Towards generalization guarantees for SGD: Data-dependent PAC-Bayes priors

Gintare Karolina Dziugaite
University of Cambridge; Element AI

Gabriel Arpino
University of Toronto

Daniel M. Roy
University of Toronto; Vector Institute

1 Introduction

Generalization bounds are one of a handful of tools we have for explaining the empirical performance of learning algorithms like stochastic gradient descent (SGD). However, at present, there are no generalization bounds for neural networks trained by SGD that are (known to be) nonvacuous (i.e., less than one) when applied to data sets and architectures used in practice. Despite the logical disconnect, unsupported claims connecting vacuous generalization bounds to empirical performance abound. While the mathematical development of (potentially vacuous in practice) generalization bounds is essential, premature attempts to explain practice may be slowing progress towards understanding.

Several recent papers describe nonvacuous bounds for networks learned by SGD *after compression and/or randomization*. For example, Dziugaite and Roy [3] obtain nonvacuous generalization bounds for *stochastic* neural networks on MNIST. These stochastic neural networks are defined by adding Gaussian noise to the weights of the SGD solution, where the mean and variance of the noise is optimized to minimize a PAC-Bayes generalization bound. The second example we are aware of is due to Zhou et al. [15]. They combine the techniques of Dziugaite and Roy [3] and existing state-of-the-art network compression schemes to produce nonvacuous bounds for stochastic versions of the compressed networks. The method scales to IMAGENET, although the bounds are very loose.

PAC-Bayes bound optimization is a generalization of stochastic variational inference: a prior distribution P on weights is fixed independently of the data and then a variational objective is optimized over a family of posterior distributions. The variational objective is a linear combination of the Q-average of the empirical risk and the KL divergence from Q to P.

What do these bounds imply about the generalization of the original network? Neyshabur et al. [9] describe an approach to transfer bounds like those proposed by Dziugaite and Roy [3] back to the original network. The approach is sensitive to the Lipschitz constant of the network learned by SGD, which they bound by spectral norms of the weight matrices. (Cf. [1].) Unfortunately, the approach produces empirically vacuous bounds.

This approach to controlling the risk of the SGD solution via a PAC-Bayes bound requires that the posterior distribution, Q, be tightly concentrated around the weights learned by SGD. Empirically, we have found that typical weights from optimized posterior distributions are quite different from those of the SGD solution, even if the SGD solution itself is assigned relatively high density. The prior is the culprit. Typical applications of PAC-Bayes bounds use generic priors. For example, Dziugaite and Roy [3] choose P to be a scale mixture of Gaussians, centered at the randomly initialized weights used subsequently by SGD during training. (This choice is already a considerable improvement over a prior centered at the origin, as the random weight initialization breaks symmetries that are often not reversed.) Since the prior distribution, P, is not itself concentrated around the SGD solution, the $\mathrm{KL}(Q||P)$ term in the PAC-Bayes bounds is vacuously large if Q is forced to be concentrated near the SGD solution. An in variational inference, bound optimization causes the posterior to shrink excessively towards the generic prior and away from the SGD solution.

Generic priors are not inherent to the PAC-Bayesian framework: every data-independent prior yields a valid generalization bound, and so in the pursuit of obtaining tight and informative generalization bounds, we *must* optimize the prior in order to remove a spurious shift of the distribution of the penalty term away from zero. Indeed, fixing a learning rule $S \mapsto Q(S)$ for the posterior, Langford [7] and Catoni [2] showed that the KL term is minimized in expectation by choosing $P = \mathbb{E}[Q(S)]$, where the expectation is over the unknown distribution of the data, S. In this case, the expected value of the KL term is precisely the mutual information between S and random weights w satisfying $\mathbb{P}[w|S] = Q(S)$. While this mutual information is unknown, we see that every valid choice of a prior P yields a KL term that is an upper bound in expectation. From this perspective, PAC-Bayes bounds provide a tractable way to approach mutual-information bounds via clever priors.

In this paper, we extend a line of work on distribution-dependent priors initiated by Catoni [2]. (See also [4, 5, 6, 10, 11, 12, 13].) Technically, we use an approach first presented by [12], however, in the setting of deep learning, we must contend with the nonconvex learning problem. Our bounds rely on SGD exhibiting a coarse type of stability: namely, the weights obtained from training on a subset of the data are highly predictive of the weights obtained from the whole data set. We use this property to devise data-dependent priors and then verify empirically that the resulting PAC-Bayes bounds are much tighter.

2 Preliminaries

Consider neural network classifiers parameterized by weights $w \in \mathbb{R}^d$. Let $\mathcal{D}$ denote the unknown data distribution on the space Z of labeled examples. Let $S = (z_1, \dots, z_m)$ denote m i.i.d. samples from $\mathcal{D}$. The goal of supervised learning is to identify weights w that minimize the risk $R_{\mathcal{D}}(w) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(w, z)]$, where $\ell : \mathbb{R}^d \times Z \to [0, 1]$. Since $\mathcal{D}$ is presumed unknown, the classifier is chosen based on its empirical risk $\hat{R}_S(w) = \frac{1}{m}\sum_{i=1}^m \ell(w, z)$.

Our goal is to study stochastic gradient descent, which can be viewed abstractly as a map $S \mapsto A(S) : Z^m \to \mathbb{R}^d$. We are interested in its generalization error: $R_{\mathcal{D}}(A(S)) - \hat{R}_S(A(S))$. Our approach will be to use PAC-Bayes theorems [8, 14].

Theorem 2.1 (Linear PAC-Bayes Bound; [2, 7])**.** *Fix $\tau > 1/2$. For every $\delta > 0$, $m \in \mathbb{N}$, $\mathcal{D} \in \mathcal{M}_1(Z)$, and $P \in \mathcal{M}_1(\mathbb{R}^D)$ with probability at least $1 - \delta$, for every distribution $Q \in \mathcal{M}_1(\mathbb{R}^D)$,*

$$R_{\mathcal{D}}(Q) \le \frac{1}{1 - \frac{1}{2\tau}}\Big(\hat{R}_S(Q) + \frac{\tau}{m}\big(\mathrm{KL}(Q||P) + \log\frac{1}{\delta}\big)\Big). \tag{1}$$

3 Data-dependent priors

Poor choice of a prior in the PAC-Bayes setting results in a loose generalization bound. The basic idea introduced by Parrado-Hernández et al. [12] is to use a fraction k of the training data S to choose a prior. However, predicting the learned weights for the full data set is nontrivial because the empirical risk is nonconvex for neural networks, and SGD is nondeterministic.

Our approach is to use the dynamics of SGD itself to predict the weights of the final network, and to couple SGD running on a fraction k of the training data with SGD running on the full data in order to optimize the prior. While there are more sophisticated versions of this idea, we use the following simple variant: For $k \in (0, 1]$, let $w_p[k]$, for $p \in \mathbb{R}_+$, denote the weights learned by SGD running on the first k fraction of the full data set after p/k epochs. (The index p is real-valued and so $w_p[k]$ is piecewise constant in p, with jumps after each minibatch. Similarly k is a real-valued and so $w_p[k]$ is piecewise constant in k, with jumps as additional data points are introduced.) By construction, $(w_p[1])_{p\in[0,k]}$ and $(w_p[k])_{p\in[0,k]}$ have the same distribution for each $k \in (0, 1]$. Thus, there exists a coupling of $w_p[k]$, for $k \in (0, 1]$ and $p \in \mathbb{R}_+$ such that, for every $k \in (0, 1]$, $w_p[1] = w_p[k]$ for $p \in [0, k]$. Let w_0 denote the weights upon initialization, and, for $p > 0$, let $w_p = w_p[1]$. Then $w_k = w_k[k]$ represents the weights after a fraction k through the first epoch, which are necessarily independent of the remaining $(1 - k)$ fraction of the data. For any fixed k, we may use w_k as a prior mean, and then measure the generalization error of w_T (or a Gibbs classifier surrogate Q) on the basis of $(1 - k)\, m$ data points. We can make a data-dependent choice of k via a union bound. (Note that this coupling is trivial to generate, as it corresponds to a single run (w_p) of SGD followed

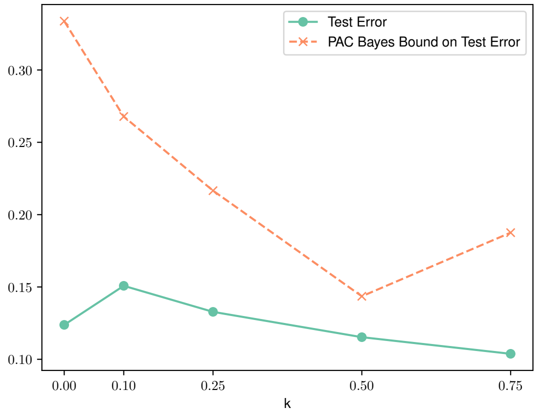

Figure 1: Test performance and PAC-Bayes bounds as a function of the fraction k of data used to learn a prior over weights. The optimal performance at 0.5 is a test error bound of approximately 15%. In contrast, Zhou et al. [15] obtain a bound greater that 40%.

by simulations of additional epochs for k fractions of the data starting from w_k, and proceeding independently of w_p for $p > k$.)

The following generalization bound is a simple extension of Theorem 2.1 to allow for data-dependence [12]. For a fraction k and dataset S, let S_k denote the first k fraction of S.

Theorem 3.1. *Fix $\lambda > 1/2$ and a finite subset $\mathcal{K} \subset [0,1]$. For all $\delta > 0$, $m \in \mathbb{N}$, $\mathcal{D} \in \mathcal{M}_1(Z)$, with probability at least $1-\delta$ over $S \sim \mathcal{D}^m$, for all $Q \in \mathcal{M}_1(\mathbb{R}^d)$, taking $P_k = \int \mathcal{N}(w_k, \lambda I)\pi(\mathrm{d}\lambda)$,*

$$R_{\mathcal{D}}(Q) \leq \frac{1}{1-\frac{1}{2\tau}} \inf_{k \in \mathcal{K}} \left(\hat{R}_{S \setminus S_k}(Q) + \frac{\tau}{(1-k)m} \mathrm{KL}(Q||P_k) + \log \frac{|\mathcal{K}|}{\delta} \right). \tag{2}$$

Other choices of data-dependent priors include choosing P_k to be centered at an estimate of the mean of $w_T[k]$ and using estimates of the covariance of $w_T[k]$ given w_k in place of λI.

4 Experimental Results

We computed our bounds on fully connected networks learned by SGD by minimizing cross-entropy loss on the full 10-class MNIST. We ran SGD until training error was 0.02. We chose $\mathcal{K} = \{0.0, 0.1, 0.25, 0.5, 0.75\}$. We used the PAC-Bayes bound optimization described in [3]. Both priors and posteriors were Gaussians, leading to tractable KL divergence formula. Thus the KL divergence with the scale mixture of Gaussian prior was upper bounded by a data-dependent choice of the correct scale from a discrete set, and a penalty coming from a union bound over that set. The bounds are presented in Fig. 1. Our bound is three times tighter than the bound produced by Zhou et al. [15] using neural network compression schemes. We do not see a significant reduction in the distance from the SGD solution to the posterior mean.

References

[1] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky. "Spectrally-normalized margin bounds for neural networks". In: *Advances in Neural Information Processing Systems*. 2017, pp. 6241–6250.

[2] O. Catoni. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*. 2007. arXiv: 0712.0248.

[3] G. K. Dziugaite and D. M. Roy. "Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data". In: *Proceedings of the 33rd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. 2017. arXiv: 1703.11008.

[4] G. K. Dziugaite and D. M. Roy. “Data-dependent PAC-Bayes priors via differential privacy”. In: *Advances in Neural Information Processing Systems (NIPS)*. Vol. 29. Cambridge, MA: MIT Press, 2018. arXiv: 1802.09583.

[5] G. Lever, F. Laviolette, and J. Shawe-Taylor. “Distribution-dependent PAC-Bayes priors”. In: *International Conference on Algorithmic Learning Theory*. Springer. 2010, pp. 119–133.

[6] G. Lever, F. Laviolette, and J. Shawe-Taylor. “Tighter PAC-Bayes bounds through distribution-dependent priors”. *Theoretical Computer Science* 473 (2013), pp. 4–28. ISSN: 0304-3975.

[7] D. A. McAllester. “A PAC-Bayesian Tutorial with A Dropout Bound”. *CoRR* abs/1307.2118 (2013).

[8] D. A. McAllester. “Some PAC-Bayesian Theorems”. *Machine Learning* 37.3 (Dec. 1999), pp. 355–363. ISSN: 1573-0565.

[9] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro. *A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks*. 2017. arXiv: 1707.09564.

[10] L. Oneto, D. Anguita, and S. Ridella. “PAC-bayesian analysis of distribution dependent priors: Tighter risk bounds and stability analysis”. *Pattern Recognition Letters* 80 (2016), pp. 200–207. ISSN: 0167-8655.

[11] L. Oneto, S. Ridella, and D. Anguita. “Differential privacy and generalization: Sharper bounds with applications”. *Pattern Recognition Letters* 89 (2017), pp. 31–38. ISSN: 0167-8655.

[12] E. Parrado-Hernández, A. Ambroladze, J. Shawe-Taylor, and S. Sun. “PAC-Bayes bounds with data dependent priors”. *Journal of Machine Learning Research* 13.Dec (2012), pp. 3507–3531.

[13] O. Rivasplata, E. Parrado-Hernandez, J. Shaws-Taylor, S. Sun, and C. Szepesvari. *PAC-Bayes bounds for stable algorithms with instance-dependent priors*. 2018. arXiv: 1806.06827.

[14] J. Shawe-Taylor and R. C. Williamson. “A PAC analysis of a Bayesian estimator”. In: *Proceedings of the tenth annual conference on Computational learning theory*. ACM. 1997, pp. 2–9.

[15] W. Zhou, V. Veitch, M. Austern, R. P. Adams, and P. Orbanz. “Compressibility and Generalization in Large-Scale Deep Learning” (2018). arXiv: 1804.05862.