A PAC-Bayesian Perspective on the Interpolating Information Criterion

Anonymous Author(s)
Affiliation
Address
email

Abstract

Deep learning is renowned for its theory–practice gap, whereby principled theory typically fails to provide much beneficial guidance for implementation in practice. This has been highlighted recently by the benign overfitting phenomenon: when neural networks become sufficiently large to interpolate the dataset perfectly, model performance appears to improve with increasing model size, in apparent contradiction with the well-known bias–variance tradeoff. While such phenomena have proven challenging to theoretically study for general models, the recently proposed Interpolating Information Criterion (IIC) provides a valuable theoretical framework to examine performance for overparameterized models. Using the IIC, a PAC-Bayes bound is obtained for a general class of models, characterizing factors which influence generalization performance in the interpolating regime. From the provided bound, we quantify how the test error for overparameterized models achieving effectively zero training error depends on the quality of the implicit regularization imposed by e.g. the combination of model, optimizer, and parameter-initialization scheme; the spectrum of the empirical neural tangent kernel; curvature of the loss landscape; and noise present in the data.

1 Introduction

A prominent curiosity in modern machine learning is the occurrence of strong generalization performance, even in the *overparameterized* setting where the number of parameters exceeds the size of the training set and models can *interpolate* even noisy data [8, 51]. This is at odds with classical theoretical arguments in line with the bias–variance tradeoff, as interpolators are typically thought to correspond to high-variance estimators in the presence of data noise, and therefore should perform poorly [19, §2.9]. Such observations have sparked renewed interest in *interpolating estimators* and the occurrence of *benign overfitting* [5, 14, 46]. One of the more celebrated realizations of benign overfitting is the *double descent* curve particularly pronounced in linear regression [8, 12, 18, 28], where model mean-squared error *decreases* monotonically in the overparameterized regime. This surprising behaviour arises due to the *implicit regularization* present in the choice of estimator [37]. However, rigorous theoretical examination of these phenomena beyond the linear setting remains a significant challenge. For example, analogous curves can become arbitrarily complicated in the kernel regression setting [10, 27, 30].

The problem of *model selection* becomes exacerbated in the overparameterized setting: how do we compare between interpolators? Classically, model selection is conducted using an *information criterion*, the most prominent of which are the AIC and BIC [26], although these all break down for overparameterized models. One recent approach to model selection in the *general* overparameterized setting is presented in [22] with the *Interpolating Information Criterion* (IIC).[1] Adopting a Bayesian

[1] This is related to but substantially more general than previous work in the linear/kernel setting [21].

setup, performance for the IIC is measured in terms of the *marginal likelihood*. Similar to [17], the interpolating regime is examined through the *cold posterior* scenario, where the temperature of the likelihood is decreased to concentrate posterior mass onto the zero-loss set of parameters. The IIC itself relies on a novel (and broadly applicable) principle of *Bayesian duality* [22]: for any overparameterized model, there exists a corresponding underparameterized model with the same marginal likelihood. Conveniently, this corresponding underparameterized model is often amenable to asymptotic approximations via Laplace's method, resulting in a tractable form of the marginal likelihood, even for complex models.

The IIC is theoretically interesting, but its utility may not be immediately obvious. While the marginal likelihood is a standard in Bayesian statistics, it is not often a metric of choice for machine learning practitioners. Several deficiencies in the marginal likelihood have been raised as detrimental to accurate examination of model quality [32]. Instead, a more popular framework for assessing model performance using Bayesian ideas is that of *PAC-Bayes bounds* [3]. These bounds on the true risk are often more straightforward to interpret in practice, and provide the tightest estimates of the test error to date [13, 31]. However, PAC-Bayes bounds are often limited by their requirement of a tractable choice of prior. Hence, previous bounds have only been capable of revealing coarse attributes (such as norm-based metrics [38, 39]) linked to generalization through specific choices of the prior [24].

Using techniques from the derivation of the IIC in [22], we construct a PAC-Bayes bound that holds for a very wide class of models *and* priors in the general overparameterized setting. In doing so, we provide a precise and *complete* characterization of how model performance for interpolators depends on the quality of the implicit regularization, the sharpness of the model about the estimator, the curvature of the zero-loss region in the loss landscape, and the noise of the data. While earlier attempts have been made to develop PAC-Bayesian generalization bounds in the cold posterior setting [41], these are again limited by strong simplifying assumptions. In constrast, our PAC-Bayes bound holds for a general class of interpolators, with minimal assumptions on the regularity of the model.

2 Interpolating Regime

Parameter estimators for regression problems are typically minimizers of an empirical risk L_n:

$$\theta^\star \in \mathcal{M} := \operatorname*{arg\,min}_{\theta\in\Theta} L_n(\theta), \quad \text{where} \quad L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i,\theta), y_i),$$

where $x_1, \dots, x_n \in \mathcal{X}$ are inputs, $y_1, \dots, y_n \in \mathbb{R}^m$ are the corresponding outputs, and $\Theta \subset \mathbb{R}^d$ is the parameter space. For example, in deep learning, $f : \mathcal{X} \times \mathbb{R}^d \to \mathbb{R}^m$ prescribes a (nonlinear) neural network architecture with d weights and m outputs over the input space $\mathcal{X}$. For simplicity, assume $\Theta = \mathbb{R}^d$ and restrict our attention to the mean-squared loss $\ell(z, y) = \|z - y\|^2$, although more general loss functions can also be considered.

When the number of parameters d exceeds the size of the dataset mn and the model can *interpolate* the data exactly, $\mathcal{M}$ is often uncountable. So, which $\theta \in \mathcal{M}$ should be chosen? A convenient approach is to select an estimator within $\mathcal{M}$ that is the solution to a constrained optimization problem involving a regularizer R [7].

Definition 1. *An interpolator is an estimator of the form $\theta^\star = \arg\min_{\theta\in\Theta} R(\theta)$ subject to $f(x_i, \theta) = y_i$ for all $i = 1, \dots, n$, where $R : \mathbb{R}^d \to \mathbb{R}$.*

We assume that R has both a unique minimizer over $\mathbb{R}^d$ *and* a unique minimizer over $\mathcal{M}$. As observed in [22], interpolators as prescribed in Definition 1 arise naturally in the Bayesian context, as we now demonstrate. To start, consider the usual Bayesian posterior $\rho_\gamma(\theta) \propto \exp(-\frac{1}{\gamma}L_n(\theta))\pi(\theta)$ formed from the Gibbs likelihood with temperature γ and prior π. In the limit as $\gamma \to 0^+$, ρ_γ will concentrate around regions where $L_n(\theta)$ is minimized, namely $\mathcal{M}$[2]. This is the *cold posterior* setting, which has surprisingly been observed to enhance predictive performance [48]. In this setting, the role of π in prescribing mass to estimators on $\mathcal{M}$ is enhanced. If we now choose $\pi(\theta) \propto \exp(-\frac{1}{\tau}R(\theta))$ to be the Gibbs measure corresponding to R with temperature τ, then regions with high probability under π correspond to smaller values of R. In this way, R acts as a regularizer over the set of interpolators. By taking $\tau \to 0^+$, the cold posterior concentrates on the minimizer $\theta^\star$ of R on $\mathcal{M}$.

[2] This limiting behaviour in the posterior was quantified and investigated in [11].

The order of the limits ($\gamma \to 0^+$, and then $\tau \to 0^+$) is of great importance, as the limiting behaviour of ρ_γ varies greatly depending on the relative rates with which these two limits are taken [15]. For example, if γ and τ reduce at the same rate, then ρ_γ concentrates around a *maximum a posteriori* (MAP) estimator, and if $\tau \to 0^+$ first, then ρ_γ concentrates on the unique minimizer of R.

In modern machine learning, the limit $\gamma \to 0^+$ corresponds to the procedure of optimizing the model to zero empirical risk. These asymptotics can be equally viable for models which achieve *sufficiently small* training error. If $\theta^\star$ represents the trained weights, then R is the implicit regularizer of the model, comprising all the factors (choice and hyperparameters of the optimizer, initialization, etc.) which dictate the particular solution reached at the end of training. By examining the true risk over the posterior ρ_γ under the limits $\gamma \to 0^+$, and then $\tau \to 0^+$, we reveal a localized estimate of test error for large neural networks at the end of training. Within $\mathcal{M}$, R is the primary measurement distinguishing between estimators, and plays an analogous role to the “risk” in the bound to follow.

3 PAC-Bayes Bounds

Performance in machine learning is typically analyzed through the *true risk function*, $L(\theta)$. Assuming that each (x_i, y_i) is an iid realization from a distribution $\mathcal{D}$, we let $L(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\ell(f(x,\theta), y)$. The difference between $L_n(\theta)$ and $L(\theta)$ is referred to as *generalization error*. A small value for the error for well-trained models is typically thought to be indicative of good real-world performance, and hence high model quality. A Bayesian analogue of the PAC framework, called *PAC-Bayes theory*, was first introduced by McAllester [36] and has become recognized as a promising approach for potentially-practical non-vacuous bounds on the generalization error [13]. Following [6], and letting $\mathrm{KL}(\cdot\|\cdot)$ denote the Kullback–Leibler divergence, the Donsker–Varadhan change of measure theorem [6, Lemma 3] applied to two probability measures, ρ and π, states that

$$\mathbb{E}_{\theta\sim\rho}\phi(\theta) \leq \mathrm{KL}(\rho\|\pi) + \log \mathbb{E}_{\theta\sim\pi} e^{\phi(\theta)}, \qquad \text{for any } \phi : \Theta \to \mathbb{R}.$$

If ϕ is n-times the generalization error, then using Markov's inequality, with probability at least $1-\delta$ over the choice of $(x_i, y_i) \overset{\text{iid}}{\sim} \mathcal{D}$, it holds that

$$\mathbb{E}_{\theta\sim\rho}L(\theta) \leq \mathbb{E}_{\theta\sim\rho}L_n(\theta) + \frac{1}{n}\left[\mathrm{KL}(\rho\|\pi) + \log \mathbb{E}_{(x,y)\sim\mathcal{D}}\zeta_n + \log(1/\delta)\right], \tag{1}$$

where $\zeta_n = \mathbb{E}_{\theta\sim\pi} e^{n(L(\theta)-L_n(\theta))}$ encodes dispersion in the loss under the prior due to noise in the data. While (1) holds for arbitrary measures ρ and π, the bound is tightest when ρ is the posterior ρ_γ with $\gamma = 1/n$ [3, §2.1]. Equation (1) is the core PAC-Bayes bound: to minimize the true risk, one should optimize over the empirical risk, and choose a prior that is as close to the posterior as possible. Effective choices of priors have resulted in non-vacuous generalization bounds, even for moderately large-scale neural networks [13, 31]. However, the Kullback–Leibler term is often intractable for arbitrary priors, and so π is typically chosen to render the right-hand side explicitly computable. There are two major issues with this: (i) the true role of the implicit regularization—believed to be *critical* [37, 51]—remains opaque, as only a simplified version of this regularization can be examined; and (ii) the bound can only be as tight as one can approximate the optimal choice of prior.

An alternative approach is to trade strict upper bounds for *asymptotics* in the interpolating regime using the techniques of [22], along with the observations of [16]. By doing so, a tractable PAC-Bayes bound is obtained for almost any choice of prior, opening the door to a more precise theoretical understanding of regularization, and potentially tighter bounds. The resulting PAC-Bayes bound depends on the performance of the interpolator $\theta^\star$ under the regularizer R, and it makes explicit the dependence of model performance on three key factors:

- **Sharpness:** $S = \log\det(DF(\theta^\star)DF(\theta^\star)^\top)$; where $DF(\theta)$ is the $nm \times d$ Jacobian of F with rows $(\nabla_\theta F(x_i, \theta))_{i=1}^n$. Sharpness measures are well-known to (sometimes [49]) correlate with performance [20, 25]. Note that S is the log-determinant of the *empirical neural tangent kernel* (NTK) [23, 40].
- **Dispersion:** $P = 2n^{-1}\log \mathbb{E}_{(x,y)\sim\mathcal{D}} e^{n(L(\theta_0)-L_n(\theta_0))}$; where $\theta_0 \in \mathbb{R}^d$ is the assumed global minimizer of the regularizer R. Note that if $\ell(f(x,\theta_0), y)$ is normally distributed, then P is its variance. However, as P does not depend on the posterior, and so plays a limited role in our bound.
- **Curvature:** $K = \log\det_+ \nabla^2_{\mathcal{M}} R(\theta^\star) - \log\det \nabla^2 R(\theta_0)$; where $\det_+$ is the pseudo-determinant (product of all non-zero eigenvalues) and $\nabla^2_{\mathcal{M}}$ is the manifold Hessian over $\mathcal{M}$ [1, §5.5]. The

manifold Hessian over $\mathcal{M}$ is well-defined according to the Implicit Function Theorem, which asserts that $\mathcal{M}$ is a submanifold of dimension $d - mn$ if DF is continuous and full rank on $\mathbb{R}^d$.

The interpolating information criterion as presented in [22] is given in terms of these factors as

$$\mathrm{IIC} = \log[R(\theta^\star) - R(\theta_0)] + \frac{S+K}{mn} - \log(mn), \tag{2}$$

where a smaller IIC is indicative of better model performance. For more details on the nature of these factors, see [22, §5.1]. Following [22], our analysis operates under the following conditions. Of these, (F) is perhaps the most unusual condition, but is necessary to ensure that $\mathbb{E}_{\theta\sim\rho_\gamma} L(\theta) \neq 0$.

Assumptions. *Assume the following conditions:*

(A) F and R are $\mathcal{C}^\infty$-smooth on $\mathbb{R}^d$

(B) $\mathcal{M}$ is non-empty, and $DF(\theta)$ is full rank for all $\theta \in \mathbb{R}^d$

(C) the function $\theta \mapsto \pi(\theta)\det(DF(\theta)DF(\theta)^\top)^{-1/2}$ is integrable over $\mathbb{R}^d$

(D) the manifold Hessian $\nabla^2_{\mathcal{M}} R(\theta^\star)$ is non-singular

(E) $R(\theta) \leq M\|\theta\|^p$ for all $\theta \in \mathbb{R}^d$ for some $M, p > 0$

(F) the normalizing constant for ρ_γ is bounded as $\gamma \to 0^+$

(G) $R(\theta^\star)$ is uniformly bounded for any $n = 1, 2, \dots$

With these assumptions, we can establish the following theorem, our main result, the proof of which can be found in Appendix A.

Theorem 1 (PAC-Bayes Bound for Interpolators)**.** *Consider the cold posterior ρ_γ with prior $\pi(\theta) \propto \exp(-\frac{1}{\tau}R(\theta))$ under the choice of temperature $\tau = \frac{2}{mn}[R(\theta^\star) - R(\theta_0)]$. Then for any $0 < \delta < 1$, with probability at least $1-\delta$, as $\gamma \to 0^+$, and then $n \to \infty$,*

$$\begin{aligned}2\mathbb{E}_{\theta\sim\rho_\gamma} L(\theta) \leq{}& m\log(R(\theta^*) - R(\theta_0)) + \frac{1}{n}S + \frac{1}{n}K + P \\ &+ m\left(1 - \log\frac{mn}{2\pi}\right) + \frac{1}{n}\log\left(\delta^{-2}\right) + \mathcal{O}(n^{-2}) + \mathcal{O}(\gamma).\end{aligned} \tag{3}$$

In Theorem 1, the temperature τ is chosen so as to minimize the bound, excluding higher-order terms. The bound (3) has a similar interpretation to the IIC in [22] and so most of the discussion there is also relevant here. Indeed, in terms of the IIC in (2), the bound (3) becomes

$$2\mathbb{E}_{\theta\sim\rho_\gamma} L(\theta) \leq m\cdot(1 + \log(2\pi) + \mathrm{IIC}) + P + n^{-1}\log(\delta^{-2}) + \mathcal{O}(n^{-2}) + \mathcal{O}(\gamma).$$

4 Discussion and Conclusions

A PAC-Bayesian bound is presented in Theorem 1 for interpolators in the overparameterized regime, using the results of the IIC [22]. Our bound is quite general, imposing few restrictions on the model and the form of its implicit regularization. This is particularly advantageous in the setting of deep learning, where the precise nature of the model and the training process is often complex.

Drawing particular attention to the factor S, recall sharpness of the loss landscape is typically quantified in terms of the Hessian of the loss [50]. Multiple examinations have reported limitations to sharpness metrics computed involving the Hessian [4, 43]. One possibility for this deficiency is that the entire spectrum of the Hessian (and not only the top part) matters. The log-determinant depends not only on the largest eigenvalue, but on the decay rate of *all* the eigenvalues as well. However, for large neural networks, the Hessian is almost inevitably singular, and so its log-determinant is undefined [47, 52]. Our presented form of S has no such issues, and its relation to the Hessian is well studied [29, 43, 44]. This representation of sharpness should prove valuable in further explorations of the correlation between the eigenspectra and test performance as seen in heavy-tailed self-regularization theory [33–35] and other linearized analyses [2, 5, 42].

Finally, we remark that in view of the vast literature investigating implicit regularization of stochastic optimizers [9, 37, 45], the form of the regularizer R for neural network interpolators is a fertile ground for future research.

References

[1] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

[2] Kumar K Agrawal, Arnab Kumar Mondal, Arna Ghosh, and Blake Richards. α-req: Assessing representation quality in self-supervised learning by measuring eigenspectrum decay. *Advances in Neural Information Processing Systems*, 35:17626–17638, 2022.

[3] Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *arXiv preprint arXiv:2110.11216*, 2021.

[4] Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharpness and generalization. *Proceedings of the 40th International Conference on Machine Learning*, 202, 23–29 Jul 2023.

[5] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[6] Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. Pac-bayesian bounds based on the rényi divergence. In *Artificial Intelligence and Statistics*, pages 435–444. PMLR, 2016.

[7] Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

[8] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[9] Alberto Bietti, Grégoire Mialon, Dexiong Chen, and Julien Mairal. A kernel perspective for regularizing deep neural networks. In *International Conference on Machine Learning*, pages 664–674. PMLR, 2019.

[10] Lin Chen, Yifei Min, Mikhail Belkin, and Amin Karbasi. Multiple descent: Design your own generalization curve. *Advances in Neural Information Processing Systems*, 34:8898–8912, 2021.

[11] Valentin De Bortoli and Agnès Desolneux. On quantitative Laplace-type convergence results for some exponential probability measures, with two applications. *arXiv preprint arXiv:2110.12922*, 2021.

[12] M. Dereziński, F. Liang, and M. W. Mahoney. Exact expressions for double descent and implicit regularization via surrogate random design. Technical Report Preprint: arXiv:1912.04533, 2019.

[13] Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017*. AUAI Press, 2017.

[14] Spencer Frei, Gal Vardi, Peter Bartlett, and Nathan Srebro. Benign overfitting in linear classifiers and leaky relu networks from KKT conditions for margin maximization. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3173–3228. PMLR, 2023.

[15] W Fulks. A generalization of Laplace's method. *Proceedings of the American Mathematical Society*, 2(4):613–622, 1951.

[16] Pascal Germain, Francis Bach, Alexandre Lacoste, and Simon Lacoste-Julien. PAC-Bayesian theory meets Bayesian inference. *Advances in Neural Information Processing Systems*, 29, 2016.

[17] Boris Hanin and Alexander Zlokapa. Bayesian interpolation with deep linear networks. *Proceedings of the National Academy of Sciences*, 120(23):e2301345120, 2023.

[18] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

[19] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[20] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

[21] Liam Hodgkinson, Chris Van Der Heide, Fred Roosta, and Michael W. Mahoney. Monotonicity and double descent in uncertainty estimation with Gaussian processes. In *Proceedings of the 40th International Conference on Machine Learning*, pages 13085–13117, 2023.

[22] Liam Hodgkinson, Chris van der Heide, Robert Salomone, Fred Roosta, and Michael W Mahoney. The interpolating information criterion for overparameterized models. *arXiv preprint arXiv:2307.07785v1*, 2023.

[23] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[24] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

[25] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

[26] Sadanori Konishi and Genshiro Kitagawa. *Information criteria and statistical modeling*. Springer, 2008.

[27] Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

[28] Z. Liao, R. Couillet, and M. W. Mahoney. A random matrix analysis of random Fourier features: beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent. Technical Report Preprint: arXiv:2006.05013, 2020.

[29] Zhenyu Liao and Michael W Mahoney. Hessian eigenspectra of more realistic nonlinear models. *Advances in Neural Information Processing Systems*, 34:20104–20117, 2021.

[30] Fanghui Liu, Zhenyu Liao, and Johan Suykens. Kernel regression in high dimensions: Refined analysis beyond double descent. In *International Conference on Artificial Intelligence and Statistics*, pages 649–657. PMLR, 2021.

[31] Sanae Lotfi, Marc Finzi, Sanyam Kapoor, Andres Potapczynski, Micah Goldblum, and Andrew G Wilson. PAC-Bayes compression bounds so tight that they can explain generalization. *Advances in Neural Information Processing Systems*, 35:31459–31473, 2022.

[32] Sanae Lotfi, Pavel Izmailov, Gregory Benton, Micah Goldblum, and Andrew Gordon Wilson. Bayesian model selection, the marginal likelihood, and generalization. In *Proceedings of the 39th International Conference on Machine Learning*, pages 14223–14247, 2022.

[33] Charles H Martin and Michael W Mahoney. Traditional and heavy tailed self regularization in neural network models. In *International Conference on Machine Learning*, pages 4284–4293. PMLR, 2019.

[34] Charles H Martin and Michael W Mahoney. Heavy-tailed universality predicts trends in test accuracies for very large pre-trained deep neural networks. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 505–513. SIAM, 2020.

[35] Charles H Martin, Tongsu Peng, and Michael W Mahoney. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications*, 12(1):4122, 2021.

[36] David A McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

[37] Behnam Neyshabur. Implicit regularization in deep learning. *Ph.D. Thesis, Toyota Technological Institute at Chicago*, 2017.

[38] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

[39] Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017.

[40] Roman Novak, Jascha Sohl-Dickstein, and Samuel S Schoenholz. Fast finite width neural tangent kernel. In *International Conference on Machine Learning*, pages 17018–17044. PMLR, 2022.

[41] Konstantinos Pitas and Julyan Arbel. Cold posteriors through PAC-Bayes. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*, 2022.

[42] Tim G J Rudner, Sanyam Kapoor, Shikai Qui, and Andrew Gordon Wilson. Function-space regularization in neural networks: A probabilistic perspective. *Proceedings of the 40th International Conference on Machine Learning*, 202, 23–29 Jul 2023.

[43] Sidak Pal Sing, Thomas Hofmann, and Bernhard Schölkopf. The Hessian perspective into the Nature of Convolutional Neural Networks. *Proceedings of the 40th International Conference on Machine Learning*, 202, 23–29 Jul 2023.

[44] Sidak Pal Singh, Gregor Bachmann, and Thomas Hofmann. Analytic insights into structure and rank of neural network hessian maps. *Advances in Neural Information Processing Systems*, 34:23914–23927, 2021.

[45] Samuel L Smith, Benoit Dherin, David Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. In *International Conference on Learning Representations*, 2020.

[46] Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *J. Mach. Learn. Res.*, 24:123–1, 2023.

[47] Susan Wei, Daniel Murfet, Mingming Gong, Hui Li, Jesse Gell-Redman, and Thomas Quella. Deep learning is singular, and that's good. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[48] Florian Wenzel, Kevin Roth, Bastiaan Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the Bayes posterior in deep neural networks really? In *Proceedings of the 37th International Conference on Machine Learning*, pages 10248–10259, 2020.

[49] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W Mahoney. PyHessian: Neural Networks Through the Lens of the Hessian. In *2020 IEEE international conference on big data (Big data)*, pages 581–590. IEEE, 2020.

[50] Zhewei Yao, Amir Gholami, Qi Lei, Kurt Keutzer, and Michael W Mahoney. Hessian-based analysis of large batch training and robustness to adversaries. *Advances in Neural Information Processing Systems*, 31, 2018.

[51] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[52] Jiaxin Zhang. Modern Monte Carlo methods for efficient uncertainty quantification and propagation: a survey. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(5):e1539, 2021.

A Appendix

Proof of Theorem 1. Let $J(\theta) = DF(\theta)DF(\theta)^\top \in \mathbb{R}^{mn\times mn}$ denote the empirical NTK. We start from the core PAC-Bayes bound (1)

$$\mathbb{E}_\rho L(\theta) \leq \mathbb{E}_\rho L_n(\theta) + \frac{\log \mathbb{E}_\mathcal{D}\zeta_n}{n} + \frac{\mathrm{KL}(\rho\|\pi)}{n} + \frac{\log(1/\delta)}{n},$$

where dummy variables in the expectations have been dropped for brevity. First, since $L_n(\theta) = 0$ on $\mathcal{M}$, $\mathbb{E}_\rho L_n(\theta) = \mathcal{O}(\gamma)$. Next, taking $\pi(\theta) \propto \exp(-\frac{1}{\tau}R(\theta))$, applying Laplace's method twice,

$$\begin{aligned}\zeta_n = \int_\Theta e^{n(L(\theta)-L_n(\theta))}\pi(\theta)\mathrm{d}\theta &= \frac{\int_\Theta e^{n(L(\theta)-L_n(\theta))}e^{-\frac{1}{\tau}R(\theta)}\mathrm{d}\theta}{\int_\Theta e^{-\frac{1}{\tau}R(\theta)}\mathrm{d}\theta} \\ &= e^{n(L(\theta_0)-L_n(\theta_0))} + \mathcal{O}(\tau),\end{aligned}$$

and so $\log \mathbb{E}_\mathcal{D}\zeta_n = \frac{1}{2}nP + \mathcal{O}(\tau)$. Observe that

$$\begin{aligned}\mathrm{KL}(\rho\|\pi) &= \int_{\mathbb{R}^d} \log\left(\frac{\rho(\theta)}{\pi(\theta)}\right)\rho(\theta)\mathrm{d}\theta \\ &= \frac{1}{\mathcal{Z}_\gamma}\int_{\mathbb{R}^d} \log\left(\frac{\pi(\theta)e^{-\frac{n}{\gamma}L_n(\theta)}}{\pi(\theta)\mathcal{Z}_\gamma}\right)\pi(\theta)e^{-\frac{n}{\gamma}L_n(\theta)}\mathrm{d}\theta \\ &= \frac{1}{\mathcal{Z}_\gamma}\int_{\mathbb{R}^d}\left(-\log\mathcal{Z}_\gamma - \frac{n}{\gamma}L_n(\theta)\right)\pi(\theta)e^{-\frac{n}{\gamma}L_n(\theta)}\mathrm{d}\theta \\ &= -\log\mathcal{Z}_\gamma - \frac{n}{\gamma}\mathbb{E}_\rho L_n(\theta) \\ &\leq -\log\mathcal{Z}_\gamma.\end{aligned}$$

From the proof of [22, Theorem 1],

$$-\log\mathcal{Z}_\gamma = \frac{1}{\tau}[R(\theta^\star) - R(\theta_0)] + \frac{mn}{2}\log(\pi\tau) + \frac{S+K}{2} + \mathcal{O}(\tau).$$

In line with [22], choosing $\tau = \frac{2}{mn}[R(\theta^\star) - R(\theta_0)]$, since $R(\theta^\star) = \mathcal{O}(1)$, $\tau = \mathcal{O}(n^{-1})$ and

$$-\log\mathcal{Z}_\gamma = \frac{mn}{2}\left(1 + \log\frac{2\pi}{mn} + \log[R(\theta^\star) - R(\theta_0)]\right) + \frac{S+K}{2} + \mathcal{O}(n^{-1}).$$

Altogether,

$$\begin{aligned}\mathbb{E}_\rho L(\theta) &\leq \frac{P}{2} - \frac{\log\mathcal{Z}_\gamma}{n} + \frac{\log(1/\delta)}{n} + \mathcal{O}(\tau) + \mathcal{O}(\gamma) \\ &\leq \frac{P}{2} + \frac{m}{2}\left(1 + \log\frac{2\pi}{mn}\right) + \frac{m}{2}\log[R(\theta^\star) - R(\theta_0)] \\ &\qquad + \frac{S+K}{2n} + \frac{\log(1/\delta)}{n} + \mathcal{O}(n^{-2}) + \mathcal{O}(\gamma).\end{aligned}$$

□