Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties of Entropy-SGD and data-dependent priors

Gintare Karolina Dziugaite [1 2] **Daniel M. Roy** [3 2]

Abstract

We show that Entropy-SGD (Chaudhari et al., 2017), when viewed as a learning algorithm, optimizes a PAC-Bayes bound on the risk of a Gibbs (posterior) classifier, i.e., a randomized classifier obtained by a risk-sensitive perturbation of the weights of a learned classifier. Entropy-SGD works by optimizing the bound's prior, violating the hypothesis of the PAC-Bayes theorem that the prior is chosen independently of the data. Indeed, available implementations of Entropy-SGD rapidly obtain zero training error on random labels and the same holds of the Gibbs posterior. In order to obtain a valid generalization bound, we rely on a result showing that data-dependent priors obtained by stochastic gradient Langevin dynamics (SGLD) yield valid PAC-Bayes bounds provided the target distribution of SGLD is ε-differentially private. We observe that test error on MNIST and CIFAR10 falls within the (empirically nonvacuous) risk bounds computed under the assumption that SGLD reaches stationarity. In particular, Entropy-SGLD can be configured to yield relatively tight generalization bounds and still fit real labels, although these same settings do not obtain state-of-the-art performance.

1. Introduction

Optimization is central to much of machine learning, but generalization is the ultimate goal. Despite this, the generalization properties of many optimization-based learning algorithms are poorly understood. The standard example is stochastic gradient descent (SGD), one of the workhorses of deep learning, which has good generalization performance in many settings, even under overparametrization (Neyshabur et al., 2014), but rapidly overfits in others (Zhang et al., 2017). Can we develop high performance learning algorithms with provably strong generalization guarantees? Or is there a limit?

In this work, we study an optimization algorithm called Entropy-SGD (Chaudhari et al., 2017), which was designed to outperform SGD in terms of generalization error when optimizing an empirical risk. Entropy-SGD minimizes an objective $f : \mathbb{R}^p \to \mathbb{R}$ indirectly by approximating stochastic gradient ascent on the so-called local entropy

$$F(\mathbf{w}) \stackrel{\text{def}}{=} C(\tau) + \log \underbrace{\mathbb{E}_{\xi\sim\mathscr{N}_\tau}[e^{-\tau f(\mathbf{w}+\xi)}]}_{\int \exp(-f(\mathbf{w}+x))\mathscr{N}_\tau(\mathrm{d}x)}$$

where $\tau > 0$ is an inverse temperature, $C(\tau)$ is an additive constant, and $\mathscr{N}_\tau$ denotes a zero-mean isotropic multivariate normal distribution on $\mathbb{R}^p$ whose scale depends on τ.

Our first contribution is connecting Entropy-SGD to results in statistical learning theory, showing that maximizing the local entropy corresponds to minimizing a PAC-Bayes bound (McAllester, 1999) on the risk of the so-called Gibbs posterior. The distribution of $\mathbf{w} + \xi$ is the PAC-Bayesian "prior", and so optimizing the local entropy optimizes the bound's prior. This connection between local entropy and PAC-Bayes follows from a result due to Catoni (2007, Lem. 1.1.3) in the case of bounded risk. (See Theorem 3.1.) In the special case where τf is the empirical cross entropy, the local entropy is literally a Bayesian log marginal density. The connection between minimizing PAC-Bayes bounds under log loss and maximizing log marginal densities is the subject of recent work by Germain et al. (2016). Similar connections have been made by Zhang (2006a;b); Grünwald (2012); Grünwald & Mehta (2016).

Despite the connection to PAC-Bayes, as well as theoretical results by Chaudhari et al. suggesting that Entropy-SGD may be more stable than SGD, we demonstrate that Entropy-SGD (and its corresponding Gibbs posterior) can rapidly overfit, just like SGD. We identify two changes, motivated by theoretical analysis, that prevent overfitting.

[1]Dept. of Engineering, Univ. of Cambridge, Cambridge, UK [2]Vector Institute, Toronto, Canada [3]Dept. of Statistical Sciences, Univ. of Toronto, Toronto, Canada. Correspondence to: Gintare Karolina Dziugaite <gkd22@cam.ac.uk>, Daniel M. Roy <droy@utstat.toronto.edu>.

Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

The first change relates to the stability of the optimized prior mean, with respect to changes to the data. The PAC-Bayes theorem requires that the prior be independent of the data, and so by optimizing the prior mean, Entropy-SGD invalidates the bound. Indeed, the bound does not hold empirically. While a PAC-Bayes prior may not be chosen based on the data, it can depend on the data distribution. This suggests that if the prior depends only weakly on the data, it may be possible to derive a valid bound and control overfitting.

Indeed, Dziugaite & Roy (2018) recently formalized this idea using differential privacy (Dwork, 2006; Dwork et al., 2015b) under the assumption of bounded risk. Using existing results connecting statistical validity and differential privacy (Dwork et al., 2015b, Thm. 11), they show that an ε-differentially private prior yields a valid, though looser, PAC-Bayes bound.

Achieving strong differential privacy can be computationally intractable. Motivated by this obstruction, Dziugaite & Roy relax the privacy requirement in the case of Gaussian PAC-Bayes priors parameterized by their mean vector. They show that convergence in distribution to a differentially private mechanism suffices for generalization. This allows one to use stochastic gradient Langevin dynamics (SGLD; Welling & Teh, 2011), which is known to converge weakly to its target distribution, under regularity conditions. We will refer to the Entropy-SGD algorithm as Entropy-SGLD when the SGD step on local entropy is replaced by SGLD.

The one hurdle to using data-dependent priors learned by SGLD is that we cannot easily measure how close we are to converging. Rather than abandoning this approach, we take two steps: First, we run SGLD far beyond the point where it appears to have converged. Second, we assume convergence, but then view/interpret the bounds as being optimistic. In effect, these two steps allow us to see the potential and limitations of using private data-dependent priors to study Entropy-SGLD.

Empirically, we find that the resulting PAC-Bayes bounds are quite tight but still conservative. On MNIST, when the limiting privacy of Entropy-SGLD is tuned to contribute no more than $2\varepsilon^2 \times 100 \approx 0.2\%$ to the generalization error, the test-set error of the learned network is 3–8%, which is roughly 5–10 times higher than state-of-the-art test-set error, which for MNIST is between 0.2-1%.[1]

The second change pertains to the stability of the stochastic gradient estimate made on each iteration of Entropy-SGD. This estimate is made using SGLD. (Hence Entropy-SGD is SGLD within SGD.) Chaudhari et al. make a subtle but critical modification to the noise term in the SGLD update: the noise is divided by a factor that ranges from 10^3 to 10^4. (This factor was ostensibly tuned to produce good empirical results.) Our analysis shows that, as a result of this modification, the Lipschitz constant of the objective function is approximately 10^6–10^8 times larger, and the conclusion that the Entropy-SGD objective is smoother than the original risk surface no longer stands. This change to the noise also negatively impacts the differential privacy of the prior mean. Working backwards from the desire to obtain tight generalization bounds, we are led to divide the SGLD noise by a factor of only $\sqrt[4]{m}$, where m is the number of data points. (For MNIST, $\sqrt[4]{m} \approx 16$.) The resulting bounds are nonvacuous and tighter than those recently published by Dziugaite & Roy (2017), although it must be emphasized that the bounds are optimistic because we assume SGLD has converged. The extent to which it has not converged may inflate the bound.

We begin by introducing sufficient background so that we can make a formal connection between local entropy and PAC-Bayes bounds. We discuss additional related work in Appendix F. We then introduce several existing learning bounds that use differential privacy, including the PAC-Bayes bounds outlined above that use data-dependent priors. In Section 5, we present experiments on MNIST and CIFAR10, which provide evidence for our theoretical analysis. We close with a short discussion.

2. Preliminaries: Supervised learning, Entropy-SGD, and PAC-Bayes

We consider the batch supervised learning setting, where we are given a sample $z_1, \ldots, z_m$ drawn i.i.d. from an unknown probability distribution $\mathscr{D}$ on a space $Z = X \times Y$ of labeled examples. Given a family of classifiers, indexed by weight vectors $\mathbf{w} \in \mathbb{R}^p$, and a bounded loss function $\ell : \mathbb{R}^p \times Z \to \mathbb{R}$. the *risk* and *empirical risk* are

$$R_{\mathscr{D}}(\mathbf{w}) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{z\sim\mathscr{D}}(\ell(\mathbf{w},z)); \quad \hat{R}_S(Q) \stackrel{\text{def}}{=} \tfrac{1}{m}\textstyle\sum_{i=1}^m \mathop{\mathbb{E}}_{\mathbf{w}\sim Q}(\ell(\mathbf{w},z_i)).$$

Our goal is to learn a classifier with small risk, taking advantage of the fact that $R_{\mathscr{D}}(\mathbf{w}) = \mathbb{E}_{S\sim\mathscr{D}^m}[\hat{R}_S(\mathbf{w})]$. We consider randomized (Gibbs) classifiers, formalized as probability distributions $Q \in \mathscr{M}_1(\mathbb{R}^p)$ on the space of weight vectors. The (expected) risk of a randomized classifier is

$$R_{\mathscr{D}}(Q) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{\mathbf{w}\sim Q}(R_{\mathscr{D}}(\mathbf{w})) = \mathop{\mathbb{E}}_{z\sim\mathscr{D}}\Big(\mathop{\mathbb{E}}_{\mathbf{w}\sim Q}(\ell(\mathbf{w},z))\Big). \tag{1}$$

We will sometimes refer to elements of $\mathbb{R}^p$ and $\mathscr{M}_1(\mathbb{R}^p)$ as classifiers and randomized classifiers, respectively.

Our focus is the case of neural network that output probability vectors $p(\mathbf{w},x) = (p(\mathbf{w},x)_1, \ldots, p(\mathbf{w},x)_K)$ over K classes on input x when the weights are $\mathbf{w}$. Zero–one (0–1)

[1] These numbers must be interpreted carefully—the simple fact that the deep-learning tool chain was developed using MNIST likely implies that generalization and test set bounds are biased.

loss is $\ell(\mathbf{w},(x,y))=1$ if $y=\arg\max_k p(\mathbf{w},x)_k$ and 0 otherwise. We also use cross entropy loss as a differentiable surrogate. Cross entropy loss is $\ell(\mathbf{w},(x,y))=-\log p(\mathbf{w},x)_y$. Note that cross entropy loss is merely bounded below. We use a bounded modification (Appendix C.2). We will often refer to the (empirical) 0–1 risk as the (empirical) error.

2.1. Entropy-SGD

Entropy-SGD is a gradient-based learning algorithm proposed by Chaudhari et al. (2017) as an alternative to stochastic gradient descent on the empirical risk surface $\hat{R}_S$. The authors argue that Entropy-SGD has better generalization performance. Part of that argument is a theoretical analysis of the smoothness of the local entropy surface that Entropy-SGD optimizes in place of the empirical risk surface, as well as a uniform stability argument that they admit rests on assumptions that are violated, but to a small degree empirically. As we have mentioned in the introduction, Entropy-SGD's modifications to the noise term in SGLD result in much worse smoothness. We will modify Entropy-SGD in order to stabilize its learning and control overfitting.

Entropy-SGD is stochastic gradient ascent applied to the optimization problem $\arg\max_{\mathbf{w}\in\mathbb{R}^p} F_{\gamma,\tau}(\mathbf{w};S)$, where

$$F_{\gamma,\tau}(\mathbf{w};S)=\log\int_{\mathbb{R}^p}\underbrace{\exp(-\tau\hat{R}_S(\mathbf{w}')-\tau\frac{\gamma}{2}\|\mathbf{w}'-\mathbf{w}\|_2^2)}_{g^{\mathbf{w},S}_{\gamma,\tau}(\mathbf{w}')}\mathrm{d}\mathbf{w}'. \tag{2}$$

The objective $F_{\gamma,\tau}(\cdot;S)$ is known as the *local entropy*, and can be viewed as the log partition function of the unnormalized probability density function $g^{\mathbf{w},S}_{\gamma,\tau}$. (We will denote the corresponding distribution by $G^{\mathbf{w},S}_{\gamma,\tau}$.) Assuming that one can exchange differentiation and integration, it is straightforward to verify that

$$\nabla_{\mathbf{w}}F_{\gamma,\tau}(\mathbf{w};S)=\mathop{\mathbb{E}}_{\mathbf{w}'\sim G^{\mathbf{w},S}_{\gamma,\tau}}(\tau\gamma(\mathbf{w}-\mathbf{w}')), \tag{3}$$

and then the local entropy $F_{\gamma,\tau}(\cdot;S)$ is differentiable, even if the empirical risk $\hat{R}_S$ is not. Indeed, Chaudhari et al. show that the local entropy and its derivative are Lipschitz. Chaudhari et al. argue informally that maximizing the local entropy leads to "flat minima" in the empirical risk surface, which several authors (Hinton & van Camp, 1993; Hochreiter & Schmidhuber, 1997; Baldassi et al., 2015; 2016) have argued is tied to good generalization performance (though none of these papers gives generalization bounds, vacuous or otherwise). Chaudhari et al. propose *approximate* SGD on the local entropy, replacing the gradient $\nabla_{\mathbf{w}}F_{\gamma,\tau}(\mathbf{w};S)$ with a Monte Carlo estimate $\tau\gamma(\mathbf{w}-\mu_L)$, with $\mu_1=\mathbf{w}_1$ and $\mu_{j+1}=\alpha\mathbf{w}'_j+(1-\alpha)\mu_j$, where $\mathbf{w}'_1,\mathbf{w}'_2,\ldots$ are (approximately) i.i.d. samples from $G^{\mathbf{w},S}_{\gamma,\tau}$ and $\alpha\in(0,1)$ defines a weighted average. Obtaining samples from $G^{\mathbf{w},S}_{\gamma,\tau}$ is likely intractable when the dimensionality of the weight vector is large. The authors assume the empirical risk is differentiable and use Stochastic Gradient Langevin Dynamics (SGLD; Welling & Teh, 2011), which simulates a Markov chain whose long-run distribution converges to $G^{\mathbf{w},S}_{\gamma,\tau}$.[2] The final output of Entropy-SGD is the deterministic predictor corresponding to the final weights $\mathbf{w}^*$ achieved by several epochs of optimization.

Algorithm 1 One outerloop step of Entropy-SG(L)D

Input:
$\mathbf{w}\in\mathbb{R}^p$ ▷ Current weights
$S\in Z^m$ ▷ Data
$\ell:\mathbb{R}^p\times Z\to\mathbb{R}$ ▷ Loss
$\tau,\beta,\gamma,\eta,\eta',L,K$ ▷ Parameters
Output: Weights $\mathbf{w}$ moved along stochastic gradient

1: **procedure** ENTROPY-SG(L)D-STEP
2: $\mathbf{w}',\mu\leftarrow\mathbf{w}$
3: **for** $i\in\{1,\ldots,L\}$ **do** ▷ Run SGLD for L iterations.
4: $\eta'_i\leftarrow\eta'/i$
5: $(z_{j_1},\ldots,z_{j_K})\leftarrow$ sample minibatch of size K
6: $\mathrm{d}\mathbf{w}'\leftarrow-\frac{\tau}{K}\sum_{i=1}^K\nabla_{\mathbf{w}'}\ell(\mathbf{w}',z_{j_i})-\gamma\tau(\mathbf{w}'-\mathbf{w})$
7: $\mathbf{w}'\leftarrow\mathbf{w}'+\frac{1}{2}\eta'_i\mathrm{d}\mathbf{w}'+\sqrt{\eta'_i}N(0,I_p)$
8: $\mu\leftarrow(1-\alpha)\mu+\alpha\mathbf{w}'$ ▷ C.f. Eq. (3).
9: $\mathbf{w}\leftarrow\mathbf{w}+\frac{1}{2}\eta\tau\gamma(\mathbf{w}-\mu)+\underbrace{\sqrt{\eta/\beta}N(0,I_p)}_{\text{Entropy-SGLD only}}$
10: **return** $\mathbf{w}$

Algorithm 1 gives a complete description of the stochastic gradient step performed by Entropy-SGD. If we rescale the learning rate, $\eta'\leftarrow\frac{1}{2}\eta'\tau$, lines 6 and 7 are equivalent to

6: $\mathrm{d}\mathbf{w}'\leftarrow-\frac{1}{K}\sum_{i=1}^K\nabla_{\mathbf{w}'}\ell(\mathbf{w}',z_{j_i})-\gamma(\mathbf{w}'-\mathbf{w})$
7: $\mathbf{w}'\leftarrow\mathbf{w}'+\eta'_i\mathrm{d}\mathbf{w}'+\sqrt{\eta'_i}\sqrt{2/\tau}N(0,I_p)$

Notice that the noise term is multiplied by a factor of $\sqrt{2/\tau}$. A multiplicative factor ε—called the "thermal noise", but playing exactly the same role as $\sqrt{2/\tau}$ here—appears in the original description of the Entropy-SGD algorithm given by Chaudhari et al.. However, ε does not appear in the definition of local entropy used in their stability analysis. Our derivations highlight that scaling the noise term in SGLD update has a profound effect: the thermal noise exponentiates the density that defines the local entropy. The smoothness analysis of Entropy-SGD does not take into consideration the role of ε, which is critical because Chaudhari et al. take ε to be as small as 10^{-3} and 10^{-4}. Indeed, the conclusion that the local entropy surface is smoother no longer holds. We will see that τ controls the stability (and then the generalization error) of our variant of Entropy-SGD.

[2] Chaudhari et al. take $L=20$ steps of SLGD, using a constant step size $\eta'_j=0.2$ and weighting $\alpha=0.75$.

2.2. KL divergence and the PAC-Bayes theorem

Let $Q, P \in \mathscr{M}_1(\mathbb{R}^p)$, assume Q is absolutely continuous with respect to P, and write $\frac{\mathrm{d}Q}{\mathrm{d}P} : \mathbb{R}^p \to \mathbb{R}_+ \cup \{\infty\}$ for some Radon–Nikodym derivative of Q with respect to P. Then the Kullback–Liebler divergence from Q to P is

$$\mathrm{KL}(Q||P) \stackrel{\text{def}}{=} \int \log \frac{\mathrm{d}Q}{\mathrm{d}P} \, \mathrm{d}Q.$$

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

$$\mathrm{KL}(q||p) \stackrel{\text{def}}{=} \mathrm{KL}(\mathscr{B}_q||\mathscr{B}_p) = q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p}.$$

We now present a PAC-Bayes theorem. The first such result was established by McAllester (1999). We focus on the setting of bounding the generalization error of a (randomized) classifier on a finite discrete set of labels K. We will use the following variation of a PAC-Bayes bound, where we consider bounded loss functions.

Theorem 2.1 (Linear PAC-Bayes Bound; McAllester 2013; Catoni 2007)**.** *Fix $\lambda > 1/2$ and assume the loss takes values in an interval of length L_{max}. For every $\delta > 0$, $m \in \mathbb{N}$, $\mathscr{D} \in \mathscr{M}_1(\mathbb{R}^k \times K)$, and $P \in \mathscr{M}_1(\mathbb{R}^p)$, with probability at least $1-\delta$, for every $Q \in \mathscr{M}_1(\mathbb{R}^p)$,*

$$R_{\mathscr{D}}(Q) \leq \frac{1}{1-\frac{1}{2\lambda}} \Big(\hat{R}_S(Q) + \frac{\lambda L_{max}}{m} \Big(\mathrm{KL}(Q||P) + \log \frac{1}{\delta} \Big) \Big). \tag{4}$$

The (PAC-Bayes) prior P in the bound is data independent. Later, we introduce bounds for data-dependent priors.

3. Maximizing local entropy minimizes a PAC-Bayes bound

We now present our first contribution, a connection between the local entropy and PAC-Bayes bounds. We begin with some notation for Gibbs distributions. For a measure P on $\mathbb{R}^p$ and function $g : \mathbb{R}^p \to \mathbb{R}$, let $P[g]$ denote the expectation $\int g(h) P(\mathrm{d}h)$ and, provided $P[g] < \infty$, let P_g denote the probability measure on $\mathbb{R}^p$, absolutely continuous with respect to P, with Radon–Nikodym derivative $\frac{\mathrm{d}P_g}{\mathrm{d}P}(h) = \frac{g(h)}{P[g]}$. A distribution of the form $P_{\exp(-\tau g)}$ is generally referred to as a Gibbs distribution. In the special case where P is a probability measure, we call $P_{\exp(-\tau \hat{R}_S)}$ a "Gibbs posterior".

Theorem 3.1 (Maximizing local entropy optimizes a PAC-Bayes bound's prior)**.** *Assume the loss takes values in an interval of length L_{max}, let $\tau = \frac{m}{\lambda L_{max}}$ for some $\lambda > 1/2$, Then the set of weight $\mathbf{w}$ maximizing the local entropy $F_{\gamma,\tau}(\mathbf{w}; S)$ equals the set of weights $\mathbf{w}$ minimizing the right hand side of Eq. (4) for $Q = G_{\gamma,\tau}^{\mathbf{w},S} = P_{\exp(-\tau \hat{R}_S)}$ and P a multivariate normal distribution with mean $\mathbf{w}$ and covariance matrix $(\tau\gamma)^{-1} I_p$.*

See Appendix A for the proof. The theorem requires the loss function to be bounded, because the PAC-Bayes bound we have used applies only to bounded loss functions. Germain et al. (2016) described PAC-Bayes generalization bounds for unbounded loss functions, though it requires that one make additional assumptions about the distribution of the empirical risk, which we would prefer not to make. (See Grünwald & Mehta (2016) for related work on excess risk bounds and further references).

4. Data-dependent PAC-Bayes priors

Theorem 3.1 reveals that Entropy-SGD is optimizing a PAC-Bayes bound with respect to the prior. As a result, the prior P depends on the sample S, and the hypotheses of the PAC-Bayes theorem (Theorem 2.1) are not met. Naively, it would seem that this interpretation of Entropy-SGD cannot explain its ability to generalize. Using tools from differential privacy, Dziugaite & Roy (2018) show that if the prior term is optimized in a differentially private way, then a PAC-Bayes theorem still holds, at the cost of a slightly looser bound. We will assume basic familiarity with differential privacy. (See Dziugaite & Roy (2018) for a basic summary.) We borrow the notation $\mathscr{A} : Z \rightsquigarrow T$ for a (randomized) algorithm with an input in Z and output in T.

The key result is due to Dwork et al. (2015b, Thm. 11).

Theorem 4.1. *Let $m \in \mathbb{N}$, let $\mathscr{A} : Z^m \rightsquigarrow T$, let $\mathscr{D}$ be a distribution over Z, let $\beta \in (0,1)$, and, for each $t \in T$, fix a set $v(t) \subseteq Z^m$ such that $\mathbb{P}_{S \sim \mathscr{D}^m}(S \in v(t)) \leq \beta$. If $\mathscr{A}$ is ε-differentially private for $\varepsilon \leq \sqrt{\ln(1/\beta)/(2m)}$, then $\mathbb{P}_{S \sim \mathscr{D}^m}(S \in v(\mathscr{A}(S))) \leq 3\sqrt{\beta}$.*

Using Theorem 4.1, one can compute tail bounds on the generalization error of fixed classifiers, and then, provided that a classifier is learned from data in a differentially private way, the tail bound holds on the classifier, with less confidence. The following two tail bounds are examples of this idea due to Oneto et al. (2017, Lem. 2 and Lem. 3).

Theorem 4.2. *Let $m \in \mathbb{N}$, let $\mathscr{A} : Z^m \rightsquigarrow \mathbb{R}^p$ be ε-differentially private, and let $\delta > 0$. Then $|R_{\mathscr{D}}(\mathscr{A}(S)) - \hat{R}_S(\mathscr{A}(S))| < \bar{\varepsilon} + m^{-\frac{1}{2}}$ with probability at least $1-\delta$ over $S \sim \mathscr{D}^m$, where $\bar{\varepsilon} = \max\{\varepsilon, \sqrt{\frac{1}{m} \log \frac{3}{\delta}}\}$. The same holds for the upper bound $\sqrt{(6\hat{R}_S(\mathscr{A}(S)))(\bar{\varepsilon} + m^{-\frac{1}{2}})} + 6(\bar{\varepsilon}^2 + m^{-1})$.*

4.1. An ε-differentially private PAC-Bayes bound

The PAC-Bayes theorem allows one to choose the prior based on the data-generating distribution $\mathscr{D}$, but not on the data $S \sim \mathscr{D}^m$. Using differential privacy, one can consider a data-dependent prior $\mathscr{P}(S)$.

Theorem 4.3 (Dziugaite & Roy 2018)**.** *Under 0–1 loss, for every $\delta > 0$, $m \in \mathbb{N}$, $\mathscr{D} \in \mathscr{M}_1(\mathbb{R}^k \times K)$, and ε-differentially private data-dependent prior $\mathscr{P} : Z^m \rightsquigarrow \mathscr{M}_1(\mathbb{R}^p)$, with probability at least $1-\delta$ over $S \sim \mathscr{D}^m$, for every $Q \in$*

$\mathscr{M}_1(\mathbb{R}^p)$,

$$\mathrm{KL}(\hat{R}_S(Q)||R_{\mathscr{D}}(Q)) \leq \frac{\mathrm{KL}(Q||\mathscr{P}(S)) + \ln 2\sqrt{m} + 2\max\{\ln\frac{3}{\delta},\, m\varepsilon^2\}}{m}. \quad (5)$$

Note that the bound holds for any posterior Q, including one obtained by optimizing a *different* PAC-Bayes bound. Inverting $\mathrm{KL}(\hat{R}_S(Q)||R_{\mathscr{D}}(Q))$ allows one to obtain a two-sided confidence interval for $R_{\mathscr{D}}(Q)$. Note that, in realistic scenarios, δ is large enough relative to ε that an ε-differentially private prior $\mathscr{P}(S)$ contributes $2\varepsilon^2$ to the generalization error. Therefore, ε must be much less than one to not contribute a nontrivial amount to the generalization error. As discussed by Dziugaite & Roy, one can match the m^{-1} rate by which the KL term decays choosing $\varepsilon \in O(m^{-1/2})$. Our empirical studies use this rate.

4.2. Differentially private data-dependent priors

We have already explained that the weights learned by Entropy-SGD can be viewed as the mean of a data-dependent prior $\mathscr{P}(S)$. By Theorem 4.3 and the fact that post-processing does not decrease privacy, it would suffice to establish that the mean is ε-differentially private in order to obtain a risk bound on the corresponding Gibbs posterior classifier.

The standard (if idealized) approach for optimizing a data-dependent objective in a private way is to use the exponential mechanism (McSherry & Talwar, 2007). In the context of maximizing the local entropy, the exponential mechanism corresponds to sampling exactly from the "local entropy (Gibbs) distribution" $P_{\exp(\beta F_{\gamma,\tau}(\cdot;S))}$, where $\beta > 0$ and P is some measure on $\mathbb{R}^p$. (It is natural to take P to be Lebesgue measure, or a multivariate normal distribution, which would correspond to L2 regularization of the local entropy.) The following result establishes the privacy of a sample from the local entropy distribution:

Theorem 4.4. *Let $\gamma, \tau > 0$, and assume the range of the loss is contained in an interval of length L_{max}. One sample from the local entropy distribution $P_{\exp(\beta F_{\gamma,\tau}(\cdot;S))}$ is $\frac{2\beta L_{max}\tau}{m}$-differentially private.*

See Appendix B for proof. Sampling from exponential mechanisms exactly is generally intractable. We therefore rely on the following result due to Dziugaite & Roy (2018), which allows us to use SGLD to produce an approximate sample and obtain the same bound up to a term that depends on the degree of convergence. Let $R^{0\text{–}1}_{\mathscr{D}}(\cdot)$ denote risk with respect to 0–1 loss and let $R_{\mathscr{D}}(\cdot)$ denote risk with respect to the bounded version of cross-entropy described by Dziugaite & Roy (2018). (For completeness, the bounded version is defined in Appendix C.2.1.)

Theorem 4.5 (SGLD PAC-Bayes Bound)**.** *Let $\tau > 0$ and $\Sigma \in \mathbb{R}^{p\times p}_{\succeq 0}$. For $\mathbf{w} \in \mathbb{R}^p$ and $S \in Z^m$, let $P_{\mathbf{w}} = \mathscr{N}(\mathbf{w}, \Sigma)$, $Q^S_{\mathbf{w}} = (P_{\mathbf{w}})_{\exp(-\tau\hat{R}_S)}$, and assume $\hat{R}_S(\cdot)$ is bounded. Then, for every $\varepsilon' > 0$ and $\delta, \delta' \in (0,1)$, with probability at least $1-\delta-\delta'$ over $S \sim \mathscr{D}^m$ and a sequence $\mathbf{w}_1, \mathbf{w}_2, \ldots$ (such as produced by SGLD) converging in distribution (conditionally on S) to an ε-differentially private vector $\mathbf{w}^*(S)$, there exists $N \in \mathbb{N}$, such that, for all $n > N$,*

$$\mathrm{KL}(R^{0\text{–}1}_S(Q^S_{\mathbf{w}_n})||R^{0\text{–}1}_{\mathscr{D}}(Q^S_{\mathbf{w}_n})) \leq \frac{\mathrm{KL}(Q^S_{\mathbf{w}_n}||P_{\mathbf{w}_n}) + \ln 2\sqrt{m} + 2\max\{\ln\frac{3}{\delta},\, m\varepsilon^2\}}{m} + \varepsilon'.$$

Remark 4.6. Raginsky et al. (2017) give conditions that suffice to imply that SGLD converges in distribution. Note that the number of required iterations N of SGLD may depend on the sample S, ε, and δ. See (Dziugaite & Roy, 2018) for details and improved bounds. ◁

In summary, we optimize the local entropy $F_{\gamma,\tau}(\cdot;S)$ using SGLD, repeatedly performing the update

$$\mathbf{w} \leftarrow \mathbf{w} + \tfrac{1}{2}\eta\hat{g}(\mathbf{w}) + \sqrt{\eta/\beta}\,N(0, I_p),$$

where at each round $\hat{g}(\mathbf{w})$ is an estimate of the gradient $\nabla_{\mathbf{w}} F_{\gamma,\tau}(\mathbf{w};S)$. (Recall the identity Eq. (3).) As in Entropy-SGD, we construct biased gradient estimates via an inner loop of SGLD. Ignoring error from these biased gradients, we obtain a data-dependent prior that yields a valid PAC-Bayes bound. The only change to Entropy-SGD is the addition of noise in the outer loop. We call the resulting algorithm Entropy-SGLD. (See Algorithm 1.)

As we run SGLD longer, we obtain a tighter bound that holds with probability no less than some value approaching $1-\delta$. In practice we may not know the rate at which this convergence occurs. In our experiments, we use very long runs to approximate near-convergence and then only interpret the bounds as being optimistic. We return to these issues in Sections 5 and 6.

5. Numerical evaluations on MNIST

PAC-Bayes bounds for Entropy-SGLD are data-dependent and so the question of their utility is an empirical one that requires data. In this section, we perform an empirical study of SGD, SGLD, Entropy-SGD, and Entropy-SGLD on the MNIST and CIFAR10 data sets, using both convolutional and fully connected architectures, and comparing several numerical generalization bounds to test errors estimated based on held-out data.

The PAC-Bayes bounds we use depend on the privacy of a sample from the local entropy distribution. (Bounds for SGLD depend on the privacy of a sample from the Gibbs posterior.) For the local entropy distribution, the degree ε of privacy is determined by the product of the τ and β parameters of the local entropy distribution. (Thermal noise

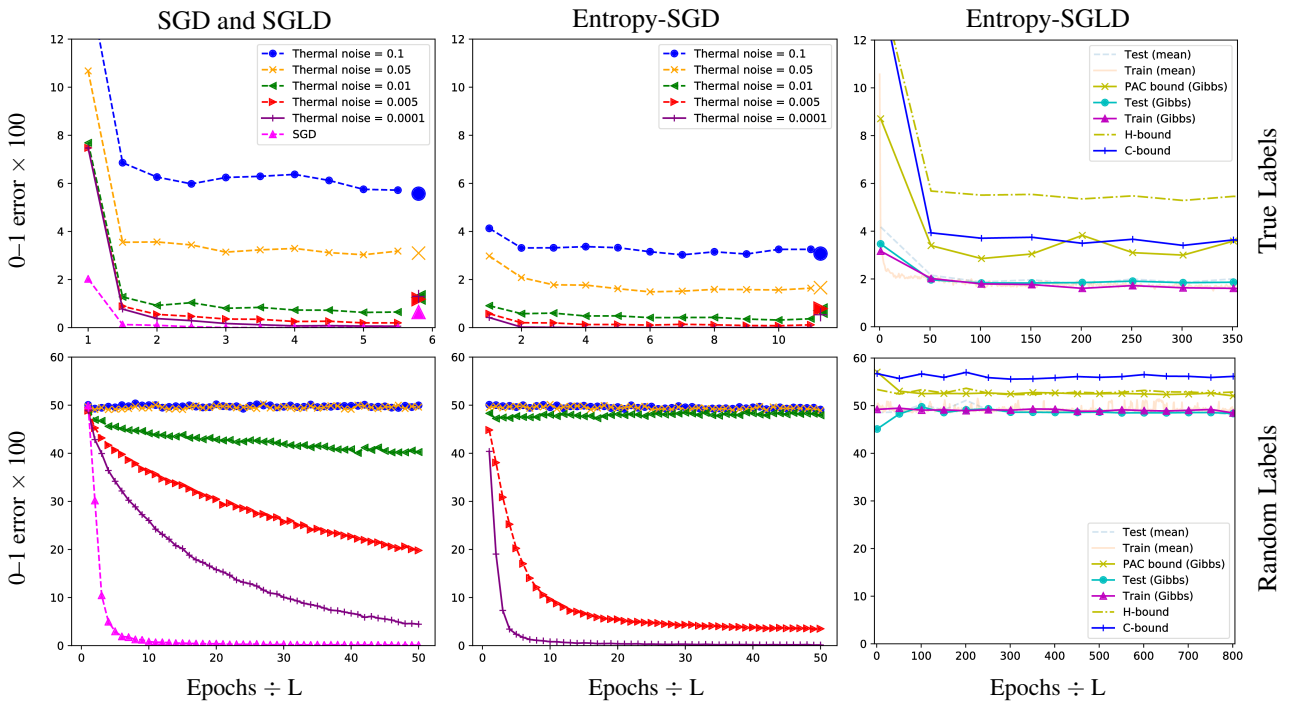

Figure 1: Results on the CONV network on two-class MNIST. **(left column)** Training error (under 0–1 loss) for SGLD on the empirical risk $-\tau\hat{R}_S$ under a variety of thermal noise $\sqrt{2/\tau}$ settings. SGD corresponds to zero thermal noise. **(top-left)** The large markers on the right indicate test error. The gap is an estimate of the generalization error. On true labels, SGLD finds classifiers with relatively small generalization error. At low thermal noise settings, SGLD (and its zero limit, SGD), achieve small empirical risk. As we increase the thermal noise, the empirical 0–1 error increases, but the generalization error decreases. At 0.1 thermal noise, risk is close to 50%. **(bottom-left)** On random labels, SGLD has high generalization error for thermal noise values 0.01 and below. (True error is 50%). **(top-middle)** On true labels, Entropy-SGD, like SGD and SGLD, has small generalization error. For the same settings of thermal noise, empirical risk is lower. **(bottom-middle)** On random labels, Entropy-SGD overfits for thermal noise values 0.005 and below. Thermal noise 0.01 produces good performance on both true and random labels. **(right column)** Entropy-SGLD is configured to approximately sample from an ε-differentially private mechanism with $\varepsilon \approx 0.0327$ by setting $\tau = \sqrt{m}$, where m is the number of training samples. **(top-right)** On true labels, the generalization error for networks learned by Entropy-SGLD is close to zero. Generalization bounds are relatively tight. **(bottom-right)** On random label, Entropy-SGLD does not overfit. See Fig. 3 for SGLD bounds at same privacy setting.

is $\sqrt{2/\tau}$.) In turn, ε increases the generalization bound. For a fixed β, theory predicts that τ affects the degree of overfitting. We see this empirically. No bound we compute is violated more frequently than it is expected to be. The PAC-Bayes bound for SGLD is expanded by an amount ε' that goes to zero as SGLD converges. We assume SGLD has converged and so the bounds we plot are optimistic. We discuss this point below in light of our empirical results, and then return to this point in the discussion.

The weights learned by SGD, SGLD, and Entropy-SGD are treated differently from those learned by Entropy-SGLD. In the former case, the weights parametrize a neural network as usual, and the training and test error are computed using these weights. In the latter case, the weights are taken to be the mean of a multivariate normal prior, and we evaluate the training and test error of the associated Gibbs posterior (i.e., a randomized classifier). We also report the performance of the (deterministic) network parametrized by these weights (the "mean" classifier) in order to give a coarse statistic summarizing the local empirical risk surface.

Following Zhang et al. (2017), we study these algorithms on MNIST with the original ("true") labels, as well as on random labels. Parameter τ that performs very well in one setting often does not perform well in the other. Random labels mimic data where the Bayes error rate is high, and where overfitting can have severe consequences.

5.1. Details

We use a two-class variant of MNIST (LeCun et al., 2010).[3] (Due to space issues, see Appendices D and E for experiments on the standard multiclass MNIST dataset and CIFAR10.) Some experiments involve random labels, i.e., labels drawn independently and uniformly at random at the start of training. We study three network architectures, abbreviated FC600, FC1200, and CONV. Both FC600 and FC1200 are 3-layer fully connected networks, with 600 and 1200 units per hidden layer, respectively. CONV is a convolutional architecture. All three network architectures are taken from the MNIST experiments by Chaudhari et al. (2017), but adapted to our two-class version of MNIST.[4] Let S and S_{tst} denote the training and test sets, respectively. For all learning algorithms we track

(i) $R_S^{0\text{–}1}(\mathbf{w})$ and $R_{S_{\text{tst}}}^{0\text{–}1}(\mathbf{w})$, i.e., the training/test error for $\mathbf{w}$.

We also track

(ii) estimates of $R_S^{0\text{–}1}(G_{\gamma,\tau}^{\mathbf{w},S})$ and $R_{S_{\text{tst}}}^{0\text{–}1}(G_{\gamma,\tau}^{\mathbf{w},S})$, i.e., the mean training and test error of the local Gibbs distribution, viewed as a randomized classifier ("Gibbs")

and, using the bound stated in Theorem 4.4, we compute

(iii) a PAC-Bayes bound on $R_{\mathscr{D}}^{0\text{–}1}(G_{\gamma,\tau}^{\mathbf{w},S})$ using Theorem 4.3 ("PAC-bound");
(iv) the mean of a Hoeffding-style bound on $R_{\mathscr{D}}^{r}(\mathbf{w}')$, where the underlying loss is the ramp loss with slope 10^6 and $\mathbf{w}' \sim P_{\exp(F_{\gamma,\tau}(\cdot;S))}$, using the first bound of Theorem 4.2 ("H-bound");
(v) an upper bound on the mean of a Chernoff-style bound on $R_{\mathscr{D}}^{r}(\mathbf{w}')$, where $\mathbf{w}' \sim P_{\exp(F_{\gamma,\tau}(\cdot;S))}$, using the second bound of Theorem 4.2 ("C-bound").

We also compute H- and C- bounds for SGLD, viewed as a sampler for $\mathbf{w}' \sim P_{\exp(-\tau\hat{R}_S)}$, where P is Lebesgue measure.

In order to get privacy guarantees for SGLD and Entropy-SGLD, we modify the cross entropy loss function to be bounded following Dziugaite & Roy (2018). (See Appendix C.2.1). With the choice of $\beta = 1$ and $\tau = \sqrt{m}$, and the loss function taking values in an interval of length $L_{\max} = 4$, the local entropy distribution is an ε-differentially private mechanism with $\varepsilon \approx 0.0327$. See Appendix C.2 for additional details. Note that, in the calculation of (iii), we do not account for Monte Carlo error in our estimate of $R_S^{0\text{–}1}(\mathbf{w})$. The effect is small, given the large number of iterations of SGLD performed for each point in the plot. Recall that

$$R_{\mathscr{D}}^{0\text{–}1}(G_{\gamma,\tau}^{\mathbf{w},S}) = \mathop{\mathbb{E}}_{\mathbf{w}'\sim G_{\gamma,\tau}^{\mathbf{w},S}}(R_{\mathscr{D}}^{0\text{–}1}(\mathbf{w}')),$$

and so we may interpret the bounds in terms of the performance of a randomized classifier or the mean performance of a randomly chosen classifier.

5.2. Results

Key results for the convolutional architecture (CONV) appear in Fig. 1. Results for FC600 and FC1200 appear in Fig. 2 of Appendix C. (Training the CONV network produces the lowest training/test errors and tightest generalization bounds. Results and bounds for FC600 are nearly identical to those for FC1200, despite FC1200 having three times as many parameters.)

The left column of Fig. 1 presents the performance of SGLD for various levels of thermal noise $\sqrt{2/\tau}$ under both true and random labels. (Assuming SGLD is close to weak convergence, we may also use SGLD to directly perform a private optimization of the empirical risk surface. The level of thermal noise determines the differential privacy of SGLD's stationary distribution and so we expect to see a tradeoff between empirical risk and generalization error. Note that, algorithmically, SGD is SGLD with zero thermal noise.) SGD achieves the smallest training and test error on true labels, but overfits the worst on random labels. In comparison, SGLD's generalization performance improves with higher thermal noise, while its risk performance worsens. At 0.05 thermal noise, SGLD achieves reasonable but relatively large risk but almost zero generalization error on both true and random labels. Other thermal noise settings have either much worse risk or generalization performance.

The middle column of Fig. 1 presents the performance of Entropy-SGD for various levels of thermal noise $\sqrt{2/\tau}$ under both true and random labels. As with SGD, Entropy-SGD's generalization performance improves with higher thermal noise, while its risk performance worsens. At the same levels of thermal noise, Entropy-SGD outperforms the risk and generalization error of SGD. At 0.01 thermal noise, Entropy-SGD achieves good risk and low generalization error on both true and random labels. However, the test-set performance of Entropy-SGD at 0.01 thermal noise is still worse than that of SGD. Whether this difference is due to SGD overfitting to the MNIST test set is unclear and deserves further study.

The right column of Fig. 1 presents the performance of Entropy-SGLD with $\tau = \sqrt{m}$ on true and random labels. (This corresponds to approximately 0.09 thermal noise.) On true labels, both the mean and Gibbs classifier learned by Entropy-SGLD have approximately 2% test error and

[3] The MNIST handwritten digits dataset (LeCun et al., 2010) consists of 60000 training set images and 10000 test set images, labeled 0–9. We transformed MNIST to a two-class (i.e., binary) classification task by mapping digits 0–4 to label 1 and 5–9 to label -1.

[4] We adapt the code provided by Chaudhari et al., with some modifications to the training procedure and straightforward changes necessary for our binary classification task.

essentially zero generalization error, which is less than predicted by the bounds evaluated. The differentially private PAC-Bayes risk bounds are roughly 3%. As expected by the theory, Entropy-SGLD, properly tuned, does not overfit on random labels, even after thousands of epochs.

We find that the PAC-Bayes bounds are generally tighter than the H- and C-bounds. All bounds are nonvacuous, though still loose. The error bounds reported here are tighter than those reported by Dziugaite & Roy (2017). However, *the bounds are optimistic because they do not include the additional term which measure how far SGLD is from its weak limit.* Despite the bounds being optimistically tight, we see almost no violations in the data. (Many violations would undermine our assumption.) While we observe tighter generalization bounds than previously reported, and better test error, we are still far from the performance of SGD. The optimistic picture we get from the bounds suggests we need to develop new approaches. Weaker notions of stability with respect to the training data/privacy may be necessary to achieve further improvement in generalization error and test error.

6. Discussion

Our work reveals that Entropy-SGD can be understood as optimizing a PAC-Bayes generalization bound in terms of the bound's prior. Because the prior must be independent of the data, the bound is invalid, and, indeed, we observe overfitting in our experiments with Entropy-SGD when the thermal noise $\sqrt{2/\tau}$ is set to 0.0001 as suggested by Chaudhari et al. for MNIST.

PAC-Bayes priors can, however, depend on the data distribution. This flexibility seems wasted, since the data sample is typically viewed as one's only view onto the data distribution. However, using results combining differential privacy and PAC-Bayes bounds, we arrive at an algorithm, Entropy-SGLD, that minimizes its own PAC-Bayes bound (though for a surrogate risk). Entropy-SGLD performs an approximately private computation on the data, extracting information about the underlying distribution, without undermining the statistical validity of its PAC-Bayes bound. The cost of using the data is a looser bound, but the gains in choosing a better prior make up for the loss. (The gains come from the KL term being much smaller on the account of the prior being better matched to the data-dependent posterior.)

Our bounds based on Theorem 4.5 are optimistic because we do not include the ε' term, assuming that SGLD has essentially converged. We do not find overt evidence that our approximation is grossly violated, which would be the case if we saw the test error repeatedly falling outside our confidence intervals. We believe that it is useful to view the bounds we obtain for Entropy-SGLD as being optimistic and representing the bounds we might be able to achieve rigorously should there be a major advance in private optimization. (No analysis of the privacy of SGLD takes advantage of the fact that it mixes weakly, in part because it's difficult to characterize how much it has converged in any real-world setting after a finite number of steps.) On the account of using private data-dependent priors (and making optimistic assumptions), the bounds we observe for Entropy-SGLD are significantly tighter than those reported by Dziugaite & Roy (2017). However, despite our bounds potentially being optimistic, the test set error we are able to achieve is still 5–10 times worse than that of SGD. Differential privacy may be too conservative for our purposes, leading us to underfit. We are able to achieve good generalization on both true and random labels under 0.01 thermal noise, despite this value of noise being too large for tight bounds. Identifying the appropriate notion of privacy/stability to combine with PAC-Bayes bounds is an important problem.

Despite Entropy-SGLD having much stronger generalization guarantees, Entropy-SGLD learns much more slowly than Entropy-SGD, the test error of Entropy-SGLD is far from state of the art, and the PAC-Bayes bounds, while much tighter than existing bounds, are still quite loose. It seems possible that we may be facing a fundamental tradeoff between the speed of learning, the excess risk, and the ability to produce a certificate of one's generalization error via a rigorous bound. Characterizing the relationship between these quantities is an important open problem.

Acknowledgments This research was carried out in part while the authors were visiting the Simons Institute for the Theory of Computing at UC Berkeley. The authors would like to thank Pratik Chaudhari, Pascal Germain, David McAllester, and Stefano Soatto for helpful discussions. GKD is supported by an EPSRC studentship. DMR is supported by an NSERC Discovery Grant, Connaught Award, Ontario Early Researcher Award, and U.S. Air Force Office of Scientific Research grant #FA9550-15-1-0074.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pp. 308–318, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4139-4. doi: 10.1145/2976749.2978318.

Achille, A. and Soatto, S. On the emergence of invariance and disentangling in deep representations, 2017.

Achille, A. and Soatto, S. Information dropout: Learning optimal representations through noisy computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. First appeared as https://arxiv.org/abs/1611.01353.

Baldassi, C., Ingrosso, A., Lucibello, C., Saglietti, L., and Zecchina, R. Subdominant dense clusters allow for simple learning and high computational performance in neural networks with discrete synapses. *Phys. Rev. Lett.*, 115:128101, Sep 2015. doi: 10.1103/PhysRevLett.115.128101. URL http://link.aps.org/doi/10.1103/PhysRevLett.115.128101.

Baldassi, C., Borgs, C., Chayes, J. T., Ingrosso, A., Lucibello, C., Saglietti, L., and Zecchina, R. Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes. *Proceedings of the National Academy of Sciences*, 113(48):E7655–E7662, 2016. doi: 10.1073/pnas.1608103113. URL http://www.pnas.org/content/113/48/E7655.abstract.

Bassily, R., Smith, A., and Thakurta, A. Differentially private empirical risk minimization: Efficient algorithms and tight error bounds, 2014.

Bassily, R., Nissim, K., Smith, A., Steinke, T., Stemmer, U., and Ullman, J. Algorithmic stability for adaptive data analysis. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1046–1059. ACM, 2016.

Catoni, O. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Lecture Notes-Monograph Series. Institute of Mathematical Statistics, 2007. doi: 10.1214/074921707000000391.

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. Entropy-sgd: Biasing gradient descent into wide valleys. In *International Conference on Learning Representations (ICLR)*, 2017.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.

Dimitrakakis, C., Nelson, B., Mitrokotsa, A., and Rubinstein, B. I. Robust and private bayesian inference. In *International Conference on Algorithmic Learning Theory*, pp. 291–305. Springer, 2014.

Dwork, C. Differential privacy. In Bugliesi, M., Preneel, B., Sassone, V., and Wegener, I. (eds.), *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10–14, 2006, Proceedings, Part II*, pp. 1–12, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. doi: 10.1007/11787006_1. URL http://dx.doi.org/10.1007/11787006_1.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, pp. 2350–2358, 2015a.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. L. Preserving statistical validity in adaptive data analysis. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pp. 117–126. ACM, 2015b.

Dziugaite, G. K. and Roy, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proc. 33rd Conf. Uncertainty in Artificial Intelligence (UAI)*, 2017.

Dziugaite, G. K. and Roy, D. M. Data-dependent pac-bayes priors via differential privacy, 2018.

Germain, P., Bach, F., Lacoste, A., and Lacoste-Julien, S. PAC-bayesian Theory Meets Bayesian Inference. In *Advances in Neural Information Processing Systems*, pp. 1884–1892, 2016.

Grünwald, P. The safe bayesian-learning the learning rate via the mixability gap. In *ALT*, pp. 169–183. Springer, 2012.

Grünwald, P. D. and Mehta, N. A. Fast rates with unbounded losses, 2016.

Hinton, G. E. and van Camp, D. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, COLT '93, pp. 5–13, New York, NY, USA, 1993. ACM. doi: 10.1145/168304.168306. URL http://doi.acm.org/10.1145/168304.168306.

Hochreiter, S. and Schmidhuber, J. Flat minima. *Neural Comput.*, 9(1):1–42, January 1997. doi: 10.1162/neco.1997.9.1.1. URL http://dx.doi.org/10.1162/neco.1997.9.1.1.

Kifer, D., Smith, A., and Thakurta, A. Private convex empirical risk minimization and high-dimensional regression. *Journal of Machine Learning Research*, 1(41):1–40, 2012.

Kingma, D. P., Salimans, T., and Welling, M. Variational dropout and the local reparameterization trick. In Cortes,

C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28*, pp. 2575–2583. 2015.

Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images, 2009. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Langford, J. *Quantitatively tight sample complexity bounds*. PhD thesis, Carnegie Mellon University, 2002.

Langford, J. and Caruana, R. (not) bounding the true error. In Dietterich, T. G., Becker, S., and Ghahramani, Z. (eds.), *Advances in Neural Information Processing Systems 14*, pp. 809–816. MIT Press, 2002.

LeCun, Y., Cortes, C., and Burges, C. J. C. MNIST handwritten digit database. http://yann.lecun.com/exdb/mnist/, 2010. URL http://yann.lecun.com/exdb/mnist.

London, B. A pac-bayesian analysis of randomized learning with applications to stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pp. 2935–2944, 2017.

McAllester, D. A. Pac-bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, COLT '99, pp. 164–170, New York, NY, USA, 1999. ACM. doi: 10.1145/307400.307435. URL http://doi.acm.org/10.1145/307400.307435.

McAllester, D. A. A PAC-Bayesian Tutorial with a Dropout Bound, 2013.

McSherry, F. and Talwar, K. Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '07, pp. 94–103, Washington, DC, USA, 2007. IEEE Computer Society. ISBN 0-7695-3010-9. doi: 10.1109/FOCS.2007.41. URL http://dx.doi.org/10.1109/FOCS.2007.41.

Minami, K., Arai, H., Sato, I., and Nakagawa, H. Differential privacy without sensitivity. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 956–964, 2016.

Mir, D. J. *Differential privacy: an exploration of the privacy-utility landscape*. PhD thesis, Rutgers University, 2013.

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: On the role of implicit regularization in deep learning, 2014. Workshop track poster at ICLR 2015.

Oneto, L., Ridella, S., and Anguita, D. Differential privacy and generalization: Sharper bounds with applications. *Pattern Recognition Letters*, 89:31–38, 2017. ISSN 0167-8655. doi: 10.1016/j.patrec.2017.02.006.

Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Proc. Conference on Learning Theory (COLT)*, 2017.

Wang, Y.-X., Fienberg, S. E., and Smola, A. J. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 2493–2502. JMLR.org, 2015. URL http://dl.acm.org/citation.cfm?id=3045118.3045383.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 681–688, 2011.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Representation Learning (ICLR)*, 2017.

Zhang, T. From ε-entropy to kl-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006a. doi: 10.1214/009053606000000704.

Zhang, T. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006b.