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

$I(S; H) = \mathbb{E}[\mathrm{KL}(Q(S)||P^*)] = \inf_{P'} \mathbb{E}[\mathrm{KL}(Q(S)||P')]$, 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 $\frac{dQ}{dP} : \mathcal{H} \to \mathbb{R}_+ \cup \{\infty\}$ for some Radon–Nikodym derivative (aka, density) of Q with respect to P. The Kullback–Liebler (KL) divergence from Q to P is $\mathrm{KL}(Q||P) = \int \ln \frac{dQ}{dP} dQ$ 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

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

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 $Q \in \mathcal{M}_1(\mathcal{H})$,*

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

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

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

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,

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

and

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

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 out-perform 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 $i = 1, \ldots, n$,
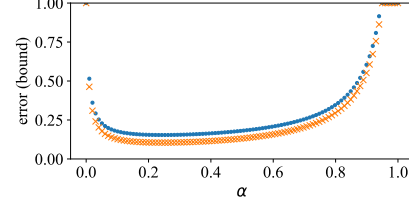

Figure 1: Lower (orange x's) and upper (blue dots) bounds on the expected value of a linear PAC-Bayes bound as a function of the fraction, α, of the 100 training data used by the data-dependent (PAC-Bayes) prior. Each bound uses the optimal (in expectation) tradeoff β and data-dependent prior $P(S_J)$, 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 $\alpha = 0.24$.).

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 $(y_i, \boldsymbol{x}_i)$.

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 al-

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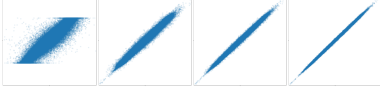
Figure 2: MNIST, FC; **x-axis**: parameter values of *base* run; **y-axis**: parameter values of α-*prefix* run; **left to right:** α values equal to $\{0, 0.1, 0.5, 0.9\}$. As α increases, the correlation between the parameters learnt by SGD on all of the data and an α fraction of the data increases.

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_α^G 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_α^G.) Note that S_α^G is also equal in dis-

tribution to S. We may then take w_α^G to be the mean of $Q(U, S_\alpha^G)$, i.e., the weights produced by SGD on the data set S_α^G using the randomness U.

By design, SGD acting on S_α^G 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_α^G and w_S are equal in distribution when conditioned on S_α and U, and so w_α^G 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_α^G, we call an SGD run training on S_α^G an α-*prefix+ghost* run, obtaining parameters w_α^G.

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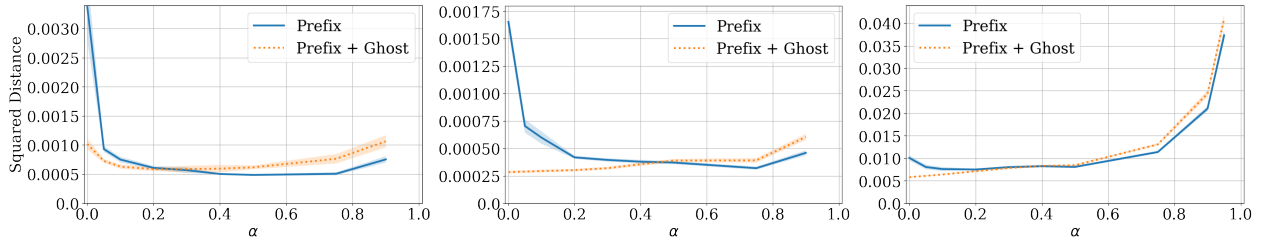

Figure 3: **left**: MNIST, LeNet-5; **center**: Fashion-MNIST, LeNet-5; **right**: MNIST, FC; **x-axis**: α used for α-*prefix* α-*prefix+ghost* runs; **y-axis**: squared L^2 distance divided by $(1 - \alpha)|S|$. For a Gaussian priors and posteriors with fixed covariance, smaller distances yields tighter 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 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 $L_S^{0-1}(w) \leq \mathcal{E}$ **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.
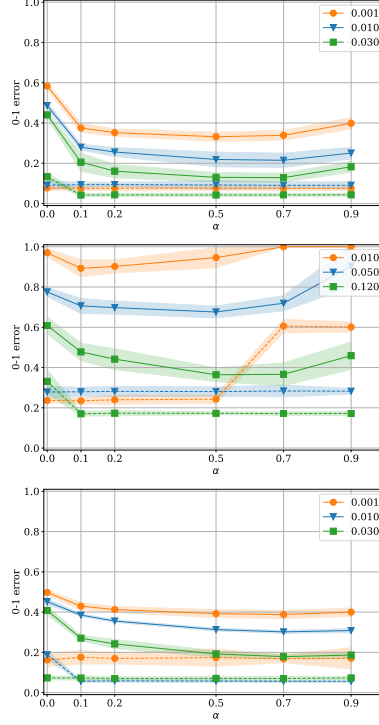
Figure 4: **top**: MNIST, LeNet-5; **center**: Fashion-MNIST, LeNet-5; **bottom**: MNIST, FC; **y-axis**: error-rate; **x-axis**: fraction α of the data used by the α-*prefix* run of SGD to predict the weights produced by the *base* run of SGD, w_S; **dashed lines**: test error; **solid lines**: error 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-
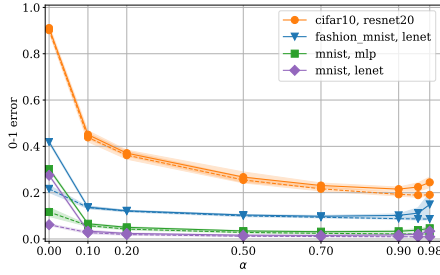
Figure 5: **Y-axis**: error-rate; **x-axis**: fraction α of the data used to learn the prior mean; **dashed lines**: test error; **solid lines**: bound on the error of a Gaussian Gibbs classifier whose mean and diagonal covariance are learned by optimizing the bound surrogate; **legend**: dataset and 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 CI-FAR10 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

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.

Supplementary Materials

A   Proof of Proposition 3.1

The proof is straightforward, following essentially from definitions. We give it here for completeness.

We pause to make two remarks about J:

1. We do not require J to have any particular distribution and so, e.g., J could be uniformly distributed among subsets of cardinality αn or could be a.s. nonrandom and equal to $[m]$.

2. The statement that $\mathbb{P}[\hat{h}|S, J] = Q(S)$ a.s. implies that $\mathbb{P}[\hat{h}|S] = Q(S)$ a.s. and that $\hat{h}$ is independent of J, both marginally and conditionally on S. Informally, any randomness in J plays no role in the determination of $\hat{h}$.

Let $\bar{J} = [n] \setminus J$, fix $\mathcal{F}$, and consider the linear PAC-Bayes bound *based on* $L_{S_{\bar{J}}}(Q(S))$, i.e., where we use the data in $S_{\bar{J}}$ to estimate the risk of $Q(S)$. By the linear PAC-Bayes theorem, we are permitted to choose our prior based on S_J, since S_J is independent of $S_{\bar{J}}$. In fact, we can also choose our prior knowing J, due to the independence outlined above in the second remark.

Conditionally on J, the expected value of the linear PAC-Bayes bound under the data-dependent oracle prior is the infimum

$$\inf_{P \in Z^m \to \mathcal{F}} \mathbb{E}_S[\Psi_{\beta,\delta}(Q(S), P(S_J); S_{\bar{J}})] \tag{8}$$

$$= \mathbb{E}_S[\beta^{-1} L_{S_{\bar{J}}}(Q(S))] + \inf_{P \in Z^m \to \mathcal{M}_1(\mathcal{H})} \frac{\mathbb{E}_S[\mathrm{KL}(Q(S)||P(S_J))] + \log \frac{1}{\delta}}{2\beta(1-\beta)(1-\alpha)n} \tag{9}$$

$$= \mathbb{E}_S[\beta^{-1} L_{S_{\bar{J}}}(Q(S))] + \frac{\mathbb{E}_S[\mathrm{KL}(Q(S)||\mathbb{E}[Q(S)|S_J])] + \log \frac{1}{\delta}}{2\beta(1-\beta)(1-\alpha)n} \tag{10}$$

$$= \mathbb{E}_S[\beta^{-1} L_{S_{\bar{J}}}(Q(S))] + \frac{I_{\mathcal{F}}^J(\hat{h}; S|S_J) + \log \frac{1}{\delta}}{2\beta(1-\beta)(1-\alpha)n}. \tag{11}$$

Here $\mathbb{E}_S$ denotes expectation over $S \sim \mathcal{D}^n$, conditional on J. Note that the optimal prior here depends on J. We arrive at the unconditional expected value of the bound by taking expectations also over J, which changes the *disintegrated* (conditional) mutual information $I_{\mathcal{F}}^J(\hat{h}; S|S_J)$ into a (conditional) mutual information $I_{\mathcal{F}}(\hat{h}; S|S_J, J)$, that is no longer a random quantity.

It follows immediately that the data-dependent risk bound is tighter, in expectation, that the bound based on $J = \emptyset$, when

$$
(1 - \beta)\, \mathbb{E}[L_S(Q(S))] + \frac{I_{\mathcal{F}}(\hat{h}; S) + \log \frac{1}{\delta}}{2n}
$$
$$
> (1 - \beta)\, \mathbb{E}[L_{S_{\bar{J}}}(Q(S))] + \frac{I_{\mathcal{F}}(\hat{h}; S|S_J, J) + \log \frac{1}{\delta}}{2(1 - \alpha)n}. \tag{12}
$$

The statement of the proposition is obtained by simple manipulations. The above inequality is equivalent to

$$
\frac{I_{\mathcal{F}}(\hat{h}; S) + \log \frac{1}{\delta}}{2n} - \frac{I_{\mathcal{F}}(\hat{h}; S|S_J, J) + \log \frac{1}{\delta}}{2(1 - \alpha)n}
$$
$$
> (1 - \beta)\, \mathbb{E}[L_{S_{\bar{J}}}(Q(S)) - L_S(Q(S))]. \tag{13}
$$

Rewriting the left-hand side,

$$
\frac{I_{\mathcal{F}}(\hat{h}; S) + \log \frac{1}{\delta}}{2n} - \frac{I_{\mathcal{F}}(\hat{h}; S|S_J, J) + \log \frac{1}{\delta}}{2(1 - \alpha)n}
$$
$$
= \frac{1}{2}\left(\frac{I_{\mathcal{F}}(\hat{h}; S)}{n} - \frac{I_{\mathcal{F}}(\hat{h}; S|S_J, J)}{(1 - \alpha)n}\right) - \frac{\log \frac{1}{\delta}}{2n}\left(\frac{\alpha}{1 - \alpha}\right). \tag{14}
$$

Therefore, we prefer a data-dependent prior based on J when

$$
\left(\frac{I_{\mathcal{F}}(\hat{h}; S)}{n} - \frac{I_{\mathcal{F}}(\hat{h}; S|S_J, J)}{(1 - \alpha)n}\right)
$$
$$
> 2(1 - \beta)\, \mathbb{E}[L_{S_{\bar{J}}}(Q(S)) - L_S(Q(S))] + \frac{\log \frac{1}{\delta}}{n}\left(\frac{\alpha}{1 - \alpha}\right). \tag{15}
$$

The result follows by the definition of the information rate gain and excess bias. Note that, if J is (a.s.) nonrandom, then $I_{\mathcal{F}}(\hat{h}; S|S_J, J)$ is simply $I_{\mathcal{F}}(\hat{h}; S|S_J)$.

B   Proof of Lemma 3.3

We begin with a proof sketch.

Sketch. With J and β fixed, the minimization over $P(S_J)$ meets the hypotheses of Theorem 2.2 and so we may simplify the objective by taking $P(S_J) = \mathbb{E}[Q(S)|S_J] = \mathbb{P}[W|S_J]$. The KL term then becomes a conditional mutual information $I(W; S \setminus S_J | S_J)$. Due to linearity of expectation, we may then optimize β explicitly, leaving only a minimization over subsets J,

$$
\inf_{J \subseteq [n]}\left(\Phi(J) := R(J) + C(J) + \sqrt{2R(J)C(J) + C^2(J)}\right)
$$

where $R(J) = \mathbb{E}[L_{S \setminus S_J}(Q)]$ and $C(J) = (I(W; S \setminus S_J | S_J) + \log \frac{1}{\delta})/|S \setminus S_J|$.

One can show that $I(W; S \setminus S_J | S_J) = \frac{D}{2} \ln \phi_{\bar{J}}/\kappa$, where $\phi_{\bar{J}}$ is the variance contribution from $S \setminus S_J$ and ξ. Using sub-Gaussian and sub-exponential tail bounds, one can establish that $R(J) \leq \overline{R} = \exp\{-D/16\} + \exp\{-\tau^2/(4\phi_{[n]} \sigma^2)\}$, where $\phi_{[n]}$ is due to variance in S, ξ, and $\tau = (\sum_{i=1}^n \eta_i)\|u\|^2$.

Choosing $n = 100$, $D = 1000$, $\sigma = 8$, $\kappa = 4$, $\tau = 64$, and $\delta = 0.05$, we obtain $\Phi(\emptyset) \geq 2C(\emptyset) \approx 1.1$, while $\min_J \Phi(J) \lesssim 0.15$. Our upper bound is achieved by $J = [24]$, i.e., by using the initial 24 data points to obtain a data-dependent (oracle) prior. □

B.1 Complete proof and bounds on the objective

We now provide a complete rigorous proof. For subsets J (of $[n]$), let $\bar{J} = [n] \setminus J$; let $\eta_J^p = \sum_{i \in J} \eta_i^p$ for $p \in \{1, 2\}$; let $\phi_J = \eta_J^2 \sigma^2 / D + \kappa$; and let $\phi_{-i} = \phi_{[n] \setminus \{i\}}$.

By Theorem 2.2 and linearity of expectation, for every subset J and $\beta \in (0, 1)$, Theorem 2.2 implies that the optimal prior is $P_J(S_J) = \mathbb{P}[W | S_J]$, and so we can simplify Eq. (6) by choosing this prior. In particular, now $\mathbb{E}[\text{KL}(Q || P_J(S_J))] = I(W; S_{\bar{J}} | S_J)$.

Define $R(J) = \mathbb{E}[L_{S_{\bar{J}}}(Q)]$ and $C(J) = (I(W; S_{\bar{J}} | S_J) + \log \frac{1}{\delta}) / |S_{\bar{J}}|$. By linearity of expectation, we can remove the infimum over $\beta \in (0, 1)$ by explicit minimization. As a result, we see that Eq. (6) is equivalent to

$$\inf_{J \subseteq [n]} \underbrace{R(J) + C(J) + \sqrt{2R(J)C(J) + C^2(J)}}_{\Phi(J)}. \tag{16}$$

Pick some $J \subseteq [n]$. Then the optimal prior conditioned on S_J is

$$P_J(S_J) = \mathbb{E}[Q(S) | S_J] = \delta_{\eta_{[n]}^1 \boldsymbol{u}} \otimes N_J, \tag{17}$$

where $N_J = \mathcal{N}(\sum_{i \in J} \eta_i y_i \boldsymbol{x}_{i,2}, \phi_{\bar{J}} I_D)$. Let $\psi(r) = r - 1 - \ln r$ for $r > 0$. Then

$$\text{KL}(Q(S) || P_J(S_J)) = D\psi(\kappa/\phi_{\bar{J}})/2 + \frac{1}{2\phi_{\bar{J}}} \sum_{j=1}^{D} \left(\sum_{i \notin J} \eta_i y_i \boldsymbol{x}_{i,2,j} \right)^2. \tag{18}$$

Taking expectations,

$$I(W; S_{\bar{J}} | S_J) = \mathbb{E}[\text{KL}(Q(S) || P_J(S_J))] = \frac{D}{2} \left(\psi(\kappa/\phi_{\bar{J}}) + \frac{\sigma^2 \eta_{\bar{J}}^2 / D}{\phi_{\bar{J}}} \right) \tag{19}$$

$$= \frac{D}{2} \left(\psi(\kappa/\phi_{\bar{J}}) + (1 - \kappa/\phi_{\bar{J}}) \right) \tag{20}$$

$$= \frac{D}{2} \ln \phi_{\bar{J}}/\kappa. \tag{21}$$

It remains to control the empirical risk term. To that end, pick $i \in [n]$ and let $\tau = \eta_{[n]}^1 \|u\|^2$. Then

$$\mathbb{E}\ell(W, z_i) = \mathbb{P}[y_i \langle W, \boldsymbol{x}_i \rangle \le 0] = \mathbb{P}[\tau + y_i \langle w_{n,2}, \boldsymbol{x}_{i,2} \rangle + y_i \langle \xi, \boldsymbol{x}_{i,2} \rangle \le 0], \tag{22}$$

where

$$y_i \langle w_{n,2}, \boldsymbol{x}_{i,2} \rangle = \eta_i \|\boldsymbol{x}_{i,2}\|^2 + \sum_{j \ne i} \eta_j y_i y_j \langle \boldsymbol{x}_{j,2}, \boldsymbol{x}_{i,2} \rangle. \tag{23}$$

Rearranging and exploiting the chain rule of conditional expectation and symmetry of the normal distribution,

$$\mathbb{E}\ell(W, z_i) = \mathbb{E}\mathbb{P}^{\boldsymbol{x}_{i,2}} \left[\sum_{j \ne i} \langle \eta_j \, \boldsymbol{x}_{j,2}, \boldsymbol{x}_{i,2} \rangle + \langle \xi, \boldsymbol{x}_{i,2} \rangle \ge \tau + \eta_i \|\boldsymbol{x}_{i,2}\|^2 \right],$$

where the conditional probability is a tail bound on a univariate Gaussian with mean zero and variance $\|\boldsymbol{x}_{i,2}\|^2 \phi_{-i}$.

Applying the standard (sub-)Gaussian tail bound,

$$\mathbb{E}\ell(W, z_i) \le \mathbb{E}\exp\left\{ -\frac{1}{2} \frac{(\tau + \eta_i \|\boldsymbol{x}_{i,2}\|^2)^2}{\|\boldsymbol{x}_{i,2}\|^2 \phi_{-i}} \right\} \le \mathbb{E}\exp\left\{ -\frac{\tau^2}{2\|\boldsymbol{x}_{i,2}\|^2 \phi_{-i}} \right\}, \tag{24}$$

where the last inequality is crude, but suffices for our application. Note that $D\|\boldsymbol{x}_{i,2}\|^2 / \sigma^2$ is a chi-squared random variable with D degrees of freedom, hence sub-exponential. Indeed, with probability at least $1 - c$,

$$D\|\boldsymbol{x}_{i,2}\|^2 / \sigma^2 \le D + 2\sqrt{D \log(1/c)} + 2\log(1/c). \tag{25}$$

Rearranging,

$$\|\boldsymbol{x}_{i,2}\|^2 \leq \frac{\sigma^2}{D}(D + 2\sqrt{D\log(1/c)} + 2\log(1/c)) \tag{26}$$

$$\leq \sigma^2(1 + 4\sqrt{\log(1/c)/D}) =: B(c), \tag{27}$$

where the second inequality holds assuming $c \geq \exp\{-D\}$, which we will ensure from this point on. So

$$\mathbb{E}\ell(W, z_i) \leq \inf_{c \geq e^{-D}} \left\{ \frac{1}{c} + \left(1 - \frac{1}{c}\right)\exp\{-\tau^2/(2\phi_{-i}B(c))\}. \right\}. \tag{28}$$

Taking $c = \exp\{-D/16\}$, we have $B(c) = 2\sigma^2$. Then, using $\phi_{-i} \leq \phi_{[n]}$,

$$\mathbb{E}\ell(W, z_i) \leq \underbrace{\exp\{-D/16\} + \exp\{-\tau^2/(4\phi_{[n]}\,\sigma^2)\}}_{\overline{R}}. \tag{29}$$

We may now obtain a bound

$$R(J) = \mathbb{E}L_{S_{\bar{J}}}(Q(S)) = \frac{1}{n - |J|} \sum_{i \notin J} \mathbb{E}\ell(W, z_i) \leq \max_{i \notin J} \mathbb{E}\ell(W, z_i) = \overline{R}. \tag{30}$$

Thus

$$\Phi(J) \leq \overline{R} + C(J) + \sqrt{2\overline{R}\,C(J) + C(J)^2} \tag{31}$$

At the same time, we have $\Phi(J) \geq 2C(J)$ for all $J \subseteq [n]$. (**Note that these two bounds are used to produce Fig. 1.**)

In particular, noting $\log 1/\delta > 0$,

$$\Phi(\emptyset) \geq \frac{D}{m} \ln \frac{\sigma^2 \eta_{[n]}^2/D + \kappa}{\kappa}. \tag{32}$$

The result can be seen to follow from these bounds by evaluation using the particular values. In particular, one can see that taking J to be a nonempty initial segment of $[n]$, we have $\Phi(J) < 2C(\emptyset) \leq \Phi(\emptyset)$.

C   Analytic form of the KL for an approximate data-dependent oracle bound

In this section, we explore one possible analytic bound for a KL term for a PAC-Bayes bound, based on the setup in Section 4. We assume $\operatorname{tr}(\Sigma)$ and $\det(\Sigma)$ are nonrandom. In an application, one would have to cover a set of possible values to handle the random case.

The KL divergence between Gaussians $Q(U, S) = \mathcal{N}(w_S, \Sigma)$ and $P = \mathcal{N}(w_\alpha, \Sigma_\alpha)$ takes the form

$$2\mathrm{KL}(Q(U,S)\|P) = \underbrace{\|w_S - w_\alpha\|^2_{\Sigma_\alpha^{-1}}}_{\text{mean component}} + \underbrace{\operatorname{tr}(\Sigma_\alpha^{-1}(\Sigma)) - p + \ln\frac{\det\Sigma_\alpha}{\det\Sigma}}_{\text{variance component}}. \tag{33}$$

Specializing to an isotropic prior, i.e., $\Sigma_\alpha = \sigma_P I$, we obtain

$$2\mathrm{KL}(Q(U,S)\|P) = \underbrace{\frac{1}{\sigma_P}\|w_S - w_\alpha\|^2}_{\text{mean component}} + \underbrace{\frac{1}{\sigma_P}\operatorname{tr}(\Sigma) - p + p\ln\sigma_P - \ln\det\Sigma}_{\text{variance component}}. \tag{34}$$

Note that

$$\operatorname{tr}(\mathrm{cov}^{S_\alpha, U}(w_S)) = \inf_{w_\alpha} \mathbb{E}^{S_\alpha, U}[\|w_S - w_\alpha\|^2]. \tag{35}$$

Consider

$$\sigma_P = \frac{1}{p}\big(\text{tr}(\text{cov}^{S_\alpha,U}(w_S)) + \text{tr}(\Sigma)\big). \tag{36}$$

Substituting above, for some random variable Z such that $\mathbb{E}^{S_\alpha,U}[Z] = 1$,

$$2\text{KL}(Q(U,S)\|P) = Zp - p + p\ln\left\{\frac{\frac{1}{p}\text{tr}(\text{cov}^{S_\alpha,U}(w_S)) + \frac{1}{p}\text{tr}(\Sigma)}{(\det\Sigma)^{1/p}}\right\} \tag{37}$$

$$\leq Zp - p + p\frac{\frac{1}{p}\text{tr}(\text{cov}^{S_\alpha,U}(w_S)) + \frac{1}{p}\text{tr}(\Sigma) - (\det\Sigma)^{1/p}}{(\det\Sigma)^{1/p}}. \tag{38}$$

Taking expectations, conditional on S_α, U,

$$2\mathbb{E}^{S_\alpha,U}[\text{KL}(Q(U,S)\|P)] \leq \frac{\text{tr}(\text{cov}^{S_\alpha,U}(w_S))}{(\det\Sigma)^{1/p}} + p\frac{\frac{1}{p}\text{tr}(\Sigma) - (\det\Sigma)^{1/p}}{(\det\Sigma)^{1/p}}. \tag{39}$$

Further, if we assume $\Sigma = \sigma I$, then

$$\mathbb{E}^{S_\alpha,U}[\text{KL}(Q(U,S)\|P)] \leq \frac{1}{2\sigma}\text{tr}(\text{cov}^{S_\alpha,U}(w_S)). \tag{40}$$

D   Variational KL bound

The linear PAC-Bayes bound requires one to specify a value of β. For a particular posterior kernel Q, the optimal value of β depends on the likely value of the empirical risk term. However, the value of β must be chosen independently of the data used to evaluate the bound.

In the proof of Lemma 3.3 in Appendix B, the linear PAC-Bayes bound is optimized, in expectation. Since the expected value of the bound is independent of the data, and since the constant β can be pulled outside the expectations, we can choose the value of beta that minimizes the bound in expectation. The result is Eq. (16), with $C(J)$ defined in terms of an expected KL, as the mutual information appears only when the prior is chosen to be the oracle prior.

In this section, we describe how the bound due to Maurer (2004) can be approximated to reveal a high-probability tail bound with the same form as if we optimized β. The cost is a $O(\log\sqrt{m}/m)$ term.

Let $\mathcal{B}_p$ denote the Bernoulli distribution on $\{0,1\}$ with mean p. For $p, q \in [0,1]$, we abuse notation and define

$$\text{kl}(q\|p) \overset{\text{def}}{=} \text{KL}(\mathcal{B}_q\|\mathcal{B}_p) = q\ln\frac{q}{p} + (1-q)\ln\frac{1-q}{1-p}.$$

The following PAC-Bayes bound for bounded loss is due to Maurer (2004). The same result for 0–1 loss was first established by Langford and Seeger (2001), building off the seminal work of McAllester (1999a). See also (Langford, 2002) and (Catoni, 2007).

Theorem D.1 (PAC-Bayes; Maurer 2004, Thm. 5)**.** *Under bounded loss $\ell \in [0,1]$, for every $\delta > 0$, $m \in \mathbb{N}$, distribution $\mathcal{D}$ on Z, and distribution P on $\mathcal{H}$,*

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left((\forall Q)\ \text{kl}(L_S(Q)\|L_\mathcal{D}(Q)) \leq \frac{\text{KL}(Q\|P) + \ln\frac{2\sqrt{m}}{\delta}}{m}\right) \geq 1 - \delta. \tag{41}$$

One can recover the bound by McAllester (1999a) via Pinsker's inequality, resulting in a (looser) bound on $|L_S(Q) - L_\mathcal{D}(Q)|$. Maurer's bound behaves like the bound in Theorem 2.1, except that it holds for all β simultaneously, at the cost of a $\frac{1}{m}\log\sqrt{m}$ term.

D.1   Inverting the KL bound

Here we derive a novel PAC-Bayes bound that is an upper bound on the inverted KL bound (Theorem D.1) and that is used during optimization in our empirical work. The bound is the piecewise combination of two bounds. In independent work, Rivasplata, Tankasali, and Szepesvari (2019) derive the first of the two parts, which they call a "quad bound". The second part is a consequence of Pinsker's inequality.

Theorem D.2 (Variational KL bound). *With probability at least $1 - \delta$ over S,*

$$L_{\mathcal{D}}(Q) \le \min \begin{cases} L_S(Q) + B + \sqrt{B(B + 2L_S(Q))}, \\ L_S(Q) + \sqrt{\frac{B}{2}}, \end{cases} \tag{42}$$

where

$$B = \frac{\mathrm{KL}(Q||P) + \log \frac{2\sqrt{m}}{\delta}}{|S|}. \tag{43}$$

The variational KL bound takes the minimum value of the *moment bound* (top) and the *Pinsker bound* (bottom).

Proof. Let $\mathrm{kl}(L_S(Q)||L_{\mathcal{D}}(Q))$ be KL between two Bernoulli random variables with success probabilities $L_S(Q)$ and $L_{\mathcal{D}}(Q)$, respectively. Then by Theorem D.1, with probability greater than $1 - \delta$,

$$\mathrm{kl}(L_S(Q)||L_{\mathcal{D}}(Q)) \le \frac{\mathrm{KL}(Q||P) + \log \frac{2\sqrt{|S|}}{\delta}}{|S|}. \tag{44}$$

Let B denote the right hand side of the inequality. By Donsker–Varadhan we get

$$\mathrm{kl}(L_S(Q)||L_{\mathcal{D}}(Q)) \ge \lambda L_S(Q) - \log \underset{x \sim \mathrm{Ber}(L_{\mathcal{D}}(Q))}{\mathbb{E}} [e^{\lambda x}] \tag{45}$$

for any λ. The final term is the moment generating function of a Bernoulli random variable and so

$$\mathrm{kl}(L_S(Q)||L_{\mathcal{D}}(Q)) \ge \lambda L_S(Q) - \log(1 - L_S(Q) + L_S(Q)e^{\lambda}). \tag{46}$$

We can use this lower bound on $\mathrm{kl}(L_S(Q)||L_{\mathcal{D}}(Q))$ in Eq. (44). After rearranging, we obtain

$$-L_{\mathcal{D}}(Q)(1 - e^{\lambda}) \ge e^{\lambda L_S(Q) - B} - 1. \tag{47}$$

Take $\lambda \le 0$. Then

$$L_{\mathcal{D}}(Q) \le \frac{1 - e^{\lambda L_S(Q) - B}}{1 - e^{\lambda}}. \tag{48}$$

Using the inequality $1 - e^{-x} \le -x$ in the numerator of Eq. (48), we finally arrive at

$$L_{\mathcal{D}}(Q) \le L_S(Q) + B + \sqrt{B(B + 2L_S(Q))}. \tag{49}$$

Also, note that by Pinsker's inequality,

$$\mathrm{kl}(L_S(Q)||L_{\mathcal{D}}(Q)) \ge 2(L_S(Q) - L_{\mathcal{D}}(Q))^2, \tag{50}$$

and so

$$L_{\mathcal{D}}(Q) \le L_S(Q) + \sqrt{\frac{B}{2}}. \tag{51}$$

Both Eq. (51) and Eq. (49) are upper bounds on risk obtained from the inverted kl bound. Taking the minimum of the two bounds gives us the final result. □

The inverted KL bound is visualized in Fig. 6. We see that depending on the empirical risk and KL, either the moment or the Pinsker bound is tighter. The inverted KL bound is the minimum of the two and so is tight in both regimes. By taking the minimum of two bounds, we obtain a bound this is tighter over a wider range of values for the empirical risk and KL terms.
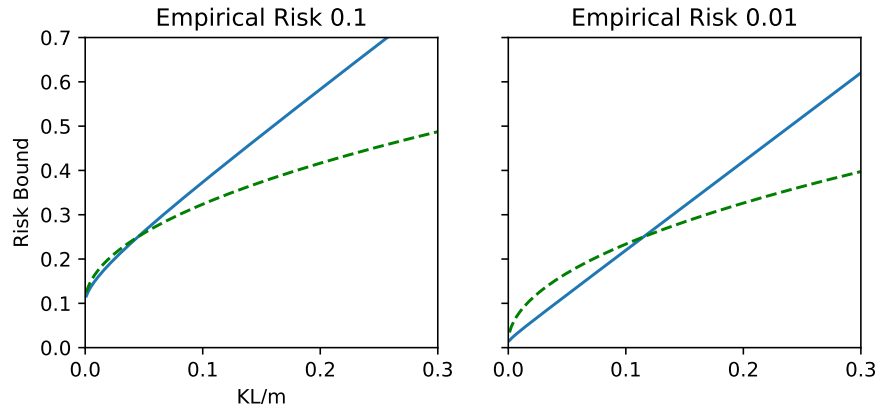
Figure 6: The two components of Variational KL bound visualized for $L_S(Q) = 0.1$ (left) and $L_S(Q) = 0.01$ (right). **Blue solid line:** moment bound; **Green dashed line:** Pinsker bound. The inverted KL bound on the risk is the minimum of the two lines.

E   Experimental details: datasets

We use three datasets in our experiments: 1) The MNIST dataset (LeCun, Cortes, and Burges, 1998), which consists of 28×28 grayscale images of handwritten decimal digits. 2) The Fashion-MNIST dataset (Xiao, Rasul, and Vollgraf, 2017), which consists of 28×28 grayscale images each associated with one of 10 categories (clothing and accessories). 3) The CIFAR-10 dataset (Krizhevsky, 2009), which consists of 32×32 RGB images each associated with one of ten categories (airplane, automobile, bird, etc.). For all datasets we use the standard training and test splits. This results in 60,000 training data for MNIST and Fashion-MNIST, 50,000 training data for CIFAR-10, and 10,000 test data for all three datasets. For CIFAR-10 we standardize all images according to the training split's statistics.

F   Experimental details: architectures

We use fully connected feed-forward multilayer perceptrons with ReLU activations for MNIST. We study networks with architecture 784–600–600–10 (featuring two hidden layers) in order to compare to Rivasplata, Tankasali, and Szepesvari (2019). Such a network has 837,610 parameters.

We also borrow the modified LeNet-5 architecture used by Zhou et al. (2019) in order to compare our bounds on SGD-trained classifiers. The network has 431,080 parameters. We use this architecture for MNIST and Fashion-MNIST.

We use the ResNet-20 architecture (He et al., 2016) for CIFAR-10. It has 269,722 parameters. For consistency with the other experiments, we use neither data augmentation nor weight decay.

G   Experimental details: training details

The bounds are evaluated on the 0–1 loss, which is not differentiable. To enable gradient-based optimization, we replace this with the cross entropy loss divided by the log number of classes, which gives a tight upper bound on the 0–1 loss.

We use SGD with momentum as the optimizer. We use one learning rate for the α-*prefix* and *base* runs and another, lower learning rate for the bound optimization. For experiments on MNIST and Fashion-MNIST, the momentum is 0.95 and the batch size is 256. For MNIST, the learning rate for the α-*prefix* and *base* runs is 0.003 and the learning rate for bound optimization is 0.0003; for Fashion-MNIST, they are respectively 0.01 and 0.003. We sweep over the prior variance $\sigma_P \in \{3 \times 10^{-8}, 1 \times 10^{-7}, 3 \times 10^{-7}, \ldots, 1 \times 10^{-2}\}$. Via a union bound argument, our hyperparameter sweeps contribute a negligible amount to the bounds.

For the best hyperparameter setting, Algorithm 1 (right) was repeated 50 times with different data-orders and

w_0. In all figures any shaded area corresponds to 2 standard deviations around the mean as computed from the 50 runs.

H   More details on direct risk bound minimization

We evaluated the performance of a learning algorithm baed on optimizing a PAC-Bayes bound based on a data-dependent prior. Our proposed algorithm gets nearly state-of-the-art performance and produces a valid and tight PAC-Bayes bound on risk.

Let $Q = \mathcal{N}(w, \Sigma_\alpha)$ and $P = \mathcal{N}(w_\alpha, \Sigma_\alpha)$. The algorithm starts with the coupling and α-*prefix* runs as before. Then the *base* run is replaced with SGD minimizing the PAC-Bayes bounds $\Psi^\dagger_\delta(Q(\theta_Q), P; S \setminus S_\alpha)$ with respect to the posterior mean w. Here $\Psi^\dagger_\delta(Q(\theta_Q), P; S \setminus S_\alpha)$ is the same bound as $\Psi^*_\delta(Q(\theta_Q), P; S \setminus S_\alpha)$ but with risk evaluate on a differentiable surrogate loss. The procedure is outlined in Algorithm 1 (left).

Similarly as before, for each α we choose the learning rate and prior variance that yield the tightest bounds. For a fixed set of hyperparameters, we repeat the optimization 50 times.

The results on 4 different networks and 3 different datasets appear in Fig. 5. The risk bounds and test errors drop dramatically with α up to $\alpha \approx 0.9$ for all the networks tested. For MNIST and Fashion-MNIST, the momentum and batch size is the same as above. For CIFAR-10, the momentum is 0.9 and the batch size is 128. For MNIST and Fashion-MNIST, the α-*prefix* run learning rate is 0.01; for CIFAR-10 it is 0.03. For all datasets, we sweep the direct bound optimization learning rate over $\{1 \times 10^{-6}, 3 \times 10^{-6}, 1 \times 10^{-5}, \ldots, 3 \times 10^{-3}\}$ and the prior variance over $\{1 \times 10^{-9}, 3 \times 10^{-9}, 1 \times 10^{-8}, \ldots, 3 \times 10^{-3}\}$.

H.1   Comparison to PAC-Bayes by Backprop

When $\alpha = 0$, the setting of our direct bound optimization experiments aligns closely to that considered by Rivasplata, Tankasali, and Szepesvari (2019): evaluating a PAC-Bayes bound-based learning algorithm using a prior centered at random initialization. This work reports a test error of 0.014 and a risk bound of 0.023 on MNIST with a 784–600–600–10 fully-connected network architecture, a Gaussian prior, and a PAC-Bayes bound expression similar to ours. Despite correspondence with the authors, we were unable to reproduce these results. For direct comparison, our $\alpha = 0$ baseline results with the same network architecture are a mean test error of 0.116 and a mean risk bound of 0.303 over 10 random seeds. Using a data-dependent prior learnt with proportion $\alpha = 0.7$ of the training data, this improves to a mean test error of 0.022 and a mean risk bound of 0.031 over 10 random seeds.

I   Optimal prior variance

Our data-dependent priors do not attempt to minimize the variance component of the KL bound. For a fixed Σ_P, the variance component in Eq. (33) (see Appendix C) increases if posterior variance Σ deviates from Σ_P. When the prior is isotropic, our empirical study shows that the optimized posterior variance is also close to isotropic. However, an isotropic structure may not describe the local minima found by SGD well. We are thus also interested in a hypothetical experiment, where we allow the prior variance to be optimal for any given diagonal Gaussian Q. While this produces an invalid bound, it reveals the contribution to the risk bound due to the prior variance. Optimizing Eq. (33) w.r.t. diagonal Σ_P yields a prior P^Σ_α with optimal variance, and $\mathrm{KL}(Q||P^\Sigma_\alpha)$ expression reduces to

$$\tfrac{1}{2} \sum_{i=1}^p \log(1 + (w^i - w^i_\alpha)/\sigma^2_i), \tag{52}$$

where σ^2_i is the ith component of the diagonal of Σ.

Computing these hypothetical bounds with P^Σ_α as a prior requires some minor modifications to Algorithm 1 (right). As in Algorithm 1 (right), the posterior is set to $Q = \mathcal{N}(w_S, \Sigma)$, with a diagonal covariance matrix Σ that is initialized to $\sigma^2_P I_p$. The prior P is centered at w^0_α, and the variance is automatically determined by the posterior variance. The KL then takes the form stated in Eq. (52). The α-*prefix* run in Algorithm 1 (right) is followed by another SGD run minimizing $\Psi^*_\delta(Q(\theta_Q), P; S \setminus S_\alpha)$ with respect to diagonal covariance Σ.

The results with the optimal prior covariance can be found in Fig. 7. At $\alpha = 0$, the optimal prior variance decreases the bound substantially. However, at larger values of α, the effect diminishes. In particular, at the

values of α that produce the lowest risk bound with a fixed isotropic prior variance, optimal prior variance makes little to no improvement. Interestingly, the optimized posterior variance remains close to isotropic.

J   Frequently Asked Questions

Data-dependent priors are already a known heuristic for obtaining tight bounds, right?   Ambroladze, Parrado-Hernández, and Shawe-Taylor (2007) proposed to use data-dependent priors and observed that they led to much tighter bounds than those produced by generic priors. However, these bounds were motivated by trying to approximate distribution-dependent oracle priors (producing so-called localized PAC-Bayes bounds (Catoni, 2007; Langford and Blum, 2003)). Indeed, in their work, the data-dependent prior is based on an estimate of the data-independent oracle prior. The bound contains a penalty for the error in this estimate, and so it will not be tighter in expectation.

In contrast, in this work, we observe that data-dependent priors can be superior to distribution-dependent oracle priors. This is not folklore. The effect we are observing is due to the fact that not all data are created equal. As demonstrated by Proposition 3.1 or Example 3.2/Lemma 3.3, using data that have a particularly strong a priori dependence on the posterior can dramatically tighten the risk bound. In our example, the initial data have greater dependence because of the decreasing step size. In the end, data-dependence is the difference between a vacuous bound (based on the distribution-dependent oracle prior) and a nonvacuous bound (based on a data-dependent oracle prior).

How does using a data-dependent prior compare to using held-out data in a, e.g., Chernoff bound?   There is a critical difference: in the former, the posterior $Q(S)$ depends on the entire sample, S. In the latter, the held-out data would not be used by the posterior. The held-out bound *cannot* be used to explain why generalization is occurring. It simply reports that generalization has occurred. In contrast, bounds based on data-dependent priors lend themselves to arguments in terms of distribution-dependent stability. We don't pursue the interpretation of these bounds here, though we discuss related issues in Section 4.

What's the relationship between direct risk bound optimization (last subsection of Section 6) and the goal of explaining SGD?   This final subsection is not directed towards understanding SGD. It presents a novel learning algorithm, though the idea of minimizing a PAC-Bayes bound is a standard one. The use of a data-dependent prior and in particular this one based on a run of SGD on an initial segment of data is new.

The paper studies minimizing high-probability PAC-Bayes bounds in expectation. Shouldn't we be using bounds on the expected generalization error?   Both approaches are sensible. Note that a PAC-Bayes bound controls the generalization error in terms of the KL divergence between the posterior and the prior. (Other types of PAC-Bayes bounds exist, but we will focus on this standard setting here.) The posterior is data dependent and so the KL divergence is a random variable, in general. If we want to develop a tight bound, we want to minimize the KL divergence term, but since it is a random variable, there's no unique way to minimize it. In this work, we minimize the contribution of the (random) KL divergence by minimizing its expectation. Since the PAC-Bayes bound holds with high probability, it would also be interesting to minimize a tail bound on the KL divergence. We do not pursue that here, but it is interesting future work.

What is the meaning of (i) $\Psi_{\beta,\delta}(Q(S), P_{\mathcal{F}}^*(S_J); S \setminus S_J)$ and (ii) $\Psi_{\beta,\delta}(Q(S), P_{\mathcal{F}}^*; S)$ in Proposition 3.1?   From Theorem 2.1, we see $\Psi_{\beta,\delta}(Q, P; S)$ is the linear PAC-Bayes bound on the risk of a posterior $Q(S)$, based on the estimate $L_S(Q(S))$, using the prior P. The terms above are, therefore,

(i) the linear PAC-Bayes bound on the risk of $Q(S)$, based on the estimate $L_{S \setminus S_J}(Q(S))$, using the (data-dependent) prior $P_{\mathcal{F}}^*(S_J)$; and

(ii) the linear PAC-Bayes bound on the risk of $Q(S)$, based on the estimate $L_S(Q(S))$, using the (data-independent) prior $P_{\mathcal{F}}^*$.

Therefore, the theorem is telling us when a linear PAC-Bayes bound can be improved using a data dependent prior. The Ψ term in Lemma 3.3 can be interpreted in the same way.

There are a wide variety of PAC-Bayes bounds. Why do you use linear PAC-Bayes bounds and do your findings generalize to other types of PAC-Bayes bounds? Our focus on linear PAC-Bayes bounds allows us to simplify the analysis considerably. Indeed, by the linearity of expectation, the expected value of the linear PAC-Bayes bound depends on the expected value of the KL term, which then gives us the connection to mutual information.

Regarding other styles of PAC-Bayes bounds, by Jensen's inequality, priors that minimize the expected value of the KL terms will lead to upper bounds on the classic sqrt-style PAC-Bayes bound. Minimizing an upper bound provides a weaker but still valid approach to controlling such bounds. There is also a connection between linear PAC-Bayes bounds and nonlinear PAC-Bayes bounds, such as that derived in Appendix D.1. In particular, from the logic that leads to Eq. (16), we see that optimizing the value of β leads to a bound of the same form as Eq. (42). What we can glean from this correspondence is that the argument in Lemma 3.3 is choosing the value of β that is optimal in expectation.

Note that β must be chosen independently of the data used in the estimate of the risk. However, one can consider a discrete range of beta values, derive a data-dependent prior (and then linear PAC-Bayes bound) for each such value, and then combine these bounds (i.e., taking the tightest one) using a union bound argument. The result is a nonlinear bound, though the final bound is no longer being minimized in expectation. To optimize a nonlinear bound directly, the easiest approach may be to control the tails of the KL term. We think this is an interesting avenue for future work.

Do these ideas extend to PAC-Bayesian bounds based on divergences other than KL? Alquier and Guedj (2018) and Bégin et al. (2016) describe PAC-Bayes bounds based on f-divergences and Rényi α-divergences, respectively. Focusing on the latter for brevity, we think the extension to Rényi divergence based on Begin et al. may be straightforward, *once* you identify a choice for Begin et al.'s Δ function so that the resulting bound is linear: notions of information rate gain would be defined in terms of a Rényi version of mutual information and then Proposition 3.1 would likely carry through. The implications of this change though seem far less clear to us and worthy of study! Clearly Example 3.2 works for $\alpha = 1$, as that's just KL. It would be interesting to understand for which values of α the example still works. We think the critical aspect would be conveying intuition (perhaps through empirical work) for key values of α. We think this is an interesting direction for future work. We don't want to suggest in this paper that it is straightforward or incremental because, until one works out all the details, it might turn out to be a really important variation.

Which is more important/fundamental: Proposition 3.1 or Example 3.2/Lemma 3.3? We think these results enhances each other, and that they are both important.

When we showed colleagues Lemma 3.3 alone, they were then eager to see a general characterization of when data-dependence led to tighter bounds. Proposition 3.1 provides necessary and sufficient conditions. As is the case with necessary and sufficient conditions, the result simply presents a different, but equivalent, perspective. In this case, it shows how the superiority of a data-dependent prior comes from the relative values of the "information rate gain" and the "excess bias". While Proposition 3.1 is immediate from definitions, we believe it provides guidance as to how to choose J. In particular, Example 3.2 and Lemma 3.3 demonstrate that one may arrive at much tighter bounds by identifying samples J that have a priori high dependence on the posterior.

How should one choose the subset, J, of data used to build the data-dependent prior? We believe that Proposition 3.1 provides guidance: we need the information rate gain to exceed (some multiple of) the excess bias plus a variance term.

The variance term will be quite small unless the number of data used in the prior, m, is quite large. Focusing then on the information rate gain, we will maximize this term if we build our prior using samples that exhibit strong a priori dependence with the final posterior. This will produce a tighter bound, provided that the samples that we leave to use in the risk estimate are not too biased.

In Example 3.2, there is strong dependence with the initial data because the step size is largest at the start. At the same time, the learned weights still produce good predictions for all the data. The excess bias term captures the effect of removing this data from the risk estimate, which is sufficiently small in this case.

Is your goal to improve the generalization of SGD? Our goal is *not* to improve the generalization of the SGD algorithm, but to improve our ability to simultaneously (i) choose posteriors Q that closely approximate the performance of SGD and (ii) derive tight generalization bounds for these posteriors. This hope is to shed light on SGD itself, eventually. As the paper argues, the roadblock here is (ii)—in particular, the priors in the KL terms in PAC-Bayes bounds are too far from the posteriors to yield numerically tight bounds. The final two paragraphs of the paper describe a learning algorithm, but this is simply a short aside, and not the main focus of the paper.

When using ghost samples for the prior, shouldn't one compare to the same algorithm that also has access to the ghost samples? No. We would agree with this sentiment if we were designing a learning algorithm or model selection method. However, we are instead probing the generalization properties of posteriors concentrated around the weights learned by SGD. Our experiment reveals that the advantage of distribution-dependence (provided via ghost samples) may be dwarfed by the advantage of data-dependence. Note that the posterior $Q(S)$ does not depend on either the ghost samples or prior.

Section 3 rigorously identifies a setting where data-dependence is required for nonvacuous (linear PAC-Bayes) bounds. In Section 4, Fig. 3, we turn to the question of teasing apart how much data- and distribution-dependence helps. Ghost samples provide distribution-dependence. Section 4.1 shows how to use ghost samples to estimate the optimal Gaussian prior mean (for a fixed variance). The curve with ghost data is therefore an estimate of the *actual* tradeoff in α. The curve without ghost data is indeed heuristic. The fact that this heuristic no-ghost-data curve nearly matches highlights a convenient empirical fact: we can maybe ignore ghost data for this class of Gaussian priors.

You use an isotropic Gaussian prior. Isn't this quite limited in the type of dependence it can capture? We agree that isotropic Gaussians are limited. For this reason, we actually went ahead and evaluated using diagonal Gaussian priors. Unlike isotropic Gaussian priors, which only have one parameter than needs tuning, diagonal Gaussians have too many to optimize naively using a union bound argument, without obtaining a vacuous bound. What are clever ways of choosing a prior's diagonal covariance in a data- and distribution-dependent way? We actually worked on this for a while before we realized that we had better check that it was going to be worth all the effort! And so, in Appendix I of the supplement, we evaluate an *oracle* diagonal Gaussian prior bound where we allow ourselves to optimize the diagonal variance *a posteriori*. When the posterior is also a Gaussian with a diagonal covariance, one can write down a closed-form expression for the optimal diagonal prior variance. This is totally invalid as a bound, but it does provide a hard limit on the best we could hope to achieve through some clever valid approach with diagonal Gaussian priors and posteriors. Our results suggest that the gains are not negligible, but they are also not massive. We conclude that one may need to account for dependence between parameters to see big gains.

How do the approaches you describe for neural networks scale to ImageNet? Different architectures? These are good questions and they deserve further study. Regarding the prospects of evaluating these techniques on ImageNet. We are not quite at the point in terms of the software engineering to scale our experiments to ImageNet. We suspect that it may be even more important to use nondiagonal Gaussians for ImageNet: the only nonvacuous bounds on ImageNet in the literature are ones for a *compressed* network (Zhou et al., 2019) and this compression scheme does not treat weights independently as a diagonal Gaussian prior does implicitly. Even so, their bounds are only slightly better than the random guessing rate! As for understanding the role of architecture, we believe there

Are your empirical results in keeping with earlier work on data-dependent priors? In short, earlier work also finds meaningful gains from using data-dependent priors. Parrado-Hernández et al. (2012) demonstrates a clear advantage of data-dependence for linear predictors when compared to uninformative (nonoracle) priors (Fig. 1). Note, however, that some approaches based on using all the data to choose the prior often yield bounds that are strictly weaker than data-independent oracle bounds because they are essentially oracle bounds with penalties for data-dependence. Our theory suggests that the effect of using only some of the data can be to actually outperform distribution-dependent oracle bounds. Dziugaite and Roy (2018) use differential privacy to select a prior using data and obtain distribution-dependence. They study a two-stage learning process in neural networks where the prior is centered at weights learned using SGLD in an approximately private way and then

the posterior is optimized based on this prior. There is a similar U-shaped tradeoff seen between optimizing the prior and the tightness of the bound.

Could these techniques be useful for continual learning? In continual learning, it's not uncommon to use posteriors from earlier on as priors for later data/tasks. This bears some resemblance to our methodology here for data-dependent oracle priors. Note, however, that the effect we are exploiting is when certain training data have outsized influence on the learned predictor. It's useful to condition on this overly influential training data because the effect it has on the learned predictor is atypical / hard to predict. In a continual or meta-learning setting, we might expect some relationship between tasks. In this case, we're hoping to learn typical properties of upcoming tasks from prior tasks, and so the role of data-dependence is quite different. Of course, we *also* want our PAC-Bayes priors to be tuned to the data distribution. Oracle priors capture all the relevant distribution dependence. Even so, they miss out on the atypical parts that are training-data specific, which is why data-dependent oracle priors can be superior. If obtaining generalization bounds for continual and meta-learning problems also suffer from the hard to predict influence of data, similar techniques may be applicable.
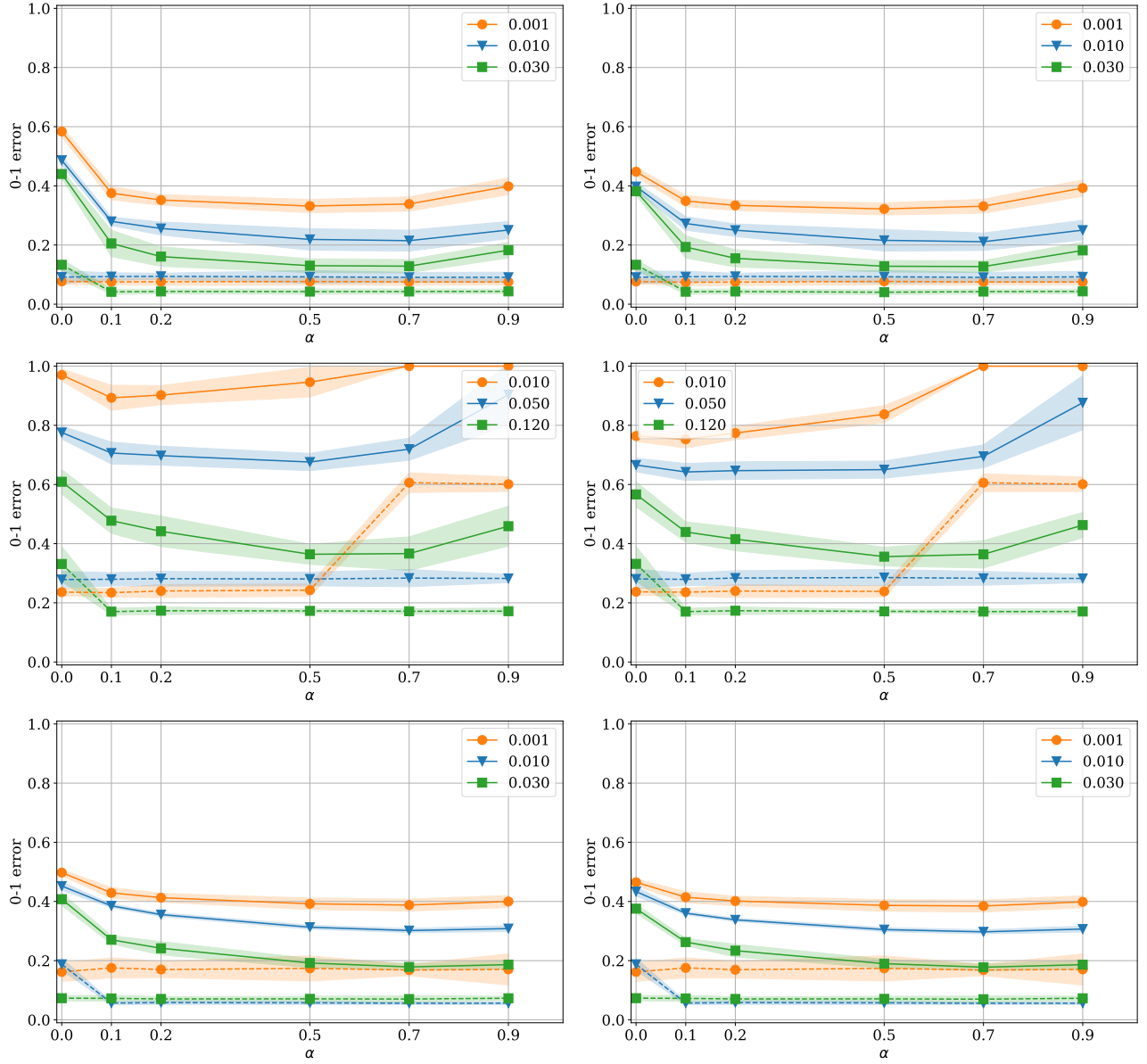
Figure 7: **Top row**: MNIST, LeNet-5; **middle row**: Fashion-MNIST, LeNet-5; **bottom row**: MNIST, FC; **y-axis**: error-rate; **x-axis**: fraction α of the data used by the α-prefix run of SGD to predict the weights produced by the *base* run of SGD; **dashed lines**: test error; **solid lines**: bound on the error of a Gaussian Gibbs classifier whose mean is the weights learned by the *base* run of SGD and whose covariance has been optimized to minimize a PAC-Bayes risk bound; **legend**: training error that was used as the stopping criterion for the *base* run; **left column:** test error and PAC-Bayes generalization bounds with isotropic prior covariance; **right column:** hypothetical bounds with diagonal prior variance set to optimal. The improvement is seen only for low α values. At higher α values, the bounds are similar to the ones obtained with isotropic prior variance. The best test error bound on MNIST (approximately 11%) is significantly better than the 46% bound by Zhou et al. (2019).
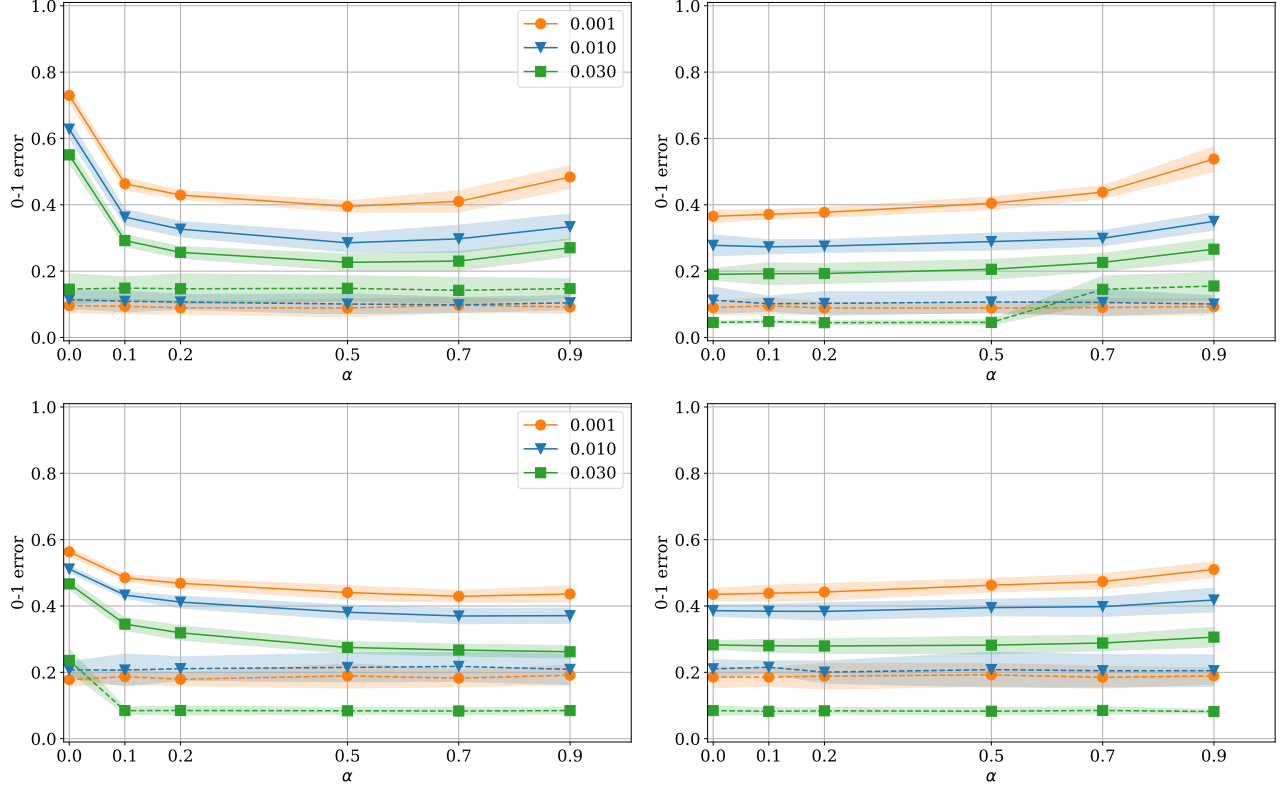
Figure 8: **Top row**: MNIST, LeNet-5; **bottom row**: MNIST, FC; **y-axis**: error-rate; **x-axis**: fraction α of the data used by the α-prefix run of SGD to predict the weights produced by the *base* run of SGD; **left column**: test error and PAC-Bayes error bounds with isotropic prior covariance using half of MNIST data; **right column**: data and oracle prior bounds, where the prior is an isotropic Gaussian. The oracle prior is approximated by using ghost samples. When using ghost samples, some improvement on the bounds is seen for small values of α (below 0.2). For large values of α (at around 0.9) and some stopping times, the bound with a data and oracle dependent prior is worse than with a data-dependent prior.