\|^2].$$

We can now exploit the Lipschitz property to conclude the proof:

$$\mathrm{Var}[f(x)] = \frac{1}{2}\mathbb{E}_x\mathbb{E}_{x'}[\|f(x) - f(x')\|^2] \leq \frac{\ell^2}{2}\mathbb{E}_x\mathbb{E}_{x'}[\|x - x'\|^2] = \ell^2\mathrm{Var}[x].$$

■

We are now ready to upper bound the variance of the outputs of a neural network.

Proposition A.3 (Bounded activation functions)**.** *Consider a feedforward fully connected neural network with T layers $h_{W_{1:T}}(x) = W_T g_{T-1}(W_{T-1} g_{T-2}(\cdots g_1(W_1 x)))$ where each non-linear function $g_t : \mathbb{R}^{d_t} \to \mathbb{R}^{d_t}$ is ℓ_t-Lipschitz. We further assume that these functions are bounded: $\|g_t(z)\| \leq M_t$ for all $z \in \mathbb{R}^{d_t}$. We let $d_T := C$ be the number of classes. Assume that $\|x\| \leq M_0$ and let Q be an isotropic Gaussian distribution centered at $\overline{W}_1, \cdots, \overline{W}_T$ with covariance matrix given by σ^2 times the identity matrix. Then,*

$$\mathrm{Var}_Q[h_{W_{1:T}}(x)] \leq \sigma^2 \sum_{t=1}^{T} d_t M_{t-1}^2 \prod_{t'=t}^{T-1} \ell_{t'}^2 \|\overline{W}_{t'+1}\|_2^2,$$

where the product is equal to 1 if $t = T$.

Proof: The proof is done by induction on the number of layers T. For $T = 1$, we have $h_{W_1}(x) = W_1 x$. If w_i denotes the ith row of W_1, then $w_i \sim \mathcal{N}(\overline{w}_i, \sigma^2 I)$ and therefore $\mathrm{Var}(w_i^T x) = \sigma^2 \|x\|^2$. Summing over the d_1 output coordinates gives

$$\mathrm{Var}_Q[h_{W_1}(x)] = \sum_{i=1}^{d_1} \mathrm{Var}(w_i^T x) = d_1 \sigma^2 \|x\|^2 \leq d_1 \sigma^2 M_0^2,$$

which matches the claimed bound since the product in the statement equals 1 when $T = 1$. Assume that the result is true for $T - 1$ layers. We have

$$\mathrm{Var}_Q[h_{W_{1:T}}(x)] = \mathbb{E}_{W_{1:T-1}} \mathbb{E}_{W_T}[\|W_T g_{T-1}(h_{W_{1:T-1}}(x)) - \overline{W}_T \mathbb{E}_{W_{1:T-1}}[g_{T-1}(h_{W_{1:T-1}}(x))]\|^2].$$

To simplify notations, write $W := W_T$, $u := g_{T-1}(h_{W_{1:T-1}}(x))$, $\overline{W} := \overline{W}_T$ and $\overline{u} := \mathbb{E}_{W_{1:T-1}}[g_{T-1}(h_{W_{1:T-1}}(x))]$. Then,

$$\mathbb{E}_W[\|Wu - \overline{W}\overline{u}\|^2] = \sum_i \mathbb{E}_{w_i}[(w_i^T u - \overline{w}_i^T \overline{u})^2] = \sum_i \mathbb{E}_{w_i}[(w_i^T u)^2 - 2(w_i^T u)(\overline{w}_i^T \overline{u}) + (\overline{w}_i^T \overline{u})^2],$$

where w_i is the vector corresponding to the ith row of W and $\overline{w}_i$ is the vector corresponding to the ith row of $\overline{W}$. Since $\mathbb{E}_{w_i}[(w_i^T u)^2] = \sigma^2 \|u\|^2 + (\overline{w}_i^T u)^2$, we get

$$\begin{aligned}
\mathbb{E}_W[\|Wu - \overline{W}\overline{u}\|^2] &= \sum_i \left[\sigma^2 \|u\|^2 + (\overline{w}_i^T u)^2 - 2(\overline{w}_i^T u)(\overline{w}_i^T \overline{u}) + (\overline{w}_i^T \overline{u})^2\right] \\
&= d_T \sigma^2 \|u\|^2 + \sum_i (\overline{w}_i^T u - \overline{w}_i^T \overline{u})^2 \\
&= d_T \sigma^2 \|u\|^2 + \|\overline{W} u - \overline{W}\overline{u}\|^2 \\
&\leq d_T \sigma^2 M_{T-1}^2 + \|\overline{W}\|_2^2 \|u - \overline{u}\|^2.
\end{aligned}$$

Therefore,

$$\mathrm{Var}_Q[h_{W_{1:T}}(x)] \leq d_T \sigma^2 M_{T-1}^2 + \|\overline{W}_T\|_2^2 \mathrm{Var}_Q[u] \leq d_T \sigma^2 M_{T-1}^2 + \|\overline{W}_T\|_2^2 \ell_{T-1}^2 \mathrm{Var}_Q[h_{W_{1:T-1}}(x)],$$

where we used Lemma A.2 for the last inequality. Using the induction hypothesis, we get

$$\begin{aligned}
\mathrm{Var}_Q[h_{W_{1:T}}(x)] &\leq d_T \sigma^2 M_{T-1}^2 + \sigma^2 \|\overline{W}_T\|_2^2 \ell_{T-1}^2 \sum_{t=1}^{T-1} d_t M_{t-1}^2 \prod_{t'=t}^{T-2} \ell_{t'}^2 \|\overline{W}_{t'+1}\|_2^2 \\
&= d_T \sigma^2 M_{T-1}^2 + \sigma^2 \sum_{t=1}^{T-1} d_t M_{t-1}^2 \prod_{t'=t}^{T-1} \ell_{t'}^2 \|\overline{W}_{t'+1}\|_2^2 \\
&= \sigma^2 \sum_{t=1}^{T} d_t M_{t-1}^2 \prod_{t'=t}^{T-1} \ell_{t'}^2 \|\overline{W}_{t'+1}\|_2^2.
\end{aligned}$$

■

Proposition A.4 (Unbounded activation functions)**.** *Consider a feedforward fully connected neural network with T layers*

$$h_{W_{1:T}}(x) = W_T g_{T-1}(W_{T-1} g_{T-2}(\cdots g_1(W_1 x))),$$

where each non-linear function $g_t : \mathbb{R}^{d_t} \to \mathbb{R}^{d_t}$ is ℓ_t-Lipschitz and satisfies $g_t(0) = 0$. We let $d_T := C$ be the number of classes. Assume that $\|x\| \leq M_0$, and let Q be an isotropic Gaussian distribution centered at $\overline{W}_1, \cdots, \overline{W}_T$ with covariance matrix given by σ^2 times the identity matrix. Then,

$$\mathrm{Var}_Q[h_{W_{1:T}}(x)] \leq \sigma^2 M_0^2 \sum_{t=1}^{T} d_t \left(\prod_{t'=1}^{t-1} \ell_{t'}^2 \big(\sigma^2 d_{t'} + \|\overline{W}_{t'}\|_2^2\big) \right) \left(\prod_{t'=t}^{T-1} \ell_{t'}^2 \|\overline{W}_{t'+1}\|_2^2 \right),$$

where the first product is equal to 1 if $t = 1$ and the second product is equal to 1 if $t = T$.

Proof: Define recursively

$$u_0 := x, \qquad u_t := g_t(W_t u_{t-1}).$$

The proof is almost identical to that of Proposition A.3, but M_{t-1}^2 is replaced by

$$\mathbb{E}\|u_{t-1}\|^2 \leq M_0^2 \prod_{t'=1}^{t-1} \ell_{t'}^2 \big(\sigma^2 d_{t'} + \|\overline{W}_{t'}\|_2^2\big).$$

This inequality is true since

$$\begin{aligned}
\mathbb{E}\|u_{t-1}\|^2 &\leq \ell_{t-1}^2 \, \mathbb{E}\Big[\mathbb{E}_{W_{t-1}}\big[\|W_{t-1} u_{t-2}\|^2\big]\Big] \\
&= \ell_{t-1}^2 \, \mathbb{E}\Big[d_{t-1}\sigma^2 \|u_{t-2}\|^2 + \|\overline{W}_{t-1} u_{t-2}\|^2\Big] \\
&\leq \ell_{t-1}^2 \, \mathbb{E}\Big[\big(\sigma^2 d_{t-1} + \|\overline{W}_{t-1}\|_2^2\big)\|u_{t-2}\|^2\Big] \\
&= \ell_{t-1}^2 \big(\sigma^2 d_{t-1} + \|\overline{W}_{t-1}\|_2^2\big) \, \mathbb{E}\|u_{t-2}\|^2 \\
&\leq \cdots \\
&\leq M_0^2 \prod_{t'=1}^{t-1} \ell_{t'}^2 \big(\sigma^2 d_{t'} + \|\overline{W}_{t'}\|_2^2\big).
\end{aligned}$$

■

Propositions A.3 and A.4 control the variance term appearing in Lemma A.1 for deep networks. The next two theorems exploit these results to bound the Jensen gap of the loss under smoothness assumptions on the activation functions.

Theorem A.5 (Bounded activation functions)**.** *Consider the same assumptions as in Proposition A.3 and also assume that each non-linear function $g_t : \mathbb{R}^{d_t} \to \mathbb{R}^{d_t}$ is β_t-smooth. Then, we have*

$$|\mathrm{Jen}_Q[L(h_{W_{1:T}}(x), y)]| \leq \frac{\sigma^2}{2} \sum_{t=1}^{T} \sum_{u=1}^{t} \beta_t d_u M_{u-1}^2 \Big(\prod_{t'=u}^{t-1} \ell_{t'}^2 \|\overline{W}_{t'+1}\|_2^2 \Big) \Big(\prod_{t'=t}^{T-1} \ell_{t'+1} \|\overline{W}_{t'+1}\|_2 \Big),$$

where the first product is equal to 1 if $u = t$ and the second product is equal to 1 if $t = T$.

Proof: Let $g_T(z) := L(z, y)$. The dependency on y does not matter for the proof. The β-smoothness of the loss function $L(z, y)$ will be denoted by β_T and the Lipschitz constant by ℓ_T. We will show by induction on the number of layers that

$$|\mathrm{Jen}_Q[g_T(h_{W_T}(x))]| \leq \sum_{t=1}^{T} \frac{\beta_t}{2} \mathrm{Var}_Q[h_{W_{1:t}}(x)] \prod_{t'=t}^{T-1} \ell_{t'+1} \|\overline{W}_{t'+1}\|_2,$$

where the product is equal to 1 if $t = T$. For $T = 1$, the result follows immediately from Lemma A.1. Assume that the result is true for $T-1$ layers. From Lemma A.1 and the induction hypothesis, we have

$$\begin{aligned}
|\mathrm{Jen}_Q[g_T(h_{W_{1:T}}(x))]| &\leq \ell_T \|\mathrm{Jen}_Q[h_{W_{1:T}}(x)]\| + \frac{\beta_T}{2}\mathrm{Var}_Q[h_{W_{1:T}}(x)] \\
&\leq \ell_T \|\overline{W}_T\|_2 \|\mathrm{Jen}[g_{T-1}(h_{W_{1:T-1}}(x))]\| + \frac{\beta_T}{2}\mathrm{Var}_Q[h_{W_{1:T}}(x)] \\
&\leq \ell_T \|\overline{W}_T\|_2 \sum_{t=1}^{T-1} \frac{\beta_t}{2}\mathrm{Var}_Q[h_{W_{1:t}}(x)] \prod_{t'=t}^{T-2} \ell_{t'+1}\|\overline{W}_{t'+1}\|_2 + \frac{\beta_T}{2}\mathrm{Var}_Q[h_{W_{1:T}}(x)] \\
&= \sum_{t=1}^{T-1} \frac{\beta_t}{2}\mathrm{Var}_Q[h_{W_{1:t}}(x)] \prod_{t'=t}^{T-1} \ell_{t'+1}\|\overline{W}_{t'+1}\|_2 + \frac{\beta_T}{2}\mathrm{Var}_Q[h_{W_{1:T}}(x)] \\
&= \sum_{t=1}^{T} \frac{\beta_t}{2}\mathrm{Var}_Q[h_{W_{1:t}}(x)] \prod_{t'=t}^{T-1} \ell_{t'+1}\|\overline{W}_{t'+1}\|_2.
\end{aligned}$$

The proof of the Theorem is now completed by using the upper bound on $\mathrm{Var}_Q[h_{W_{1:t}}(x)]$ obtained in Proposition A.3. ■

Theorem A.6 (Unbounded activation functions)**.** *Consider the same assumptions as in Proposition A.4 and also assume that each non-linear function $g_t : \mathbb{R}^{d_t} \to \mathbb{R}^{d_t}$ is β_t-smooth. Then, we have*

$$\begin{aligned}
|\mathrm{Jen}_Q[L(h_{W_{1:T}}(x), y)]| \leq \frac{\sigma^2 M_0^2}{2} \sum_{t=1}^{T}\sum_{u=1}^{t} \beta_t d_u \\
\times \left(\prod_{t'=1}^{u-1} \ell_{t'}^2 \left(\sigma^2 d_{t'} + \|\overline{W}_{t'}\|_2^2\right)\right)\left(\prod_{t'=u}^{t-1} \ell_{t'}^2 \|\overline{W}_{t'+1}\|_2^2\right)\left(\prod_{t'=t}^{T-1} \ell_{t'+1}\|\overline{W}_{t'+1}\|_2\right),
\end{aligned}$$

where the first product is equal to 1 if $u = 1$, the second product is equal to 1 if $u = t$, and the third product is equal to 1 if $t = T$.

Proof: The proof is the same as that of Theorem A.5, but using Proposition A.4 instead of Proposition A.3. ■

B Deferred proofs for Sections 3-4 and 7

Proof of Lemma 3.8. The quantity inside the expectation depends on $W_{1:T}$ only through W_T. Writing P_T for the marginal distribution of W_T under P, we have

$$\mathbb{E}_{W_{1:T}\sim P} \exp\left(\frac{\lambda^2 \ell^2 a_{W_{1:T}}^2}{2n}\right) = \mathbb{E}_{W_T \sim P_T} \exp\left(\frac{\lambda^2 \ell^2 M^2 \|W_T\|^2}{2n}\right).$$

Let

$$t = \frac{M^2 \lambda^2 \ell^2}{2n}.$$

Since $W_T \sim \mathcal{N}(0, \sigma^2 I_{m_T})$, we have

$$\mathbb{E}_{W_T \sim P_T} \exp\left(t\|W_T\|^2\right) = (1 - 2t\sigma^2)^{-m_T/2},$$

when $2t\sigma^2 < 1$. For the chosen value

$$\lambda = \frac{\sqrt{n}}{M\ell\sigma\sqrt{m_T + 1}},$$

we have

$$2t\sigma^2 = \frac{1}{m_T + 1}.$$

Therefore,

$$\mathbb{E}_{W_{1:T}\sim P}\exp\left(\frac{\lambda^2\ell^2 a^2_{W_{1:T}}}{2n}\right) = \left(\frac{m_T+1}{m_T}\right)^{m_T/2}.$$

Hence,

$$\frac{1}{\lambda}\ln\mathbb{E}_{W_{1:T}\sim P}\exp\left(\frac{\lambda^2\ell^2 a^2_{W_{1:T}}}{2n}\right) = \frac{M\ell\sigma\sqrt{m_T+1}}{\sqrt{n}}\frac{m_T}{2}\ln\left(1+\frac{1}{m_T}\right).$$

Using

$$m_T\ln\left(1+\frac{1}{m_T}\right)\le 1,$$

we obtain

$$\frac{1}{\lambda}\ln\mathbb{E}_{W_{1:T}\sim P}\exp\left(\frac{\lambda^2\ell^2 a^2_{W_{1:T}}}{2n}\right) \le \frac{M\ell\sigma\sqrt{m_T+1}}{2\sqrt{n}}.$$

■

Proof of Theorem 4.1. Let the prior P and posterior Q be isotropic Gaussian distributions with the same covariance matrix $\sigma^2 I_{m\times m}$, with prior mean 0 and posterior mean W. Then,

$$D_{\mathrm{KL}}(Q\|P) = \frac{\|W\|_F^2}{2\sigma^2}.$$

We apply Theorem 3.7 to the linear hypothesis class $h_W(x) = Wx$ with parameter set $\Theta = \{W \in \mathbb{R}^{C\times d} : \|W\|_F \le R\}$ and with $\|x\| \le M$. The data dependent complexity term appearing in Theorem 3.7 satisfies

$$2c\,\frac{R\sigma}{\sqrt{n}}\sqrt{\beta^2 J_{S,\sigma} + \ell^2 H_{S,\sigma}} \;\le\; 2\sqrt{2}\,\frac{\beta R\sigma C M^2}{\sqrt{n}}$$

(see Remark 6.5 and Theorem 6.1 in Section 6).

Moreover, specializing the Jensen gap upper bound in Theorem A.5 from Section A to the linear case ($T = 1$) yields

$$|\mathrm{Jen}_Q[L(h_W(x), y)]| \le \frac{\sigma^2}{2}\,\beta\, C\, M^2.$$

Thus, the uniform assumption $|\mathrm{Jen}_Q[L(h_W(x), y)]| \le \sigma^2 B$ in Theorem 3.7 holds with

$$B = \frac{\beta C M^2}{2}.$$

The corresponding concentration term therefore becomes

$$8\sigma^2 B\sqrt{\frac{2\ln(8/\delta)}{n}} = 4\beta C M^2\sigma^2\sqrt{\frac{2\ln(8/\delta)}{n}}.$$

Finally, using Lemma 3.8 and the above value of the KL divergence, the PAC-Bayes complexity term is bounded by

$$\frac{M\ell\sqrt{m+1}}{2\sigma\sqrt{n}}\|W\|_F^2 + \frac{M\ell\sigma\sqrt{m+1}}{\sqrt{n}}\ln\frac{2}{\delta} + \frac{M\ell\sigma\sqrt{m+1}}{2\sqrt{n}}.$$

■

Proof of Lemma 4.3. Let us write

$$L^{\text{unh}}(z, y) = a\left(-z_y + \frac{1}{C}\sum_{k=1}^{C} z_k\right) + b$$

for some constants $a, b > 0$.

We want $\inf_{\|z\|\leq r} L^{\text{unh}}(z, y) = 0$. This infimum can be computed by standard Lagrangian optimization. The solution is

$$\inf_{\|z\|\leq r} L^{\text{unh}}(z, y) = -ar\sqrt{\frac{C-1}{C}} + b.$$

To make the infimum equal to 0, we set:

$$b = ar\sqrt{\frac{C-1}{C}}.$$

We now turn our attention towards upper bounding the 0-1 loss. Suppose there exists $k \neq y$ such that $z_k \geq z_y$, so the 0-1 loss is 1 at this value of (z, y). We want to compute the infimum of $-z_y + \frac{1}{C}\sum_{k=1}^{C} z_k$ under the constraints $\|z\| \leq r$ and $z_k \geq z_y$. At a global minimizer, we must actually have $z_k = z_y$ (this can be seen by a simple perturbation argument). We may therefore restrict the optimization to the set

$$B(0, r) \cap \{z : z_k = z_y\}.$$

Both the objective and the constraint are invariant under permutations of the coordinates $\{z_j\}_{j\neq y,k}$. Therefore, if z is a minimizer, any permutation of these coordinates is also a minimizer. Averaging over all such permutations gives another feasible point satisfying

$$z_j = t \qquad (\forall j \neq y, k)$$

for some $t \in \mathbb{R}$, and

$$z_y = z_k = s$$

for some $s \in \mathbb{R}$. The optimization problem thus reduces to

$$\min_{s,t\in\mathbb{R}} \frac{C-2}{C}(t - s)$$

subject to

$$2s^2 + (C-2)t^2 \leq r^2.$$

The solution of this problem can be obtained using Lagrange multipliers. One obtains

$$t - s = -r\sqrt{\frac{C}{2(C-2)}}.$$

Therefore,

$$\inf_{\substack{\|z\|\leq r \\ z_k \geq z_y}} \left(-z_y + \frac{1}{C}\sum_{i=1}^{C} z_i\right) = -r\sqrt{\frac{C-2}{2C}}$$

and

$$L^{\text{unh}}(z, y) \geq -ar\sqrt{\frac{C-2}{2C}} + b.$$

Substituting $b = ar\sqrt{\frac{C-1}{C}}$, we obtain

$$L^{\text{unh}}(z, y) \geq ar\left(\sqrt{\frac{C-1}{C}} - \sqrt{\frac{C-2}{2C}}\right).$$

Therefore, $L^{0\text{-}1}(z,y) \le L^{\text{unh}}(z,y)$ provided

$$a \ge \frac{1}{r\left(\sqrt{\frac{C-1}{C}} - \sqrt{\frac{C-2}{2C}}\right)} = \frac{\sqrt{C}}{r\big(\sqrt{C-1} - \sqrt{\frac{C-2}{2}}\big)}.$$

With the choice of a in the statement of the Lemma, we thus indeed have

$$L^{0\text{-}1}(z,y) \le L^{\text{unh}}(z,y).$$

Finally, the gradient of the function $-z_y + \frac{1}{C}\sum_{k=1}^{C} z_k$ has norm $\sqrt{\frac{C-1}{C}}$, so the Lipschitz constant of L^{unh} is

$$\sup_z \|\nabla_z L^{\text{unh}}(z,y)\| = a\sqrt{\frac{C-1}{C}}.$$

Simplifying concludes the proof. ■

Proof of Lemma 7.1. Fix y. We start by proving that the infimum of Φ_R on the domain $\|z\| \le R$ equals 0. At a minimizer of $L(z,y)$, we must have $z_y \ge z_k$ for $k \ne y$ (this can be seen by a simple perturbation argument). We may therefore restrict the optimization to the set

$$B(0,R) \cap \{z : z_y \ge z_k \text{ for all } k \ne y\}.$$

This set is convex. Moreover, on this domain the arguments of σ satisfy $z_k - z_y \le 0$, and since σ is convex on $(-\infty, 0]$, the function

$$L(z,y) = \sum_{k \ne y} \sigma(z_k - z_y)$$

is convex on this domain. Both the objective and the constraint are invariant under permutations of the coordinates $\{z_k\}_{k \ne y}$. Therefore, if z is a minimizer, any permutation of the incorrect-class coordinates is also a minimizer. Averaging over all such permutations gives another feasible point satisfying

$$z_k = t \qquad (\forall k \ne y)$$

for some $t \in \mathbb{R}$. By convexity of $L(\cdot, y)$ on the feasible set, this averaged point achieves a value of the objective no larger than the original one. Therefore, there exists a minimizer with all incorrect-class coordinates equal. The optimization problem thus reduces to

$$\min_{s,t \in \mathbb{R}} (C-1)\sigma(t-s) \qquad \text{subject to} \qquad s^2 + (C-1)t^2 \le R^2, \quad s \ge t.$$

Since σ is increasing, this is equivalent to

$$\max_{s,t \in \mathbb{R}} (s-t) \qquad \text{subject to} \qquad s^2 + (C-1)t^2 \le R^2, \quad s \ge t.$$

The solution of this problem can be obtained by using Lagrange multipliers. The infimum of $L(z,y)$ is then found to be given by

$$(C-1)\sigma\left(-R\sqrt{\frac{C}{C-1}}\right) = m_R$$

and (ii) for the bounded domain follows. For the unbounded domain, the result is obtained by observing that the infimum of $L(z,y)$ is 0 on $\mathbb{R}^C$. Applying the same argument to $-L(z,y)$ allows to find the supremum of the function $L(z,y)$ on the domain $\|z\| \le R$. It is given by

$$(C-1)\sigma\left(R\sqrt{\frac{C}{C-1}}\right).$$

Substituting into the expression for $\Phi_R(z,y)$ and simplifying proves (iii) for the bounded domain. For the unbounded domain $\mathbb{R}^C$, the result follows from observing that the supremum of $L(z,y)$ is then $C-1$. We now prove (i). If $L^{0\text{-}1}(z,y)=1$, then $z_k \geq z_y$ for some $k \neq y$, so

$$\sigma(z_k - z_y) \geq \tfrac{1}{2}.$$

Minimizing the remaining terms under $\|z\| \leq R$ gives

$$L(z,y) \geq \frac{1}{2} + (C-2)\sigma\left(-R\sqrt{\frac{C}{2(C-2)}}\right) = q_R$$

(this can be obtained by the same argument as in the proof of (ii)). Therefore,

$$\Phi_R(z,y) = a(R)(L(z,y) - m_R) \geq 1,$$

concluding the proof of (i). In order to prove (iv), we compute the gradient of $L(z,y)$:

$$\nabla_z L(z,y) = \sum_{k \neq y} \sigma'(z_k - z_y)(e_k - e_y).$$

On one hand, the norm of the gradient evaluated at 0 is equal to $\frac{1}{4}\sqrt{C(C-1)}$. On the other hand,

$$\|\nabla_z L(z,y)\|^2 = \sum_{k \neq y} \sigma'(z_k - z_y)^2 + \left(\sum_{k \neq y} \sigma'(z_k - z_y)\right)^2 \leq (C-1)\frac{1}{16} + \left((C-1)\frac{1}{4}\right)^2 = \frac{C(C-1)}{16},$$

where we used that $0 \leq \sigma'(x) \leq 1/4$. We conclude that the Lipschitz constant of $L(z,y)$ is equal to $\frac{1}{4}\sqrt{C(C-1)}$. Multiplying by $a(R)$ then gives the Lipschitz constant of Φ_R.

Finally, in order to prove (v), we compute the Hessian of $L(z,y)$:

$$\nabla_z^2 L(z,y) = \sum_{k \neq y} \sigma''(z_k - z_y)(e_k - e_y)(e_k - e_y)^\top.$$

This Hessian has a convenient simple form. We get the following upper bound on its spectral norm:

$$\|\nabla_z^2 L(z,y)\|_2 \leq \max_{k \neq y} |\sigma''(z_k - z_y)| \cdot \|\sum_{k \neq y} (e_k - e_y)(e_k - e_y)^\top\|_2.$$

The eigenvalues and eigenvectors of $\sum_{k \neq y}(e_k - e_y)(e_k - e_y)^\top$ are straightforward to obtain. We have the eigenvalue 0 with eigenvector $\mathbf{1}$, an eigenvalue C with eigenvector $v = (1, \ldots, 1, \underbrace{-(C-1)}_{y\text{-th coordinate}}, 1, \ldots, 1)$. Also, the subspace defined by

$$v_y = 0, \qquad \sum_{k \neq y} v_k = 0$$

is an eigenspace of dimension $C-2$ associated with an eigenvalue of 1. We conclude that the spectral norm of $\sum_{k \neq y}(e_k - e_y)(e_k - e_y)^\top$ is equal to C. Multiplying by $a(R)$ and using $\sup_{x \in \mathbb{R}} |\sigma''(x)| = \frac{1}{6\sqrt{3}}$ concludes the proof. ■

Proof of Proposition 4.4. Consider the multi-class unhinged surrogate $L_r^{\text{unh}}(z,y)$ with $r = M$. We will omit the subscript r to simplify notation. Then, for any $\overline{S}$,

$$\begin{aligned}
L_D^{0\text{-}1}(\mu_{\overline{S}}^{unh}) = L_D^{0\text{-}1}\left(\frac{\mu_{\overline{S}}^{unh}}{\|\mu_{\overline{S}}^{unh}\|}\right) &\leq L_D^{\text{unh}}\left(\frac{\mu_{\overline{S}}^{unh}}{\|\mu_{\overline{S}}^{unh}\|}\right) \\
&= \frac{1}{1-p}\left[L_{\overline{D}}^{\text{unh}}\left(\frac{\mu_{\overline{S}}^{unh}}{\|\mu_{\overline{S}}^{unh}\|}\right) - \frac{p}{C}\mathop{\mathbb{E}}_{x \sim D_x}\sum_{y=1}^{C} L^{\text{unh}}\left(\frac{\mu_{\overline{S}}^{unh}}{\|\mu_{\overline{S}}^{unh}\|}x, y\right)\right] \\
&= \frac{1}{1-p}\left[L_{\overline{D}}^{\text{unh}}\left(\frac{\mu_{\overline{S}}^{unh}}{\|\mu_{\overline{S}}^{unh}\|}\right) - pb_C\right].
\end{aligned}$$

For any $0 < \sigma < 1$ and with $m = dC$, applying Theorem 4.1 with $\beta = 0$ and $\ell = b_C/M$, we obtain that with probability at least $1 - \delta$ over the draw $\overline{S} \sim \overline{D}^n$,

$$L_D^{0\text{-}1}(\mu_{\overline{S}}^{unh}) \leq \frac{1}{1-p}\left[L_{\overline{S}}^{\text{unh}}\left(\frac{\mu_{\overline{S}}^{unh}}{\|\mu_{\overline{S}}^{unh}\|}\right) - pb_C\right] + \frac{b_C\sqrt{dC+1}}{2(1-p)\sigma\sqrt{n}} + \frac{b_C\sigma\sqrt{dC+1}}{(1-p)\sqrt{n}}\left(\ln\frac{2}{\delta} + \frac{1}{2}\right).$$

The proof is completed by exploiting

$$L_{\overline{S}}^{\text{unh}}(W) = -\frac{b_C}{M}\sqrt{\frac{C}{C-1}}\text{Trace}(\mu_{\overline{S}}^{unh}W^T) + b_C$$

with $W = \frac{\mu_{\overline{S}}^{unh}}{\|\mu_{\overline{S}}^{unh}\|}$ and simplifying. ■

Proof of Corollary 4.5. The absolute loss is 1-Lipschitz in its prediction argument and Proposition 3.2 applies to $\widetilde{L}(z, y) = L(z, y) - L(0, y)$ with $a_w = M\|w\|$. Applying Theorem 3.1 and Lemma 3.8 to $\widetilde{L}$, with confidence parameter $\delta/2$, we obtain that, with probability at least $1 - \delta/2$ over $S \sim D^n$, for all probability distributions Q on $\mathbb{R}^d$,

$$\mathbb{E}_{w\sim Q}L_D(h_w) \leq \mathbb{E}_{w\sim Q}L_S(h_w) + \frac{M\sigma\sqrt{d+1}}{\sqrt{n}}\left[D_{\text{KL}}(Q\|P) + \ln\frac{2}{\delta} + \frac{1}{2}\right] + \mathbb{E}_D[|y|] - \frac{1}{n}\sum_{i=1}^n |y_i|.$$

Applying Hoeffding's inequality with a union bound completes the proof. ■

Proof of Corollary 4.6. For every y, the map $z \mapsto L_\rho(z, y)$ is 1-Lipschitz and $1/\rho$-smooth. We can apply Theorem 4.1 with $C = 1$ (see Remark 4.2), $\ell = 1$ and $\beta = 1/\rho$ to $\widetilde{L}_\rho(z, y) = L_\rho(z, y) - L_\rho(0, y)$. We then obtain that, with probability at least $1 - \delta/2$ over $S \sim D^n$, uniformly for all $w \in \Theta$,

$$L_{\rho,D}(h_w) \leq L_{\rho,S}(h_w) + \frac{2\,R\sigma M^2}{\rho\sqrt{n}} + \frac{4M^2\sigma^2}{\rho}\sqrt{\frac{2\ln(16/\delta)}{n}} + \mathbb{E}_D[L_\rho(0, y)] - \frac{1}{n}\sum_{i=1}^n L_\rho(0, y_i) + \frac{M\sqrt{d+1}}{2\sigma\sqrt{n}}\|w\|^2 + \frac{M\sigma\sqrt{d+1}}{\sqrt{n}}\left(\ln\frac{4}{\delta} + \frac{1}{2}\right).$$

Applying Hoeffding's inequality with a union bound completes the proof. ■

C Technical proofs for Section 6

We will use the following contraction result to prove the linear case.

Lemma C.1 (Contraction inequality in the multi-class case (Maurer, 2016))**.** *Let $x_1, \ldots, x_n$ be fixed and let $\mathscr{F}$ be a class of functions $f : \mathscr{X} \to \mathbb{R}^C$ with components f_k. For each $i \in \{1, \ldots, n\}$, let $\psi_i : \mathbb{R}^C \to \mathbb{R}$ satisfy*

$$|\psi_i(u) - \psi_i(v)| \leq L_i\|u - v\|_2, \qquad \forall u, v \in \mathbb{R}^C,$$

for some constants $L_i > 0$. Then,

$$\mathbb{E}_\varepsilon\left[\sup_{f\in\mathscr{F}}\sum_{i=1}^n \varepsilon_i\,\psi_i(f(x_i))\right] \;\leq\; \sqrt{2}\,\mathbb{E}_{\varepsilon_{ik}}\left[\sup_{f\in\mathscr{F}}\sum_{i=1}^n\sum_{k=1}^C \varepsilon_{ik}\,L_i\,f_k(x_i)\right],$$

where (ε_i) and (ε_{ik}) are independent Rademacher variables.

Proof of Lemma C.1. This is a slight extension to Corollary 4 of Maurer (2016) where we allow the Lipschitz constants to depend on i. Assume $L_i > 0$ for all i. Define $h_i : \mathbb{R}^C \to \mathbb{R}$ by $h_i(z) := \psi_i(z/L_i)$. Then, each h_i is 1-Lipschitz. Let

$$\mathscr{G} := \{g : \mathscr{X} \to \mathbb{R}^C : \exists f \in \mathscr{F} \text{ s.t. } g(x_i) = L_i f(x_i) \ \forall i\}.$$

For any $f \in \mathscr{F}$, letting $g \in \mathscr{G}$ satisfy $g(x_i) = L_i f(x_i)$ gives $h_i(g(x_i)) = \psi_i(f(x_i))$, hence

$$\sup_{f\in\mathscr{F}} \sum_{i=1}^n \varepsilon_i \, \psi_i(f(x_i)) = \sup_{g\in\mathscr{G}} \sum_{i=1}^n \varepsilon_i \, h_i(g(x_i)).$$

Applying Maurer's contraction inequality (Maurer, 2016) to the h_i's yields

$$\mathbb{E}_\varepsilon\Big[\sup_{f\in\mathscr{F}} \sum_{i=1}^n \varepsilon_i \, \psi_i(f(x_i))\Big] \le \sqrt{2}\, \mathbb{E}_{\varepsilon_{ik}}\Big[\sup_{g\in\mathscr{G}} \sum_{i=1}^n \sum_{k=1}^C \varepsilon_{ik} \, g_k(x_i)\Big].$$

Finally, by the definition of $\mathscr{G}$,

$$\sup_{g\in\mathscr{G}} \sum_{i=1}^n \sum_{k=1}^C \varepsilon_{ik} \, g_k(x_i) = \sup_{f\in\mathscr{F}} \sum_{i=1}^n \sum_{k=1}^C \varepsilon_{ik} \, L_i \, f_k(x_i).$$

■

Proof of Theorem 6.1. Let $\overline{W} \in \Theta$ and let $W = \overline{W} + U$, where $U \sim \mathscr{N}(0, \sigma^2 I_{Cd\times Cd})$. For each i, set $z_i = \overline{W} x_i \in \mathbb{R}^C$ and $U_i = U x_i \in \mathbb{R}^C$, so that $h_W(x_i) = z_i + U_i$. By the first-order integral theorem,

$$L(z_i + U_i, y_i) - L(z_i, y_i) = \int_0^1 \langle \nabla_z L(z_i + tU_i, y_i), \, U_i\rangle \, dt,$$

and therefore

$$F_\sigma(\overline{W}; x_i, y_i) = -\mathbb{E}_U \int_0^1 \langle \nabla_z L(z_i + tU_i, y_i), \, U_i\rangle \, dt.$$

Using $\sup \mathbb{E} \le \mathbb{E} \sup$ and $\sup \int \le \int \sup$, we obtain

$$\mathfrak{R}_S(\mathscr{F}_\sigma) \le \int_0^1 \mathbb{E}_{\varepsilon,U}\left[\sup_{\overline{W}\in\Theta} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \Big(- \langle \nabla_z L(z_i + tU_i, y_i), \, U_i\rangle\Big)\right] dt,$$

where $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n) \in \{\pm 1\}^n$ are Rademacher random variables. For fixed t and $U_1, \ldots, U_n$, define $\psi_i(z) := -\langle \nabla_z L(z + tU_i, y_i), U_i\rangle$. By β-smoothness of $L(\cdot, y_i)$,

$$|\psi_i(z) - \psi_i(z')| \le \beta \, \|U_i\| \, \|z - z'\|,$$

so ψ_i is $(\beta\|U_i\|)$-Lipschitz. Applying Maurer's vector contraction inequality (Lemma C.1) to the class $f_{\overline{W}}(x) = \overline{W} x$ with Lipschitz constants $L_i = \beta\|U_i\|$, we obtain

$$\mathfrak{R}_S(\mathscr{F}_\sigma) \le \frac{\sqrt{2}\beta}{n} \, \mathbb{E}_{\varepsilon_{ik},U}\left[\sup_{\|\overline{W}\|_F \le R} \sum_{i=1}^n \sum_{k=1}^C \varepsilon_{ik} \, \|U_i\| \, (\overline{W} x_i)_k\right].$$

The supremum over $\overline{W}$ can be obtained explicitly:

$$\sup_{\|\overline{W}\|_F \le R} \sum_{i=1}^n \sum_{k=1}^C \varepsilon_{ik} \, \|U_i\| \, (\overline{W} x_i)_k = R \Big\| \sum_{i=1}^n \|U_i\| \, \varepsilon_i^{(C)} x_i^\top \Big\|_F,$$

where $\varepsilon_i^{(C)} = (\varepsilon_{i1}, \ldots, \varepsilon_{iC})^\top \in \mathbb{R}^C$. This is true since

$$\begin{aligned}\sum_{i=1}^{n}\sum_{k=1}^{C} \varepsilon_{ik} \|U_i\| (\overline{W}x_i)_k &= \sum_{i=1}^{n} \|U_i\| (\varepsilon_i^{(C)})^\top \overline{W} x_i \\ &= \sum_{i=1}^{n} \left\langle \overline{W}, \|U_i\| \varepsilon_i^{(C)} x_i^\top \right\rangle_F \\ &= \left\langle \overline{W}, \sum_{i=1}^{n} \|U_i\| \varepsilon_i^{(C)} x_i^\top \right\rangle_F,\end{aligned}$$

where $\left\langle \cdot, \cdot \right\rangle_F$ denotes the Frobenius inner product. Therefore, the empirical Rademacher complexity is upper bounded as follows:

$$\mathfrak{R}_S(\mathscr{F}_\sigma) \leq \frac{\sqrt{2}\beta R}{n} \mathbb{E}_{\varepsilon_{ik},U} \Big\| \sum_{i=1}^{n} \|U_i\| \varepsilon_i^{(C)} x_i^\top \Big\|_F.$$

By Jensen's inequality and independence of Rademacher random variables,

$$\mathbb{E}_{\varepsilon_{ik},U} \Big\| \sum_{i=1}^{n} \|U_i\| \varepsilon_i^{(C)} x_i^\top \Big\|_F \leq \sqrt{C \sum_{i=1}^{n} \mathbb{E}_U[\|U_i\|^2] \, \|x_i\|^2}.$$

This is true since

$$\begin{aligned}\mathbb{E}_\varepsilon \Big\| \sum_{i=1}^{n} \|U_i\| \varepsilon_i^{(C)} x_i^\top \Big\|_F^2 &= \sum_{k=1}^{C} \mathbb{E}_\varepsilon \sum_{i,j} \|U_i\| \|U_j\| \varepsilon_{ik}\varepsilon_{jk} \langle x_i, x_j \rangle \\ &= \sum_{k=1}^{C}\sum_{i=1}^{n} \|U_i\|^2 \|x_i\|^2 = C \sum_{i=1}^{n} \|U_i\|^2 \|x_i\|^2.\end{aligned}$$

Since $U_i = Ux_i \sim \mathscr{N}(0, \sigma^2 \|x_i\|^2 I_C)$, we have

$$\mathbb{E}_U \|U_i\|^2 = C\sigma^2 \|x_i\|^2,$$

and therefore

$$\mathbb{E}_{\varepsilon_{ik},U} \Big\| \sum_{i=1}^{n} \|U_i\| \varepsilon_i^{(C)} x_i^\top \Big\|_F \leq \sigma C \sqrt{\sum_{i=1}^{n} \|x_i\|^4}.$$

Substituting into the previous bound gives

$$\mathfrak{R}_S(\mathscr{F}_\sigma) \leq \frac{\sqrt{2}\beta R \sigma C}{n} \sqrt{\sum_{i=1}^{n} \|x_i\|^4}.$$

If $\|x_i\| \leq M$ for all i, then $\sum_{i=1}^{n} \|x_i\|^4 \leq nM^4$, and thus

$$\mathfrak{R}_S(\mathscr{F}_\sigma) \leq \frac{\sqrt{2}\beta R \sigma C \, M^2}{\sqrt{n}}.$$

■

Proof of Lemma 6.2. Fix i and abbreviate $z_i(w) := h_w(x_i)$ and $g_i(w) := \nabla_z L(z_i(w), y_i) \in \mathbb{R}^C$. By the chain rule,

$$\nabla_w L(z_i(w), y_i) = J_i(w)^\top g_i(w).$$

Differentiating under the expectation gives

$$\nabla F_i(w) = J_i(w)^\top g_i(w) - \mathbb{E}_u\left[J_i(w+\sigma u)^\top g_i(w+\sigma u)\right].$$

Add and subtract $\mathbb{E}_u[J_i(w)^\top g_i(w+\sigma u)]$ to obtain

$$\nabla F_i(w) = J_i(w)^\top \Big(g_i(w) - \mathbb{E}_u g_i(w+\sigma u)\Big) + \mathbb{E}_u\left[(J_i(w) - J_i(w+\sigma u))^\top g_i(w+\sigma u)\right].$$

Fix $v \in \mathbb{R}^m$ and write $\langle \nabla F_i(w), v\rangle = T_{1,i}(w,v) + T_{2,i}(w,v)$ with

$$\begin{aligned} T_{1,i}(w,v) &:= \langle g_i(w) - \mathbb{E}_u g_i(w+\sigma u),\ J_i(w)v\rangle, \\ T_{2,i}(w,v) &:= \mathbb{E}_u \langle g_i(w+\sigma u),\ (J_i(w) - J_i(w+\sigma u))v\rangle. \end{aligned}$$

Then, $(a+b)^2 \le 2a^2 + 2b^2$ gives

$$\langle \nabla F_i(w), v\rangle^2 \le 2T_{1,i}(w,v)^2 + 2T_{2,i}(w,v)^2.$$

We first bound $T_{1,i}$. By Cauchy–Schwarz and β-smoothness of $L(z,y)$,

$$|T_{1,i}(w,v)| \le \|J_i(w)v\|\, \mathbb{E}_u\|g_i(w) - g_i(w+\sigma u)\| \le \beta\, \|J_i(w)v\|\, \mathbb{E}_u\|z_i(w) - z_i(w+\sigma u)\|.$$

Using the fundamental theorem of calculus,

$$z_i(w+\sigma u) - z_i(w) = \sigma \int_0^1 J_i(w + t\sigma u)\, u\, dt,$$

so Jensen's inequality gives

$$\mathbb{E}_u\|z_i(w+\sigma u) - z_i(w)\|^2 \le \sigma^2 \int_0^1 \mathbb{E}_u\left[u^\top J_i(w+t\sigma u)^\top J_i(w+t\sigma u) u\right] dt \le \sigma^2 A_{i,\sigma}.$$

Thus,

$$\mathbb{E}_u\|z_i(w+\sigma u) - z_i(w)\| \le \sigma\sqrt{A_{i,\sigma}}.$$

Since $\|J_i(w)v\|^2 = v^\top J_i(w)^\top J_i(w) v \le v^\top G_i v$, we obtain

$$2T_{1,i}(w,v)^2 \le 2\beta^2\sigma^2 A_{i,\sigma}\, v^\top G_i v.$$

We now bound $T_{2,i}$. Since $L(z, y_i)$ is ℓ-Lipschitz and differentiable, $\|g_i(w)\| \le \ell$, so by Cauchy–Schwarz and Jensen,

$$|T_{2,i}(w,v)| \le \mathbb{E}_u\left[\|g_i(w+\sigma u)\|\, \|(J_i(w+\sigma u) - J_i(w))v\|\right] \le \ell\, \sqrt{\mathbb{E}_u\|(J_i(w+\sigma u) - J_i(w))v\|^2}.$$

Applying the fundamental theorem of calculus yields

$$J_i(w+\sigma u) - J_i(w) = \sigma \int_0^1 \nabla_w^2 h_{w+t\sigma u}(x_i)[u]\, dt,$$

where

$$\nabla_w^2 h_{w'}(x_i)[u] := \begin{pmatrix} (H_{i,1}(w')u)^\top \\ \vdots \\ (H_{i,C}(w')u)^\top \end{pmatrix}.$$

By Jensen's inequality,

$$\|(J_i(w+\sigma u) - J_i(w))v\|^2 \le \sigma^2 \int_0^1 \left\|\nabla_w^2 h_{w+t\sigma u}(x_i)[u]v\right\|^2 dt.$$

For any w', the vector $\nabla_w^2 h_{w'}(x_i)[u]v \in \mathbb{R}^C$ has k-th coordinate $u^\top H_{i,k}(w')v$. Therefore,

$$\left\|\nabla_w^2 h_{w'}(x_i)[u]v\right\|^2 = \sum_{k=1}^{C}(u^\top H_{i,k}(w')v)^2 = v^\top \left(\sum_{k=1}^{C} H_{i,k}(w')^\top uu^\top H_{i,k}(w')\right) v.$$

Consequently,

$$\mathbb{E}_u\|(J_i(w+\sigma u) - J_i(w))v\|^2 \le \sigma^2 v^\top B_{i,\sigma}(w)v \le \sigma^2 v^\top B_{i,\sigma} v.$$

Therefore,

$$2T_{2,i}(w,v)^2 \le 2\ell^2\, \mathbb{E}_u\|(J_i(w+\sigma u) - J_i(w))v\|^2 \le 2\ell^2\sigma^2\, v^\top B_{i,\sigma} v.$$

Combining the bounds gives, for all $w \in \Theta$ and all $v \in \mathbb{R}^m$,

$$\langle \nabla F_i(w), v\rangle^2 \le v^\top \Big(2\beta^2\sigma^2 A_{i,\sigma} G_i + 2\ell^2\sigma^2 B_{i,\sigma}\Big) v.$$

Now fix $w_1, w_2 \in \Theta$, set $v = w_1 - w_2$, and define $w_t = w_2 + tv$. By the fundamental theorem of calculus,

$$F_i(w_1) - F_i(w_2) = \int_0^1 \langle \nabla F_i(w_t), v\rangle\, dt,$$

and Jensen's inequality implies

$$(F_i(w_1) - F_i(w_2))^2 \le \int_0^1 \langle \nabla F_i(w_t), v\rangle^2\, dt.$$

Using the bound above yields

$$(F_i(w_1) - F_i(w_2))^2 \le v^\top \Big(2\beta^2\sigma^2 A_{i,\sigma} G_i + 2\ell^2\sigma^2 B_{i,\sigma}\Big) v.$$

Summing over i gives

$$d_S(w_1, w_2)^2 \le v^\top \left(2\sigma^2 \sum_{i=1}^{n} \Big(\beta^2 A_{i,\sigma} G_i + \ell^2 B_{i,\sigma}\Big)\right) v,$$

which is the claimed bound. ■

Proposition C.2 (Talagrand comparison inequality (Vershynin, 2018, Cor. 8.5.6))**.** *Let $(X_t)_{t\in T}$ be a mean-zero random process and $(Y_t)_{t\in T}$ a mean-zero Gaussian process. Assume that for all $t, s \in T$,*

$$\|X_t - X_s\|_{\psi_2} \le K\|Y_t - Y_s\|_{L_2}.$$

Then,

$$\mathbb{E}\sup_{t\in T} X_t \le cK\, \mathbb{E}\sup_{t\in T} Y_t,$$

where $c > 0$ is an absolute constant. For a real-valued random variable X, the L_2 norm is defined as

$$\|X\|_{L_2} := \big(\mathbb{E}[X^2]\big)^{1/2}.$$

The ψ_2 norm of X is defined by

$$\|X\|_{\psi_2} := \inf\Big\{t > 0 \;:\; \mathbb{E}\Big[\exp\Big(\frac{X^2}{t^2}\Big)\Big] \le 2\Big\}.$$

Proof of Proposition 6.3. The proof is a direct application of Proposition C.2 together with a standard computation of the Gaussian width of an ellipsoid (see e.g. (Vershynin, 2018, Chapter 7)). Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables and define

$$X_w := \sum_{i=1}^{n} \varepsilon_i \, f\big(w, (x_i, y_i)\big).$$

Then, $\mathbb{E}_\varepsilon[X_w] = 0$ for every w and

$$\mathfrak{R}_S(\mathscr{F}) = \frac{1}{n}\mathbb{E}_\varepsilon\Big[\sup_{w\in\Theta} X_w\Big].$$

For $w_1, w_2 \in \Theta$, set $\Delta_i := f(w_1, (x_i, y_i)) - f(w_2, (x_i, y_i))$ and $Z := \sum_{i=1}^n \varepsilon_i \Delta_i$. The random variable Z is subgaussian with variance proxy $\sum_{i=1}^n \Delta_i^2 = d_S(w_1, w_2)^2$. This is true since the moment generating function of Z satisfies

$$\mathbb{E}_\varepsilon e^{\lambda Z} = \prod_{i=1}^{n} \mathbb{E}_\varepsilon e^{\lambda \varepsilon_i \Delta_i} = \prod_{i=1}^{n} \cosh(\lambda \Delta_i) \le \prod_{i=1}^{n} \exp\Big(\frac{\lambda^2 \Delta_i^2}{2}\Big) = \exp\Big(\frac{\lambda^2}{2}\sum_{i=1}^{n} \Delta_i^2\Big),$$

where we used $\cosh(u) \le e^{u^2/2}$ for all $u \in \mathbb{R}$. It follows that there exists a constant $c_0 > 0$ such that $\|Z\|_{\psi_2} \le c_0 \sqrt{\sum_{i=1}^n \Delta_i^2} = c_0 d_S(w_1, w_2)$ (see for example exercice 2.40 from Vershynin (2018)). From our assumption on $d_S(w_1, w_2)$, we therefore have

$$\|X_{w_1} - X_{w_2}\|_{\psi_2} \le c_0 \sqrt{(w_1 - w_2)^\top M_S (w_1 - w_2)}.$$

In order to apply the Talagrand comparison inequality, we need to specify a mean-zero Gaussian process $(Y_w)_{w\in\Theta}$. Let $g \sim \mathscr{N}(0, I_m)$ and define the Gaussian process

$$Y_w := \langle g, M_S^{1/2} w\rangle, \qquad w \in \Theta.$$

For $\Delta := w_1 - w_2$ and using $\mathbb{E}[gg^\top] = I_m$, we get

$$\|Y_{w_1} - Y_{w_2}\|_{L_2}^2 = \mathbb{E}\big(g^\top M_S^{1/2}\Delta\big)^2 = \mathbb{E}\big[\Delta^\top M_S^{1/2} g g^\top M_S^{1/2}\Delta\big] = \Delta^\top M_S \Delta.$$

Hence, $\|Y_{w_1} - Y_{w_2}\|_{L_2} = \sqrt{(w_1 - w_2)^\top M_S (w_1 - w_2)}$. We can therefore apply the Talagrand comparison inequality with $K = c_0$. Thus, there exists a constant $c_1 > 0$ such that

$$\mathbb{E}_\varepsilon \sup_{w\in\Theta} X_w \le c_1 c_0 \, \mathbb{E}_g \sup_{w\in\Theta} Y_w.$$

Furthermore,

$$\mathbb{E}_g \sup_{w\in\Theta} Y_w = R\,\mathbb{E}\|M_S^{1/2} g\|_2 \le R\sqrt{\mathbb{E}\|M_S^{1/2} g\|_2^2} = R\sqrt{\mathrm{Trace}(M_S)},$$

Combining the above inequalities and dividing by n yields

$$\mathfrak{R}_S(\mathscr{F}) = \frac{1}{n}\mathbb{E}_\varepsilon \sup_{w\in\Theta} X_w \le \frac{c_1 c_0}{n} R\sqrt{\mathrm{Trace}(M_S)} = c\,\frac{R}{n}\sqrt{\mathrm{Trace}(M_S)},$$

with $c := c_1 c_0 > 0$ a constant. ■

Proof of Corollary 6.4. By Lemma 6.2, the pseudo-metric induced by the Jensen gap class is dominated by the quadratic form associated with

$$M_{S,\sigma} = 2\sigma^2 \sum_{i=1}^{n} \Big(\beta^2 A_{i,\sigma} G_i + \ell^2 B_{i,\sigma}\Big).$$

Applying Proposition 6.3 gives

$$\mathfrak{R}_S(\mathscr{F}_\sigma) \le c\frac{R}{n}\sqrt{\mathrm{Trace}(M_{S,\sigma})}.$$

Since

$$\mathrm{Trace}(M_{S,\sigma}) = 2\sigma^2 n\left(\beta^2 J_{S,\sigma} + \ell^2 H_{S,\sigma}\right),$$

the result follows after absorbing the factor $\sqrt{2}$ into the universal constant c. ■

Remark C.3 (Explicit constant). Corollary 6.4 leaves us with an unspecified universal constant c. In the present setting, it is possible to alternatively obtain the explicit constant $c = \sqrt{\pi}$, which is useful for numerical evaluation. First, we have the following comparison between the Rademacher and Gaussian complexity:

$$\mathbb{E}_\varepsilon \sup_{w\in\Theta} X_w \le \sqrt{\frac{\pi}{2}}\,\mathbb{E}_g \sup_{w\in\Theta}\sum_{i=1}^n g_i f(w,(x_i,y_i)).$$

This follows by writing $g_i = \varepsilon_i|g_i|$ and applying Jensen's inequality together with $\mathbb{E}|g_i| = \sqrt{2/\pi}$. Second, for every $w_1, w_2 \in \Theta$, the L_2 norm satisfies

$$\Big\|\sum_{i=1}^n g_i\big(f(w_1,(x_i,y_i)) - f(w_2,(x_i,y_i))\big)\Big\|_{L_2} = d_S(w_1,w_2).$$

Third, by assumption, this quantity is upper bounded by the L_2 norm of $Y_{w_1} - Y_{w_2}$. Finally, by the Gaussian comparison inequality (Sudakov–Fernique; see (Vershynin, 2018, Theorem 7.2.8)),

$$\mathbb{E}_g \sup_{w\in\Theta}\sum_{i=1}^n g_i f(w,(x_i,y_i)) \le \mathbb{E}_g \sup_{w\in\Theta} Y_w,$$

which yields the result after considering the factor $\sqrt{2}$ that was absorbed into the universal constant c.

Proof of Proposition 6.6. Using Theorem 26.5 from Shalev-Shwartz & Ben-David (2014) and Corollary 6.4, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over $S \sim D^n$, we have uniformly for all $\overline{w} \in \Theta$,

$$\begin{aligned}
\mathbb{E}_{(x,y)\sim D}\Big[\mathrm{Jen}_Q\big(L(h_w(x),y)\big)\Big] &\le \frac{1}{n}\sum_{i=1}^n \mathrm{Jen}_Q\big(L(h_w(x_i),y_i)\big) \\
&\quad + \; 2\,\mathfrak{R}_S(\mathscr{F}_\sigma) \; + \; 8\sigma^2 B\sqrt{\frac{2\ln(4/\delta)}{n}} \\
&\le \frac{1}{n}\sum_{i=1}^n \mathrm{Jen}_Q\big[L(h_w(x_i),y_i)\big] \\
&\quad + \; 2c\,\frac{R\sigma}{\sqrt{n}}\sqrt{\beta^2 J_{S,\sigma} + \ell^2 H_{S,\sigma}} \; + \; 8\sigma^2 B\sqrt{\frac{2\ln(4/\delta)}{n}},
\end{aligned}$$

concluding the proof. ■

Proof of Proposition 6.7. This is a standard union bound argument. For fixed $\sigma > 0$ and $\eta \in (0,1)$, define

$$\mathscr{B}(\sigma,\eta;\overline{w}) := \frac{1}{n}\sum_{j=1}^n \mathrm{Jen}_{Q_{\overline{w},\sigma}}\big(L(h_w(x_j),y_j)\big) + 2c\,\frac{R\sigma}{\sqrt{n}}\sqrt{\beta^2 J_{S,\sigma} + \ell^2 H_{S,\sigma}} + 8\sigma^2 B\sqrt{\frac{2\ln(4/\eta)}{n}}.$$

By Proposition 6.6, for every fixed σ and η,

$$\mathbb{P}_S\Big(\forall \overline{w}\in\Theta:\; \mathbb{E}\Big[\mathrm{Jen}_{Q_{\overline{w},\sigma}}\big(L(h_w(x),y)\big)\Big] \le \mathscr{B}(\sigma,\eta;\overline{w})\Big) \ge 1-\eta.$$

For $i \geq 1$, choosing $\eta = \delta_i := \delta/(2i^2)$ and $\sigma = \sigma_i$, we then have

$$\mathbb{P}_S\Big(\forall \overline{w} \in \Theta : \ \mathbb{E}\Big[\mathrm{Jen}_{Q_{\overline{w},\sigma_i}}\big(L(h_w(x), y)\big)\Big] \leq \mathscr{B}(\sigma_i, \delta_i; \overline{w})\Big) \geq 1 - \delta_i.$$

Taking a union bound over $i \geq 1$ and using $\sum_{i\geq 1} \delta_i \leq \delta$ gives that with probability at least $1 - \delta$, the inequality holds simultaneously for all $i \geq 1$ and all $\overline{w} \in \Theta$. It remains to simplify the confidence term:

$$\ln\Big(\frac{4}{\delta_i}\Big) = \ln\Big(\frac{8i^2}{\delta}\Big) = \ln\Big(\frac{8}{\delta}\Big) + 2\ln(i).$$

■

D Implementation and experimental details for the regularizer

Since, for any matrix A,

$$\|A\|_F^2 = \mathrm{Tr}(AA^\top) = \mathbb{E}_v\, v^\top AA^\top v = \mathbb{E}_v\, \|A^\top v\|_2^2,$$

for any random vector v satisfying $\mathbb{E}[vv^\top] = I$, the squared Frobenius norms in $J(w_{\mathrm{eff},B'}, S_{B'})$ and $H(w_{\mathrm{eff},B'}, S_{B'})$ can be estimated using random matrix-vector products. In practice, this only requires Jacobian-vector and Hessian-vector products, and avoids explicitly forming or storing the full Jacobian or Hessian matrices.

To estimate the Jacobian term, we sample M independent random vectors $v_1, \ldots, v_M$ in the output space of the minibatch score vector

$$\big(h_{w_{\mathrm{eff},B'}}(x_1), \ldots, h_{w_{\mathrm{eff},B'}}(x_{B'})\big),$$

and compute

$$\nabla_{w_{\mathrm{eff},B'}}\left[\frac{1}{\sqrt{B'}}\Big\langle v_m,\ \big(h_{w_{\mathrm{eff},B'}}(x_1), \ldots, h_{w_{\mathrm{eff},B'}}(x_{B'})\big)\Big\rangle\right], \qquad m = 1, \ldots, M.$$

The average of the squared norms of these quantities gives a stochastic estimate of the corresponding squared Frobenius norm.

To estimate the Hessian term, for each m, we additionally sample an independent random vector u_m in parameter space and differentiate, with respect to $w_{\mathrm{eff},B'}$, the scalar product between the vector-Jacobian product above and u_m. This yields the required Hessian-vector product without explicitly forming the Hessian. During training, these stochastic estimates are kept in the computational graph, so PyTorch autograd can differentiate through the vector-Jacobian and Hessian-vector products and optimize the resulting regularized objective.

We train an eight-layer convolutional neural network on CIFAR-10, where the eight layers consist of six convolutional layers and two fully connected layers. Each convolutional layer and the first fully connected layer are followed by a BatchNorm layer and a centered Softplus activation. The loss function is the cross-entropy and training is performed with Adam, using a cosine annealing learning-rate schedule preceded by a linear warm-up phase of 3 epochs. The total number of epochs is set to 40, except in the experiments that adjust the training time as a function of the batch size. In that case, for a batch size $B \in \{256, 512, 1024, 2048, 4096\}$, we use

$$40\sqrt{B/256}$$

epochs. For each batch size, we first tune the learning rate for the baseline model. We then keep this learning rate fixed and tune only the regularization coefficient for the JH-regularized model. The learning rate grid consists of 10 logarithmically spaced values between 10^{-3} and 10^{-1}, generated using `np.geomspace`. The regularization grid consists of 10 logarithmically spaced values between 10^{-5} and 5×10^{-4}, except for the $\mathrm{JH}(w/\|w\|; S_{B'})$ method in the first experiment. For this method, the appropriate search grid consisted of 10 logarithmically spaced values between 10^{-7} and 5×10^{-6}. Hyperparameters are selected using validation accuracy on 5000 examples and the final selected configurations are rerun over three seeds. In all experiments,

the regularizer is computed on a subset minibatch of size $B' = 50$ with $M = 4$ random vectors except for the experiments reported in Table 3 below, where we investigate the effect of these hyperparameters under the variable-epoch protocol with batch size $B = 4096$. Increasing B' led to a small improvement in test accuracy, while increasing M beyond 4 did not provide a clear additional benefit.

Table 3: CIFAR-10 test accuracy under the variable-epoch protocol with batch size $B = 4096$. We compare the baseline with JH-regularized models using different regularizer minibatch sizes B' and different numbers M of random vectors. Results are reported as mean $\pm$ standard deviation over three seeds.

Method	B'	M	Test accuracy (%)
Baseline	–	–	78.84 ± 0.21
$\mathrm{JH}(w_{\mathrm{eff},B'}; S_{B'})$	50	1	81.19 ± 0.43
$\mathrm{JH}(w_{\mathrm{eff},B'}; S_{B'})$	50	4	81.46 ± 0.46
$\mathrm{JH}(w_{\mathrm{eff},B'}; S_{B'})$	50	16	81.40 ± 0.18
$\mathrm{JH}(w_{\mathrm{eff},B'}; S_{B'})$	256	4	81.68 ± 0.37
$\mathrm{JH}(w_{\mathrm{eff},B'}; S_{B'})$	1024	4	81.73 ± 0.11

Remark D.1. A natural idea suggested by the generalization bounds would be to combine the JH regularizer with a norm penalty on the effective weights $w_{\mathrm{eff},B}$. However, based on our preliminary experiments, this did not meaningfully improve performance compared with using only the JH regularizer in our experimental setting. In some sense, the effective parameters $w_{\mathrm{eff},B}$ are not arbitrary because of the BatchNorm layers. Indeed, BatchNorm induces some control over the scale of the inputs fed to the nonlinearities, although this control is data dependent and comes with the caveat that the affine BatchNorm parameters can partially undo it. This type of control may already be sufficient in some practical settings to make additional norm control unnecessary. We leave a more thorough investigation of these considerations to future work.