On the role of data in PAC-Bayes bounds

Gintare Karolina Dziugaite[1,2] **Kyle Hsu**[3,4] **Waseem Gharbieh**[1] **Gabriel Arpino**[3,4] **Daniel M. Roy**[3,4]
[1]Element AI ServiceNow [2]Mila [3]U. of Toronto [4]Vector Institute

Abstract

The dominant term in PAC-Bayes bounds is often the Kullback–Leibler divergence between the posterior and prior. For so-called linear PAC-Bayes risk bounds based on the empirical risk of a fixed posterior kernel, it is possible to minimize the expected value of the bound by choosing the prior to be the expected posterior, which we call the *oracle* prior on the account that it is distribution dependent. In this work, we show that the bound based on the oracle prior can be suboptimal: In some cases, a stronger bound is obtained by using a data-dependent oracle prior, i.e., a conditional expectation of the posterior, given a subset of the training data that is then excluded from the empirical risk term. While using data to learn a prior is a known heuristic, its essential role in optimal bounds is new. In fact, we show that using data can mean the difference between vacuous and nonvacuous bounds. We apply this new principle in the setting of nonconvex learning, simulating data-dependent oracle priors on MNIST and Fashion MNIST with and without held-out data, and demonstrating new nonvacuous bounds in both cases.

1 INTRODUCTION

In this work, we are interested in the application of PAC-Bayes bounds (McAllester, 1999b; Shawe-Taylor and Williamson, 1997) to the problem of understanding the generalization properties of learning algorithms. Our focus will be on supervised learning from i.i.d. data, although PAC-Bayes theory has been generalized far beyond this setting, as summarized in a recent survey by Guedj (2019). In our setting, PAC-Bayes bounds control the risk of Gibbs classifiers, i.e., randomized classifiers whose predictions, on each input, are determined by a classifier h sampled according to some distribution Q on the hypothesis space $\mathcal{H}$. The hallmark of a PAC-Bayes bound is a normalized Kullback–Leibler (KL) divergence, $m^{-1}\mathrm{KL}(Q||P)$, defined in terms of a Gibbs classifier P that is called a "prior" because it must be independent of the m data points used to estimate the empirical risk of Q.

In applications of PAC-Bayes bounds to generalization error, the contribution of the KL divergence often dominates the bound: In order to have a small KL with a strongly data-dependent posterior, the prior must, in essence, predict the posterior. This is difficult without knowledge of (or access to) the data distribution, and represents a significant statistical barrier to achieving tight bounds. Instead, many PAC-Bayesian analyses rely on generic priors chosen for analytical convenience.

Generic priors, however, are not inherent to the PAC-Bayes framework: every valid prior yields a valid bound. Therefore, if one does not optimize the prior to the data distribution, one may obtain a bound that is loose on the account of ignoring important, favorable properties of the data distribution.

Langford and Blum (2003) were the first to consider the problem of optimizing the prior to minimize the *expected value* of the high-probability PAC-Bayes bound. In the realizable case, they show that the problem reduces to optimizing the expected value of the KL term. More precisely, they consider a fixed learning rule $S \mapsto Q(S)$, i.e., a fixed posterior kernel, which chooses a posterior, $Q(S)$, based on a training sample, S. In the realizable case, the bound depends linearly on the KL term. Then $\mathbb{E}[\mathrm{KL}(Q(S)||P)]$ is minimized by the expected posterior, $P^* = \mathbb{E}[Q(S)]$, i.e., $P^*(B) = \mathbb{E}[Q(S)(B)]$ for measurable $B \subseteq \mathcal{H}$. Both expectations are taken over the unknown distribution of the training sample, S. We call P^* the *oracle* prior. If we introduce an $\mathcal{H}$-valued random variable H satisfying $\mathbb{P}[H|S] = Q(S)$ a.s., we see that its distribution, $\mathbb{P}[H]$, is P^* and thus, the "optimality" of the oracle P^* is an immediate consequence of the identity

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

$I(S;H) = \mathbb{E}[\mathrm{KL}(Q(S)||P^*)] = \inf_{P'} \mathbb{E}[\mathrm{KL}(Q(S)||P')]$, a well-known variational characterization of mutual information in terms of KL divergence.

For so-called linear PAC-Bayes bounds (introduced below), the oracle prior is seen to minimize the bound in expectation when all the data are used to estimate the risk. This holds even in the unrealizable setting. Thus, having settled on a learning rule $S \mapsto Q(S)$, we might seek to achieve the tightest linear PAC-Bayes bound in expectation by attempting to approximate the oracle prior, P^*. Indeed, there is a large literature aimed at obtaining localized PAC-Bayes bounds via distribution-dependent priors, whether analytically (Catoni, 2003; Catoni, 2007; Lever, Laviolette, and Shawe-Taylor, 2010; Lever, Laviolette, and Shawe-Taylor, 2013), through data (Ambroladze, Parrado-Hernández, and Shawe-Taylor, 2007; Negrea et al., 2019), or by way of concentration of measure, privacy, or stability (Dziugaite and Roy, 2018; Oneto, Anguita, and Ridella, 2016; Oneto, Ridella, and Anguita, 2017; Rivasplata, Parrado-Hernandez, et al., 2018).

One of the contributions of this paper is the demonstration that an oracle prior may not yield the tightest linear PAC-Bayes risk bound in expectation, *if we allow ourselves to consider also using only subsets of the data to estimate the risk*. Proposition 3.1 gives conditions on a learning rule for there to exist data-dependent priors that improves the bound based upon the oracle prior. This phenomenon is a hitherto unstated principle of PAC-Bayesian analysis: data-dependent priors are sometimes necessary for tight bounds. Note that, as the prior must be independent of data used to compute the bound *a posteriori*, if m training data are used to define the prior, only the remaining $n - m$ data should be used to compute the bound (i.e., compute the empirical risk term and divide the KL term). *Note that all n training data are used by the learning algorithm.* We formalize these subtleties in the body of the paper and discuss some other misconceptions in Appendix J.

We give an example of a learning problem where Proposition 3.1 implies data-dependent priors dominate. The example is adapted from a simple model of SGD in a linear model by Nagarajan and Kolter (2019b). In the example, most input dimensions are noise with no signal and this noise accumulates in the learned weights. In our version, we introduce a learning rate schedule, and so earlier data points have a larger influence on the resulting weights. Even so, there is enough variability in the posterior that the oracle prior yields a vacuous bound. By conditioning on early data points, we reduce the variability and obtain nonvacuous bounds.

The idea of using data-dependent priors to obtain tighter bounds is not new (Ambroladze, Parrado-Hernández, and Shawe-Taylor, 2007; Dziugaite and Roy, 2018; Parrado-Hernández et al., 2012; Rivasplata, Parrado-Hernandez, et al., 2018). The idea is also implicit in the luckiness framework (Shawe-Taylor, Bartlett, et al., 1996). However, the observation that using data can be essential to obtaining a tight bound, even in full knowledge of the true distribution, is new, and brings a new dimension to the problem of constructing data-dependent priors.

In addition to demonstrating the theoretical role of data-dependent priors, we investigate them empirically, by studying generalization in nonconvex learning by stochastic (sub)gradient methods. As data-dependent oracle priors depend on the unknown distribution, we propose to use held-out data ("ghost sample") to estimate unknown quantities. Unlike standard held-out test set bounds, this approach relies implicitly on a type of stability demonstrated by SGD. We also propose approximations to data-dependent oracle priors that use no ghost sample, and find, given enough data, the advantage of the ghost sample diminishes significantly. We show that both approaches yield state-of-the-art nonvacuous bounds on MNIST and Fashion-MNIST for posterior Gaussian distributions whose means are clamped to the weights learned by SGD. Our MNIST bound (11%) improves significantly on the best published bound (46%) (Zhou et al., 2019). Finally, we evaluate minimizing a PAC-Bayes bound with our data-dependent priors as a learning algorithm. We demonstrate significant improvements to both classifier accuracy and bound tightness, compared to optimizing with generic priors.

2 PRELIMINARIES

Let Z be a space of labeled examples, and write $\mathcal{M}_1(Z)$ for the space of (probability) distributions on Z. Given a space $\mathcal{H}$ of *classifiers* (e.g., neural network predictors defined by their weights w) and a bounded *loss function* $\ell : \mathcal{H} \times Z \to [0,1]$, the risk of a hypothesis $w \in \mathcal{H}$ is $L_{\mathcal{D}}(w) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(w,z)]$. We also consider *Gibbs classifiers*, i.e., elements P in the space $\mathcal{M}_1(\mathcal{H})$ of distributions on $\mathcal{H}$, where risk is defined by $L_{\mathcal{D}}(P) = \mathbb{E}_{w\sim P} L_{\mathcal{D}}(w)$. As $\mathcal{D}$ is unknown, learning algorithms often work by optimizing an objective that depends on i.i.d. training data $S \sim \mathcal{D}^n$, such as the *empirical risk* $L_S(w) = L_{\hat{\mathcal{D}}_n}(w) = \frac{1}{n}\sum_{i=1}^n \ell(w, z_i)$, where $\hat{\mathcal{D}}_n$ is the empirical distribution of S. Writing $Q(S)$ for a data-dependent Gibbs classifier (i.e., a *posterior*), our primary focus is its risk, $L_{\mathcal{D}}(Q(S))$, and its relationship to empirical estimates, such as $L_S(Q(S))$.

The PAC-Bayes framework (McAllester, 1999b; Shawe-Taylor and Williamson, 1997) provides generalization bounds on data-dependent Gibbs classifiers. Let $Q, P \in$

$\mathcal{M}_1(\mathcal{H})$ be probability measures defined on a common measurable space $\mathcal{H}$. When Q is absolutely continuous with respect to P, written $Q \ll P$, we write $\frac{\mathrm{d}Q}{\mathrm{d}P} : \mathcal{H} \to \mathbb{R}_+ \cup \{\infty\}$ for some Radon–Nikodym derivative (aka, density) of Q with respect to P. The Kullback–Liebler (KL) divergence from Q to P is $\mathrm{KL}(Q||P) = \int \ln \frac{\mathrm{d}Q}{\mathrm{d}P} \mathrm{d}Q$ if $Q \ll P$ and ∞ otherwise. Assuming Q and P admit densities q and p, respectively, w.r.t. some sigma-finite measure $\nu \in \mathcal{M}(\mathcal{H})$, the definition of the KL divergence satisfies

$$\mathrm{KL}(Q||P) = \int \log \frac{q(w)}{p(w)} q(w) \nu(\mathrm{d}w).$$

The following PAC-Bayes bound follows from (McAllester, 2013, Thm. 2), taking $\beta = 1 - 1/(2\lambda)$. (See also Catoni (2007, Thm. 1.2.6).)

Theorem 2.1 (Linear PAC-Bayes bound)**.** *Let $\beta, \delta \in (0,1)$, $n \in \mathbb{N}$, $\mathcal{D} \in \mathcal{M}_1(Z)$, and $P \in \mathcal{M}_1(\mathcal{H})$. With probability at least $1-\delta$ over $S \sim \mathcal{D}^n$, for all $Q \in \mathcal{M}_1(\mathcal{H})$,*

$$L_{\mathcal{D}}(Q) \le \Psi_{\beta,\delta}(Q, P; S) \stackrel{def}{=} \frac{1}{\beta} L_S(Q) + \frac{\mathrm{KL}(Q||P) + \log \frac{1}{\delta}}{2\beta(1-\beta)|S|}.$$

As is standard, we call P the *prior*.

Note that the KL term in the bound depends on the data S through the kernel $Q(S)$. If we are interested in obtaining the tightest possible bound for the kernel $Q(S)$, then we can seek to minimize the KL term in some distribution sense. Our control of the KL term comes from the prior P. Since the bound is valid for all priors independent from S, we can choose P by optimizing, e.g., the risk bound in expectation, as first proposed by Langford and Blum (2003):

Theorem 2.2. *Let $n \in \mathbb{N}$ and fix a probability kernel $Q : Z^n \to \mathcal{M}_1(\mathcal{H})$. For all $\beta, \delta \in (0,1)$ and $\mathcal{D} \in \mathcal{M}_1(Z)$, $\mathbb{E}_{S\sim\mathcal{D}^n} \Psi_{\beta,\delta}(Q(S), P; S)$ is minimized, in P, by the "oracle" prior $P^* = \mathbb{E}_{S\sim\mathcal{D}^n}[Q(S)]$.*

Note that, in other PAC-Bayes bounds, the KL term sometimes appears within a concave function. In this case, oracle priors can be viewed as minimizing an upper bound on bound. We focus on linear PAC-Bayes bounds here for analytical tractability.

3 DATA-DEPENDENT ORACLE PRIORS

Here we demonstrate that, for linear PAC-Bayes bounds, one may obtain a stronger bound using a "data-dependent oracle" prior, rather than the usual (data-independent) oracle prior. Further, using a data-dependent oracle prior may mean the difference between a vacuous and nonvacuous bound.

A typical PAC-Bayes generalization bound for a posterior kernel $S \mapsto Q(S)$ is based on the empirical risk $L_S(Q(S))$ computed from the same data fed to the kernel. Instead, let J be a (possibly random) subset of $[n]$ of size $m < n$, independent from S, let S_J denote the subsequence of data with indices in J, and let $S \setminus S_J$ denote the complementary subsequence. Consider now the PAC-Bayes bound based on the estimate $L_{S\setminus S_J}(Q(S))$. In this case, the prior need only be independent from $S \setminus S_J$. The *$\sigma(S_J)$-measurable data-dependent oracle prior* $P^*(S_J) = \mathbb{E}[Q(S)|S_J]$ arises as the solution of the optimization

$$\inf_{P \in Z^{|J|} \to \mathcal{M}_1(\mathcal{H})} \mathbb{E}[\mathrm{KL}(Q(S)||P(S_J))]. \tag{1}$$

Letting $\hat{w}$ be a random element in $\mathcal{H}$ satisfying $\mathbb{P}[\hat{w}|S, J] = Q(S)$ a.s., the value of Eq. (1) is the *conditional* mutual information $I(\hat{w}; S|S_J)$. This conditional mutual information represents the expected value of the KL term in the linear PAC-Bayes bound and so this data-dependent prior achieves, in expectation, the tightest linear PAC-Bayes bound based on the estimate $L_{S\setminus S_J}(Q(S))$.

We can also consider restricting the prior distribution to a family $\mathcal{F} \subseteq \mathcal{M}_1(\mathcal{H})$ of distributions, in which case the optimization in Eq. (1) is over the set of kernels $Z^{|J|} \to \mathcal{F}$. We refer to a solution of this optimization as a data-dependent oracle prior *in* $\mathcal{F}$, denoted $P^*_{\mathcal{F}}(S_J)$, and refer to the value of Eq. (1) as the conditional $\mathcal{F}$-mutual information, denoted $I_{\mathcal{F}}(\hat{w}; S|S_J)$. The unconditional $\mathcal{F}$-mutual information is defined equivalently.[1] In Section 4, we study data-dependent oracle priors in a restricted family $\mathcal{F}$ in a setting where dealing with the set of all priors is intractable.

Fix $\mathcal{F}$ and define the *information rate gain* (from using S_J to choose the prior in $\mathcal{F}$) and the *excess bias* (from using $S \setminus S_J$ to estimate the risk) to be, respectively,

$$\mathrm{R}_{\mathcal{F}}(\hat{w}; S|S_J) = \frac{I_{\mathcal{F}}(\hat{h}; S)}{|S|} - \frac{I_{\mathcal{F}}(\hat{h}; S|S_J, J)}{|S \setminus S_J|} \tag{2}$$

and

$$\mathrm{B}(\hat{w}; S|S_J) = \mathbb{E}[L_{S\setminus S_J}(\hat{w}) - L_S(\hat{w})]. \tag{3}$$

Note that, if J is chosen uniformly at random, then $\mathrm{B}(\hat{w}; S|S_J) = 0$. Using these two quantities, we can

[1] When $\mathcal{F}$ is the set of all distributions, we drop $\mathcal{F}$ from the notation. The notation $P^*(S_J)$ is understood to also specify the data S_J held out from the estimate of risk. Thus, $P^*_{\mathcal{F}}$ denotes the distribution-dependent but data-independent oracle prior when the choice of prior is restricted to $\mathcal{F}$, just as P^* represents the distribution-dependent but data-independent oracle prior when the choice of prior is unrestricted.

characterize whether a data-dependent prior can outperform the oracle prior. The following result is an immediate consequence of the above definitions. (We present the straightforward proof in Appendix A for completeness.)

Proposition 3.1. *Let $\beta, \delta \in (0,1)$, $n \in \mathbb{N}$, and $\mathcal{D} \in \mathcal{M}_1(Z)$. Fix $Q : Z^n \to \mathcal{M}_1(\mathcal{H})$ and let $J \subseteq [n]$ be a (possibly random) subset of nonrandom cardinality $m < n$, independent from $S \sim \mathcal{D}^n$. Conditional on S and J, let $\hat{w}$ have distribution $Q(S)$. Then $\mathbb{E}_J \mathbb{E}_{S\sim\mathcal{D}^n} \Psi_{\beta,\delta}(Q(S), P^*_{\mathcal{F}}(S_J); S \setminus S_J) \leq \mathbb{E}_{S\sim\mathcal{D}^n} \Psi_{\beta,\delta}(Q(S), P^*_{\mathcal{F}}; S)$ if and only if*

$$\mathrm{R}_{\mathcal{F}}(\hat{w}; S|S_J) \geq 2(1-\beta)\,\mathrm{B}(\hat{w}; S|S_J) + \frac{\log\frac{1}{\delta}}{n}\frac{m}{n-m}, \quad (4)$$

i.e., Eq. (4) holds if and only if the linear PAC-Bayes bound with a oracle (data-independent) prior is no tighter, in expectation, than that with the data-dependent oracle prior.

To interpret the proposition, consider $\beta = 1/2$: then a data-dependent prior yields a tighter bound, if the information rate gain is larger than the excess bias and a term that accounts for excess variance.

Do such situations arise naturally? In fact, they do. The following demonstration uses a linear classification problem presented by Nagarajan and Kolter (2019b). Their example was originally constructed to demonstrate potential roadblocks to studying generalization in SGD using uniform convergence arguments. We make one, but important modification: we modify the learning algorithm to have another feature of SGD in practice: a *decreasing* step size. As is the case in ordinary training, the decreasing step size causes earlier data points to have more influence. As the data are noisy, the noise coming from these early samples has an outsized effect that renders a linear PAC-Bayes bound vacuous. By leaving the initial data out of the estimate of risk, and using a data-dependent oracle prior, we achieve a tighter bound. Indeed, we obtain a nonvacuous bound, while the optimal data-independent oracle prior yields a *vacuous* bound.

Example 3.2. Consider the hypothesis class $\mathcal{H} = \mathbb{R}^d$, interpreted as linear classifiers

$$\boldsymbol{x} \mapsto \mathrm{sign}(\langle \boldsymbol{x}, \boldsymbol{w} \rangle) : \mathbb{R}^d \to \{-1, 0, 1\}, \quad \text{for } \boldsymbol{w} \in \mathbb{R}^d. \quad (5)$$

Assume that $d = K + D$, with $D \gg K$, and decompose each input $\boldsymbol{x} \in \mathbb{R}^d$ as $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$, where $\boldsymbol{x}_1 \in \mathbb{R}^K$ and $\boldsymbol{x}_2 \in \mathbb{R}^D$. (We will decompose the weights similarly.) Labels y take values in $\{\pm 1\}$ and so a prediction of 0 (i.e., on the decision boundary) is a mistake.

Consider the following n i.i.d. training data: Let $\boldsymbol{u} \in \mathbb{R}^k$ be a nonrandom vector and, for each $i = 1, \ldots, n$, choose y_i uniformly at random in $\{\pm 1\}$, let $\boldsymbol{x}_{i,1} = y_i \boldsymbol{u}$, and let $\boldsymbol{x}_{i,2}$ be multivariate normal with mean 0 and covariance $(\sigma^2/D)\, I_D$, where I_D is the $D \times D$ identity matrix. Let $\mathcal{D}$ denote the common marginal distribution of each training example $(y_i, \boldsymbol{x}_i)$.

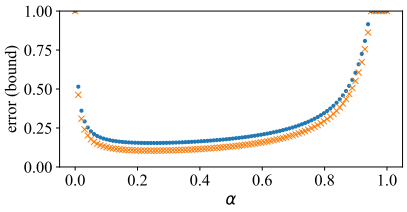

Figure 1: Lower (orange x's) and upper (blue dots) bounds on the expected value of a linear PAC-Bayes bound as a function of the fraction, α, of the 100 training data used by the data-dependent (PAC-Bayes) prior. Each bound uses the optimal (in expectation) tradeoff β and data-dependent prior $P(S_J)$, for $J = [k]$. Without using data (i.e., $J = \emptyset$), the bound is provably vacuous as the lower bound exceeds one. The upper bound is approximately 0.15 when the oracle prior is computed conditionally given the first 24 data points (i.e., $J = [24]$ and $\alpha = 0.24$.).

Consider the following one-pass learning algorithm: Let $\boldsymbol{w}_0 = 0$, then, for $t = 1, \ldots, n$ and $\eta_t = 1/t$, put $\boldsymbol{w}_t = \boldsymbol{w}_{t-1} + \eta_t y_t \boldsymbol{x}_t$. Then define the final weights to be $W = \boldsymbol{w}_n + (0, \xi)$, where ξ is an independent, zero-mean multivariate Gaussian with covariance $\kappa\, I_D$. Note that $\boldsymbol{w}_n = (\boldsymbol{w}_{n,1}, \boldsymbol{w}_{n,2})$ where $\boldsymbol{w}_{n,1} = (\sum_{i=1}^n \eta_i)\boldsymbol{u}$ and $\boldsymbol{w}_{n,2} = \sum_{i=1}^n \eta_i y_i \boldsymbol{x}_{i,2}$.

We will compare bounds based on oracle priors with those based on data-dependent oracle priors. To that end, let $S = \{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ and define Q by $\mathbb{P}[W|S] = Q(S)$ a.s. Let $[n] = \{1, \ldots, n\}$. For a subset $J \subseteq [n]$, let S_J be the corresponding subset of the data S and let $S \setminus S_J$ be the complement.

Lemma 3.3. *There are constants $n, D, \sigma, \kappa, \delta, u$ such that the infimum*

$$\inf_{J,\beta,P} \mathbb{E}\big[\Psi_{\beta,\delta}(Q(S), P(S_J); S \setminus S_J)\big], \quad (6)$$

where J ranges over subsets of $[n]$, β ranges over $(0,1)$, and P ranges over measurable functions $Z^{|J|} \to \mathcal{M}_1(\mathcal{H})$, is achieved by a nonempty set J. In particular, the optimal prior is data dependent.

Lower and upper bounds on the objective (Eq. (6)) for J of the form $\{1, \ldots, \lfloor 100\alpha \rfloor\}$, for $\alpha \in [0,1]$, are visualized in Fig. 1. Using a data-dependent prior in this scenario is critical for obtaining a nonvacuous bound. The derivation of these bounds as well as a sketch of the proof and a complete rigorous proof, can be found in Appendix B. ◁

In summary, data-dependent oracle priors, by definition, minimize linear PAC-Bayes bounds in expectation. The example above demonstrates that data-dependence can be essential in using linear PAC-Bayes bounds to obtain nonvacuous bounds. The example relies in a crucial way on the step size decreasing, so that some data points have an outsized impact on the noise that is injected into the classifier. In the remainder, we consider the problem of exploiting data dependent priors in the setting of learning with SGD.

4 DATA-DEPENDENT PRIORS FOR SGD

As the theoretical results in the previous section demonstrate, data-dependent oracle priors can lead to dramatically tighter bounds. In this section, we take the first steps towards understanding whether data-dependent priors can aid us in the study of deep learning with stochastic gradient descent (SGD).

Most attempts to build nonvacuous PAC-Bayes bounds for neural networks learned by SGD fail when the bounds are derandomized (Nagarajan and Kolter, 2019a; Neyshabur et al., 2018). In order to gain tight control on the derandomization, one requires that the posterior is concentrated tightly around the weights learned by SGD. This leads to a significant challenge as the prior must accurately predict the posterior, otherwise the KL term explodes. Can data-dependent priors allow us to use more concentrated priors? While we may not be able to achieve derandomized bounds yet, we should be able to build tighter bounds for stochastic neural networks with lower empirical risk.

In Example 3.2, we studied a posterior that depended more heavily on some data points than others. This property was introduced intentionally in order to serve as a toy model for SGD. Unlike the toy model, however, we know of no representations of the marginal distribution of the parameters learned by SGD that would allow us to optimize or compute a PAC-Bayes bound with respect to a data-dependent oracle prior. As a result, we are forced to make approximations.

Issues of tractability aside, another obstacle to using a data-dependent oracle prior is its dependence on the unknown data distribution. Ostensibly, this statistical barrier can be surmounted with extra data, although this would not make sense in a standard model-selection or self-bounded learning setup. In these more traditional learning scenarios, one has a training data set S and wants to exploit this data set to the maximum extent possible. Using some of this data to estimate or approximate (functionals of) the unknown distribution means that this data is not available to the learning algorithm or the PAC-Bayes bound. Indeed, if our goal is simply to obtain the tightest possible bound on the risk of our classifier, we ought to use most of this extra data to learn a better classifier, leaving out a small fraction to get a tight Hoeffding-style estimate of our risk.

However, if our goal is to understand the generalization properties of some posterior kernel Q (and indirectly an algorithm like SGD), we do not simply want a tight estimate of risk. *Indeed, a held-out test set bound is useless for understanding as it merely certifies that a learned classifier generalizes.* If a classifier generalizes due to favorable properties of the data distribution, then we must necessarily capture these properties in our bound. These properties may be natural side products of the learning algorithm (such as weight norms) or functionals of the unknown distribution that we must estimate (such as data-dependent oracle priors or functionals thereof). In this case, it makes sense to exploit held out data to gain insight.

4.1 Optimal isotropic Gaussian priors

In order to make progress, we begin by optimizing a prior over a restricted family $\mathcal{F}$. In particular, we consider the family of Gaussian priors when the posterior kernel chooses Gaussian posteriors. Based on empirical findings on the behavior of SGD in the literature, we propose an approximation to the data-dependent oracle prior.

Let $(\Omega, \mathcal{F}, \nu)$ be a probability space representing the distribution of a source of randomness. Our focus here is on kernels $Q : \Omega \times Z^n \to \mathcal{M}_1(\mathcal{H})$ where $Q(U, S) = \mathcal{N}(w_S, \Sigma)$ is a multivariate normal, centered at the weights $w_S \in \mathbb{R}^p$ learned by SGD (using randomness U, which we may assume without loss of generality encodes both the random initialization and the sequence of minibatches) on the full data set, S. Such posteriors underlie several recent approaches to obtaining PAC-Bayes bounds for SGD. In these bounds, the covariance matrix Σ is chosen to be diagonal and the scales are chosen to allow one to derive the bound on a deterministic classifier from the bound on a randomized classifier Q. For example, Neyshabur et al. (2018) derive deterministic classifier bounds from a PAC-Bayes bound based on (an estimate of) the Lipschitz constant of the network.

Fix some nonnegative integer $m \leq n$ and let $\alpha = m/n$. Let S_α denote the size m subset of S corresponding to the first m indices processed by SGD. (Note that these indices are encoded in U.) Writing $\mathbb{E}^{S_\alpha, U}[\cdot]$ for the conditional expectation operator given S_α, U, Theorem 2.2 implies that the tightest (linear PAC-Bayes) bound in expectation is obtained by minimizing $\mathbb{E}^{S_\alpha, U}[\mathrm{KL}(Q(U, S)||P)]$ in terms of P, which yields the

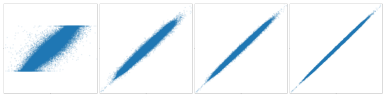

Figure 2: MNIST, FC; **x-axis**: parameter values of *base* run; **y-axis**: parameter values of α-*prefix* run; **left to right:** α values equal to $\{0, 0.1, 0.5, 0.9\}$. As α increases, the correlation between the parameters learnt by SGD on all of the data and an α fraction of the data increases.

data-dependent oracle prior $P = \mathbb{E}^{S_\alpha, U}[Q(U, S)]$. (We are permitted to condition on U because U is independent from S.)

As this prior is assumed to be intractable and the data distribution is unknown, we make a few approximations. First, as proposed in Example 3.2, we consider optimizing the prior over a family $\mathcal{F}$ of priors. Specifically, consider the identifying the isotropic Gaussian prior $P = \mathcal{N}(w_\alpha, \sigma_P I)$ that minimizes $\mathbb{E}^{S_\alpha, U}[\mathrm{KL}(Q(U, S)||P)]$. (We will revisit this simplification in Appendix I, where we consider priors and posteriors with non-isotropic diagonal covariance matrices. In short, we show that not much can be gained with diagonal priors.) If we fix σ_P, then based on the KL divergence between multivariate Gaussians (Eq. (33)), the optimization problem reduces to

$$\arg\min_{w_\alpha} \mathbb{E}^{S_\alpha, U}[\|w_S - w_\alpha\|^2]. \tag{7}$$

It follows that the mean of the *Gaussian* oracle prior (with fixed isotropic covariance) is the conditional expectation $\mathbb{E}^{S_\alpha, U}[w_S]$ of the weights learned by SGD. Under this choice, the contribution of the mean component to the bound is the value of the expectation in Eq. (7), which can be seen to be the trace of the conditional covariance of w_S given S_α, U. For the remainder of the section we will focus on the problem of approximating the oracle prior mean. The optimal choice of σ_P depends on the distribution of Σ. One approach, which assumes that we build separate bounds for different values of σ_P that we combine via a union bound argument, is outlined in Appendix C.

4.2 Ghost samples

In the setting above, the optimal Gaussian prior mean is given by the conditional expectation $\mathbb{E}^{S_\alpha, U}[w_S]$. Although the distribution $\mathcal{D}$ is presumed to be unknown, there is a natural statistical estimate for $\mathbb{E}^{S_\alpha, U}[w_S]$. Namely, consider a *ghost sample*, S^G, independent from and equal in distribution to S. Let S^G_α be the data set obtained by combining S_α with a $1-\alpha$ fraction of S^G. (We can do so by matching the position of S_α within S and within S^G_α.) Note that S^G_α is also equal in distribution to S. We may then take w^G_α to be the mean of $Q(U, S^G_\alpha)$, i.e., the weights produced by SGD on the data set S^G_α using the randomness U.

By design, SGD acting on S^G_α and randomness U will process S_α first and then start processing the data from the ghost sample. Crucially, the initial α fraction of the first epoch in both runs will be identical. By design, w^G_α and w_S are equal in distribution when conditioned on S_α and U, and so w^G_α is an unbiased estimator for $\mathbb{E}^{S_\alpha, U}[w_S]$.[2]

4.3 Terminology

We call the run of SGD on data S_α the α-*prefix* run. The run of SGD on the full data is called the *base* run. A prior is constructed from the α-*prefix* run by centering a Gaussian at the parameters obtained after T steps of optimization. Prefix stopping time T is chosen from a discrete set of values to minimize L^2 distance to posterior mean.[3] Note, that for $\alpha = 0$, $w_\alpha = w_0$, i.e., the prior is centered at random initialization as it has no access to data. This is equivalent to the approach taken by Dziugaite and Roy (2017). When the prior has access to data S^G_α, we call an SGD run training on S^G_α an α-*prefix+ghost* run, obtaining parameters w^G_α.

The procedure of running the α-*prefix* and *base* runs together for the first α-fraction of a *base* run epoch using shared information U (storing the data order) is an example of a *coupling*. This coupling is simple and does not attempt to match *base* and α-*prefix* runs beyond the first m/b iterations (where b is the batch size, which we presume divides m evenly for simplicity). It exploits the fact that the final weights have an outsized dependence on the first few iterations of SGD. More advanced coupling methods can be constructed. Such methods might attempt to couple beyond the first α–fraction of the first epoch.

As argued above, it is reasonable to use held-out data to probe the implications of a data-dependent prior as it may give us insight into the generalization properties of Q. At the same time, we may be interested in approximations to the data-dependent oracle that do not use a ghost sample. Ordinarily, we would expect two independent runs of SGD, even on the same dataset, to produce potentially quite different weights (measured, e.g., by their L^2 distance) (Nagarajan and Kolter, 2019b). Fig. 2 shows that, when we condition on an initial prefix of data, we dramatically decrease the variability of the

[2] We can minimize the variance of the KL term by producing conditionally i.i.d. copies of w^G_α and averaging, although each such copy requires an independent $n-m$-sized ghost sample.

[3] We account for these data-dependent choices via a union bound, which produces a negligible contribution.

Algorithm 1 PAC-Bayes bound computation (right) and optimization (left). **Given**:Data S, ghost data S^G (if α-*prefix+ghost*), batch size b. **Hyperparameters**:stopping criteria $\mathcal{E}$, prefix fraction α, prefix stopping time T, prior variance σ_P.

function BOUND-OPT $(\alpha, \sigma_P, T, \eta)$
- $S_\alpha \leftarrow \{z_1, .., z_{\alpha|S|}\} \subset S$ ▷ Select α-*prefix*
- $w_\alpha^0 \leftarrow \mathrm{SGD}(w_0, S_\alpha, b, \frac{|S_\alpha|}{b})$ ▷ Coupling
- $w_\alpha^S \leftarrow \mathrm{SGD}(w_\alpha^0, S, b, \infty, 0)$ ▷ α-*prefix*
- $P \leftarrow \mathcal{N}(w_\alpha^S, \sigma_P I_p)$
- $\theta_Q \leftarrow (w_\alpha^0, \sigma_P)$ ▷ Q trainable params
- ▷ Let $Q(\theta_Q) = \mathcal{N}(w_\alpha^0, \sigma_P I_p)$
- **for** $i \leftarrow 1$ to T **do**
  - Sample minibatch $S' \in S \setminus S_\alpha$, $|S'| = b$.
  - $\theta_Q \leftarrow \theta_Q - \eta \nabla_{\theta_Q} \Psi_\delta^\dagger(Q(\theta_Q), P; S \setminus S_\alpha)$
- Bound $\leftarrow \Psi_\delta^*(Q(\theta_Q), P; S \setminus S_\alpha)$
- **return** Bound

function GET-BOUND$(\mathcal{E}, \alpha, T, \sigma_P)$
- $S_\alpha \leftarrow \{z_1, .., z_{\alpha|S|}\} \subset S$
- $w_\alpha^0 \leftarrow \mathrm{SGD}(w_0, S_\alpha, b, \frac{|S_\alpha|}{b})$
- ▷ Perform *base* run
- $w_S \leftarrow \mathrm{SGD}(w_\alpha^0, S, b, \infty, \mathcal{E})$
- ▷ Perform α-*prefix+ghost* run
- $w_\alpha^G \leftarrow \mathrm{SGD}(\ w_\alpha^0, S_\alpha^G, b, T, \cdot)$
- $P \leftarrow \mathcal{N}(w_\alpha^G, \sigma_P I_p)$
- $Q \leftarrow \mathcal{N}(w_S, \sigma_P I_p)$
- Bound $\leftarrow \Psi_\delta^*(Q, P; S \setminus S_\alpha)$
- **return** Bound

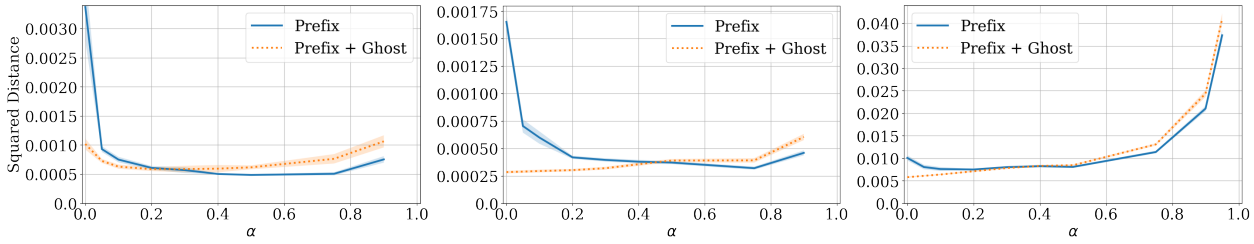

Figure 3: **left**: MNIST, LeNet-5; **center**: Fashion-MNIST, LeNet-5; **right**: MNIST, FC; **x-axis**: α used for α-*prefix* α-*prefix+ghost* runs; **y-axis**: squared L^2 distance divided by $(1-\alpha)|S|$. For a Gaussian priors and posteriors with fixed covariance, smaller distances yields tighter bounds.

learned weights. This experiment shows that we can predict fairly well the final weights of SGD on the full data set using only a fraction of the data set, implying that most of the variability in SGD comes in the beginning of training. Crucially, the two runs are *coupled* in the same manner as the ghost-sample runs: the first α-fraction of first epoch is identical. When only a fraction of the data is available, SGD treats this data as the entire data set, starting its second epoch immediately.

5 EMPIRICAL METHODOLOGY

Example 3.2 shows that a data-oracle priors can yield tighter generalization bounds than an oracle prior. In this section, we describe the experimental methodology we use to evaluate this phenomenon in neural networks trained by stochastic gradient descent (SGD).

Pseudocode. Algorithm 1 (right) describes the procedure for obtaining a PAC-Bayes risk bound on a network trained by SGD.[4] Note that the steps outlined in Lines 1–3 do not change with σ_P and therefore the best σ_P can be chosen efficiently without rerunning the optimization. If ghost data is not used, S_α^G should be replaced with S_α.

Algorithm 2 Stochastic Grad. Descent

Hyperparameters: Learning rate η

function $\mathrm{SGD}(w_0, S, b, t, \mathcal{E} = -\infty)$
- $w \leftarrow w_0$
- **for** $i \leftarrow 1$ to t **do**
  - Sample $S' \in S$, $|S'| = b$
  - $w \leftarrow w - \eta \nabla L_{S'}(w)$
  - **if** $L_S^{0-1}(w) \le \mathcal{E}$ **then** break
- **return** w

To avoid choosing β, we use a variational KL bound, described in Appendix D, which allows us to optimize β a posteriori for a small penalty. This PAC-Bayes bound on risk, denoted $\Psi_\delta^*(Q, P; S \setminus S_\alpha)$, is evaluated with $\delta = 0.05$ confidence level in all of our experiments during evaluation/optimization.

Datasets and Architectures. We use three datasets: MNIST, Fashion-MNIST and CIFAR-10. See Appendix E for more details. The architectures used are described in detail in Appendix F. For the details

[4] Algorithm 1 (right) uses a fixed learning rate and a vanilla SGD for simplicity, but the algorithm can be adapted to any variants of SGD with different learning rate schedules.

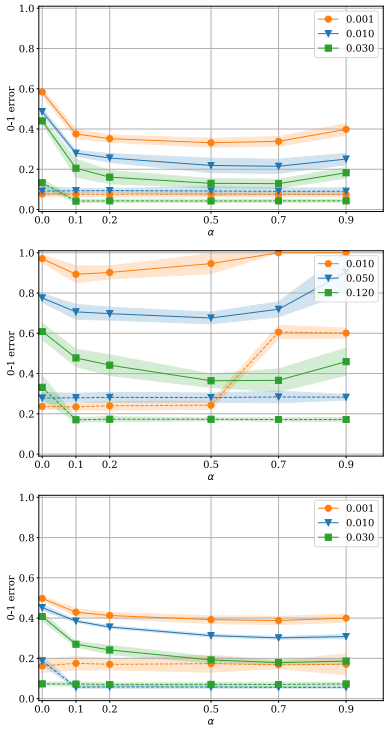

Figure 4: **top**: MNIST, LeNet-5; **center**: Fashion-MNIST, LeNet-5; **bottom**: MNIST, FC; **y-axis**: error-rate; **x-axis**: fraction α of the data used by the α-*prefix* run of SGD to predict the weights produced by the *base* run of SGD, w_S; **dashed lines**: test error; **solid lines**: error bound for a Gaussian Gibbs classifier Q, with mean w_S and isotropic covariance minimizing a PAC-Bayes risk bound; **legend**: training error used as the stopping criterion for the *base* run of SGD. The best error bound on MNIST ($\approx 11\%$) is significantly better than the 46% bound by Zhou et al. (2019).

of the training procedure, see Appendix G.

Stopping criteria. We terminate SGD optimization in the *base* run once the empirical error (L^{0-1} in Algorithms 1 and 2) measured on all of S fell below some desired value $\mathcal{E}$, which we refer to as the stopping criteria. We evaluate the results for different stopping criteria.

6 EMPIRICAL STUDY OF TRAINED NETWORKS

Evaluating data-dependent priors. A PAC-Bayes risk bound trades off empirical risk and the contribution coming from the KL term. For isotropic Gaussian priors and posteriors, the mean component in the KL is proportional to the squared difference in means normalized by the effective number of training samples not seen by the prior, i.e., $d(\alpha, S_\alpha) := \frac{\|w_S - w_\alpha\|_2^2}{(1-\alpha)|S|}$. This *scaled squared L2 distance* term determines the tightness of the bound when the prior variance and the posterior Q and data S are fixed, as the bound grows with $d(\alpha, S_\alpha)$. In this section we empirically evaluate how $d(\alpha, S_\alpha)$ and $d(\alpha, S_\alpha^G)$ vary with different values of α.

Our goal is to evaluate whether, on standard vision datasets and architectures, a data-dependent oracle prior can be superior to an oracle prior. Since we do not have access to an oracle prior, we approximate it by using a ghost sample S_α^G with $\alpha = 0$, as described in Section 4.2. Data-dependent oracle priors are approximated by using a combination of training samples and ghost samples.

Our experimental results on MNIST and Fashion-MNIST appear in Fig. 3, where we plot $d(\alpha, S_\alpha)$ and $d(\alpha, S_\alpha^G)$. The results suggest that the value of α minimizing $d(\alpha, S_\alpha^G)$ is data- and architecture-dependent. The optimal prefix size for MNIST, FC minimizing $d(\alpha, S_\alpha)$ is $\alpha > 0.2$. For MNIST, LeNet-5 and Fashion-MNIST, LeNet-5, the optimal α is between 0 and 0.1. We found that batch size affects the optimal α, whether on α-*prefix* or ghost data. As one might expect, the best α is larger for smaller batch sizes. We hypothesize that this is due to increased stochasticity of SGD.

Interestingly, at larger values of α we observe that the gap between $d(\alpha, S_\alpha)$ and $d(\alpha, S_\alpha^G)$ closes. This hap-

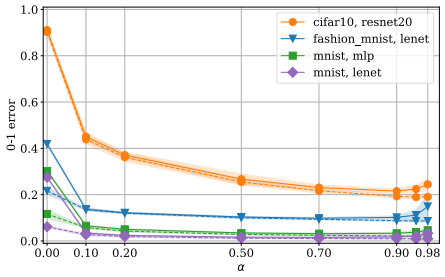

Figure 5: **Y-axis**: error-rate; **x-axis**: fraction α of the data used to learn the prior mean; **dashed lines**: test error; **solid lines**: bound on the error of a Gaussian Gibbs classifier whose mean and diagonal covariance are learned by optimizing the bound surrogate; **legend**: dataset and network architecture. For each scenario, under the optimal α, the bound is tight and test error is within a few percent of standard SGD-trained networks.

pens in all three experimental setups by $\alpha = 0.4$: we observe that the prior mean obtained with S_α training data alone is as close to final SGD weights as the prior mean obtained with S_α^G.

Generalization bounds for SGD-trained networks. We apply data-dependent priors to obtain tighter PAC-Bayes risk bounds for SGD-trained networks. We do not use ghost data in these experiments, as oracle priors are inaccessible in practice. Thus the prior mean is obtained by the α-*prefix* run on prefix data alone. See Algorithm 1 (right) and Section 5 for the details of the experiment.

From the data in Fig. 4, it is apparent that α has a significant impact on the size of the bound. In all of the three networks tested, the best results are achieved for $\alpha > 0$.

One of the clearest relationships to emerge from the data is the dependence of the bound on the stopping criterion: The smaller the error at which the *base* run was terminated, the looser the bound. This suggests that the extra optimization introduces variability into the weights that we are not able to predict well. In Appendix I, we use oracle bounds to quantify limits on how much tighter these generalization bounds could be, were we able to optimize a diagonal prior variance. The results suggest that a diagonal prior offers little advantage over an isotropic prior.

Direct risk bound minimization. One of the dominant approaches to training Gaussian neural networks is to minimize the evidence lower bound (ELBO), which essentially takes the same form as the bound in Theorem 2.1, but with a different relative weight on the KL term. Here, we optimize a PAC-Bayes bound using our data-dependent prior methodology which can be related to empirical Bayes approaches. The details of the algorithm are outlined in Algorithm 1, left, where $\Psi_\delta^\dagger(Q, P; S \setminus S_\alpha)$ denotes a PAC-Bayes bound computed with differentiable surrogate loss. We perform experiments on 3 different datasets and architectures (see Appendix H for further details).

Fig. 5 presents the error of the posterior Q (dashed line) optimized using Algorithm 1 with different values of α. It is apparent from the figure that for all the networks and datasets tested, the error of Q drops dramatically as α increases, all the way up to around $\alpha = 0.9$. Note that Q with the optimal α achieves very high performance even compared to state-of-the-art networks and at the same time comes with a valid guarantee on error. For example, ResNet20 (without data augmentation and weight decay) trained on CIFAR10 achieved error of around 0.16, and the best-performing Q in Fig. 5 gets an average error of ≈ 0.2 with a bound ≈ 0.23 that holds with 0.95 probability.

Open-source implementation. Code for replicating the main empirical results is available at https://github.com/kylehkhsu/role-of-data.

Acknowledgment

The authors would like to thank Blair Bilodeau and Jeffrey Negrea for feedback on drafts of this work. DMR is supported, in part, by an NSERC Discovery Grant. Additional resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute.

References

Pierre Alquier and Benjamin Guedj (2018). "Simpler PAC-Bayesian bounds for hostile data". *Machine Learning* 107.5, pp. 887–902.

Amiran Ambroladze, Emilio Parrado-Hernández, and John Shawe-Taylor (2007). "Tighter PAC-Bayes bounds". In *Adv. Neural Information Processing Systems*, pp. 9–16.

Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy (2016). "PAC-Bayesian bounds based on the Rényi divergence". In *Artificial Intelligence and Statistics*, pp. 435–444.

Olivier Catoni (2003). *A PAC-Bayesian approach to adaptive classification*. Tech. rep. Laboratoire de

Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7.

Olivier Catoni (2007). *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*. Vol. 56. Lecture Notes-Monograph Series. IMS. arXiv: 0712.0248.

Gintare Karolina Dziugaite and Daniel M. Roy (2017). "Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data". In *Proc. 33rd Ann. Conf. Uncertainty in Artificial Intelligence (UAI)*. arXiv: 1703.11008.

Gintare Karolina Dziugaite and Daniel M. Roy (2018). "Data-dependent PAC-Bayes priors via differential privacy". In *Adv. Neural Information Processing Systems*. arXiv: 1802.09583.

Benjamin Guedj (2019). "A Primer on PAC-Bayesian Learning". In *Proceedings of the 2nd Congress of the Société Mathématique de France*, pp. 391–414. arXiv: 1901.05353.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep residual learning for image recognition". In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

Alex Krizhevsky (2009). "Learning multiple layers of features from tiny images". URL: https://www.cs.toronto.edu/~kriz/cifar.html.

John Langford (2002). "Quantitatively tight sample complexity bounds". PhD thesis. Carnegie Mellon University.

John Langford and Avrim Blum (2003). "Microchoice bounds and self bounding learning algorithms". *Machine Learning* 51.2, pp. 165–179.

John Langford and Matthias Seeger (2001). *Bounds for Averaging Classifiers*. Tech. rep. CMU-CS-01-102. Carnegie Mellon University.

Yann LeCun, Corinna Cortes, and Christopher J. C. Burges (1998). *MNIST handwritten digit database*. URL: http://yann.lecun.com/exdb/mnist.

Guy Lever, François Laviolette, and John Shawe-Taylor (2010). "Distribution-dependent PAC-Bayes priors". In *Proc. Int. Conf. Algorithmic Learning Theory (ALT)*. Springer, pp. 119–133.

Guy Lever, François Laviolette, and John Shawe-Taylor (2013). "Tighter PAC-Bayes bounds through distribution-dependent priors". *Theoretical Computer Science* 473, pp. 4–28.

Andreas Maurer (2004). *A note on the PAC-Bayesian theorem*. arXiv: cs/0411099.

David A. McAllester (1999a). "PAC-Bayesian Model Averaging". In *Proc. 12th Ann. Conf. Computational Learning Theory*. COLT '99. Santa Cruz, California, USA: ACM, pp. 164–170. URL: http://doi.acm.org/10.1145/307400.307435.

David A. McAllester (Dec. 1999b). "Some PAC-Bayesian Theorems". *Machine Learning* 37.3, pp. 355–363. URL: https://doi.org/10.1023/A:1007618624809.

David A. McAllester (2013). *A PAC-Bayesian Tutorial with A Dropout Bound*. arXiv: 1307.2118.

Vaishnavh Nagarajan and J. Zico Kolter (2019a). *Deterministic PAC-Bayesian generalization bounds for deep networks via generalizing noise-resilience*. See also https://openreview.net/forum?id=Hygn2o0qKX.

Vaishnavh Nagarajan and J. Zico Kolter (2019b). "Uniform convergence may be unable to explain generalization in deep learning". In *Adv. Neural Information Processing Systems*.

Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M Roy (2019). "Information-Theoretic Generalization Bounds for SGLD via Data-Dependent Estimates". In *Adv. Neural Information Processing Systems*, pp. 11013–11023.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro (2018). "A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks". In *Int. Conf. Learning Representationas (ICLR)*.

Luca Oneto, Davide Anguita, and Sandro Ridella (2016). "PAC-bayesian analysis of distribution dependent priors: Tighter risk bounds and stability analysis". *Pattern Recognition Letters* 80, pp. 200–207.

Luca Oneto, Sandro Ridella, and Davide Anguita (2017). "Differential privacy and generalization:

Sharper bounds with applications". *Pattern Recognition Letters* 89, pp. 31–38.

Emilio Parrado-Hernández, Amiran Ambroladze, John Shawe-Taylor, and Shiliang Sun (2012). "PAC-Bayes bounds with data dependent priors". *J. Machine Learning Research* 13.Dec, pp. 3507–3531.

Omar Rivasplata, Emilio Parrado-Hernandez, John Shaws-Taylor, Shiliang Sun, and Csaba Szepesvari (2018). "PAC-Bayes bounds for stable algorithms with instance-dependent priors". In *Adv. Neural Information Processing Systems*.

Omar Rivasplata, Vikram M Tankasali, and Csaba Szepesvari (2019). *PAC-Bayes with Backprop*. arXiv: 1908.07380.

John Shawe-Taylor, Peter L Bartlett, Robert C Williamson, and Martin Anthony (1996). "A framework for structural risk minimisation". In *Proc. 9th Ann. Conf. Computational Learning Theory (COLT)*, pp. 68–76.

John Shawe-Taylor and Robert C Williamson (1997). "A PAC analysis of a Bayesian estimator". In *Proc. 10th Ann. Conf. Computational Learning Theory (COLT)*. ACM, pp. 2–9.

Han Xiao, Kashif Rasul, and Roland Vollgraf (2017). *Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms*. arXiv: 1708.07747.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P Adams, and Peter Orbanz (2019). "Non-vacuous Generalization Bounds at the ImageNet Scale: a PAC-Bayesian Compression Approach". In *Proc. Int. Conf. Learing Representations (ICLR)*. arXiv: 1804.05862.