Entropy-SG(L)D optimizes the prior of a (valid) PAC-Bayes bound

Gintare Karolina Dziugaite
University of Cambridge
Vector Institute

Daniel M. Roy
University of Toronto
Vector Institute

Abstract

In recent work, we have shown that Entropy-SGD (Chaudhari et al., 2017), when viewed as a learning algorithm for classifiers, optimizes a PAC-Bayes bound on the risk of the classifier, or more accurately, the Gibbs posterior, i.e., a risk-sensitive perturbation of the 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 have shown that an ε-differentially private prior yields a valid PAC-Bayes bound, a straightforward consequence of results connecting generalization with differential privacy. Using stochastic gradient Langevin dynamics (SGLD) to approximate the well-known exponential release mechanism, we observed that generalization error on MNIST (measured on held out data) falls within the (empirically nonvacuous) bounds computed under the assumption that SGLD enjoys the same privacy as an exponential release. 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 workhorse of deep learning, which has good generalization performance in many settings, 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 performing (approximate) stochastic gradient ascent on the so-called local entropy $F(\mathbf{w}) = \log \int \exp(-f(\mathbf{w}+\xi))\mathscr{N}(\mathrm{d}\xi)$, where $\mathscr{N}$ is a zero-mean isotropic multivariate normal distribution on $\mathbb{R}^p$. (See Appendix A for related work.)

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, 2013; Catoni, 2007) 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. (See our arXiv paper for a formal statement and proof.) 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. In the special case where f is the empirical cross entropy, the local entropy is literally a Bayesian log marginal density. The connection between minimizing PAC-Bayes bounds and maximizing log marginal densities is the subject of recent work by Germain et al. (2016).

Second workshop on Bayesian Deep Learning (NIPS 2017), Long Beach, CA, USA.

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 (top-right plot in Fig. 1). We identify two changes that suffice to control generalization error.

The first change relates to the stability of optimizing the prior mean. 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.

We formalize this intuition using differential privacy (Dwork, 2006; Dwork et al., 2015b). By truncating the cross entropy loss and replacing SGD with stochastic gradient Langevin dynamics (SGLD; Welling and Teh, 2011), the data-dependent prior mean can be shown to be (ε, δ)-differentially private (Wang, Fienberg, and Smola, 2015; Minami et al., 2016). Using results connecting statistical validity and differential privacy (Dwork et al., 2015b, Thm. 11), we can also show that an ε-differentially private prior mean yields a valid, though slightly expanded, generalization bound using the PAC-Bayes theorem. (We refer to the SGLD variant as Entropy-SG(L)D.)

Theorem 1.1 (PAC-Bayes with differentially private prior)**.** *Let $\mathscr{D} \in \mathscr{M}_1(\mathbb{R}^k \times \{-1, 1\})$ be the data distribution and $m \in \mathbb{N}$ be the number of training examples. Let the hypothesis space be define by neural network weights taking values in $\mathbb{R}^p$. Let $\mathscr{P}\colon Z^m \rightsquigarrow \mathscr{M}_1(\mathbb{R}^p)$ be an algorithm on the training data returning a (prior) probability distribution on the hypothesis class and let $\delta > 0$. If $\mathscr{P}$ is ε-differentially private, then*

$$\mathop{\mathbb{P}}_{S\sim\mathscr{D}^m}\left((\forall Q)\ \mathrm{KL}(\hat{R}_S(Q)||R_{\mathscr{D}}(Q)) \le \frac{\mathrm{KL}(Q||\mathscr{P}(S)) + \ln 2m + 2\max\{\ln\frac{3}{\delta},\, m\varepsilon^2\}}{m-1}\right) \ge 1-\delta. \quad (1)$$

Here $\hat{R}_S(Q)$ and $R_{\mathscr{D}}(Q)$ refer to an empirical and true risk of a randomized classifier Q.

A gap remains between pure and approximate differential privacy. In the limit as the number of iterations diverges, the distribution of SGLD's output is known to converge weakly to the corresponding stationary distribution. (See recent work by Chen, Ding, and Carin (2015) and references therein.) Weak convergence, however, falls short of implying that SGLD eventually delivers pure differential privacy. Whether and when it does is an important open problem. Regardless, we may proceed under the optimistic assumption that the privacy of SGLD is comparable to that of an exponential release, and apply our ε-differentially private PAC-Bayes bound. We find that the corresponding 95% confidence intervals are reasonably tight but still conservative in our experiments.

The second change pertains to the stability of the stochastic gradient estimate made on each iteration of Entropy-SG(L)D. This estimate is made using SGLD (hence Entropy-SG(L)D is SG(L)D with a few iterations of SGLD at every step to approximate the gradient). A subtle detail of the SGLD within Entropy-SGD is that the noise added to the gradient is intentionally divided by a factor that ranges from 1000–10000. The result is that the Lipschitz constant of the objective function is 1000–10000 times larger, making Entropy-SGD much less stable as a result. This change to the noise also negatively impacts the differential privacy of the prior mean. Working backwards from the desire to obtain reasonably tight generalization bounds, we are led to instead *multiply* the SGLD noise by a factor of $\sqrt{m}$, where m is the number of data points. The resulting bounds (which assume that SGLD implements an idealized exponential release mechanism), are nonvacuous and tighter than those recently published by Dziugaite and Roy (2017), although it must be emphasized that ours hold subject to an assumption about the privacy of the prior mean, which is certainly violated but to an unknown degree.

2 Summary of evaluations on MNIST

We evaluated Entropy-SGLD's performance and generalization bounds on a binary classification task adapted from MNIST (LeCun, Cortes, and Burges, 2010).[1] Some experiments involved random labels, i.e., labels drawn independently and uniformly at random at the start of training. We studied

[1] The MNIST handwritten digits dataset (LeCun, Cortes, and Burges, 2010) consists of 60000 training set images and 10000 test set images, labeled 0–9. We transformed MNIST to a binary classification task, combining digits 0–4 and 5–9.

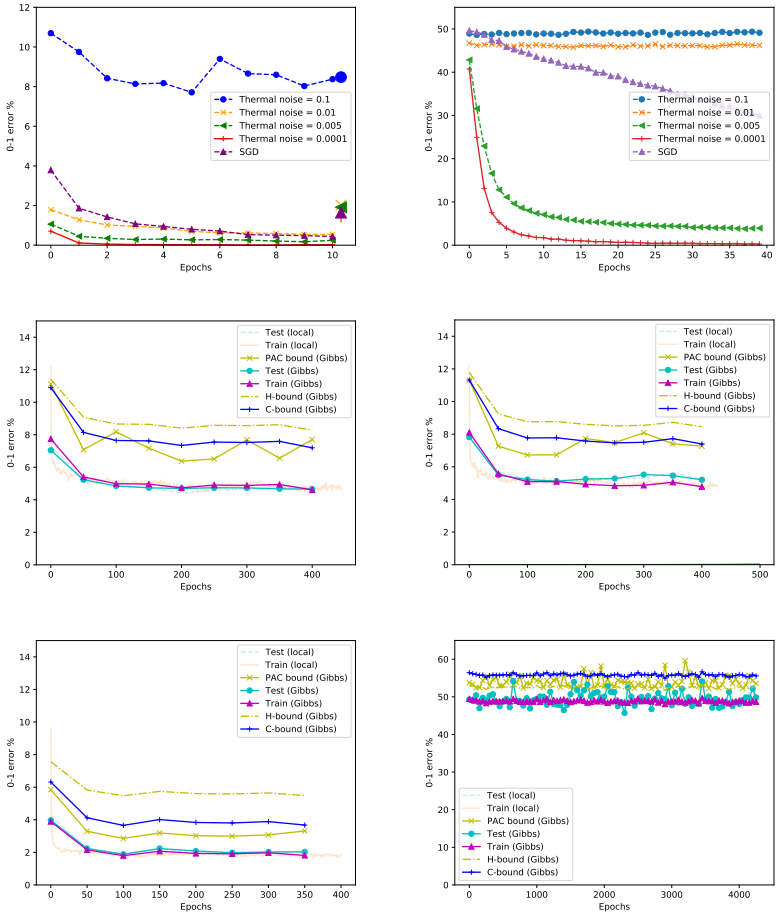

Figure 1: **(clockwise from top-left)** Entropy-SGD and vanilla-SGD performance on the training set of the binarized MNIST task. The lines indicate the 0–1 error on the training data. The larger marker at the end of 10 epochs indicates the 0–1 error on the test set, which is an empirical estimate of the true error. Thus the gap between the line (training error) and the final marker (empirical test error) is the approximate generalization error. On true labels, both algorithms find classifiers with small empirical risk and low generalization error. As we increase the thermal noise of entropy-SGD algorithm, the empirical 0–1 error increases, but the generalization gap decreases. **(clockwise from top-right)** On random labels, both algorithms exhibit high generalization error. (True error is $\sim 50\%$). **(middle-left)** Entropy-SGLD applied to FC600 network trained on true labels. **(middle-right)** Entropy-SGLD applied to FC1200 network trained on true labels. Both training error and generalization error are similar to FC600 case. Bounds are loose but nonvacuous. **(bot-right)** Entropy-SGLD applied to CONV network on true labels. Lower error and bounds than achieved with FC networks. **(bot-left)** Entropy-SGLD applied to FC600 network on random labels. The algorithm does not overfit like SGD and Entropy-SGD.

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 binary version of MNIST.[2] Let S, S_{tst} denote the training, test sets, respectively. We tracked

(i) $\hat{R}_S(\mathbf{w})$ and $\hat{R}_{S_{\text{tst}}}(\mathbf{w})$, i.e., the training and test error for $\mathbf{w}$ ("local")

(ii) estimates of $\hat{R}_S(P_{\gamma,\tau}^{\mathbf{w},S})$ and $\hat{R}_{S_{\text{tst}}}(P_{\gamma,\tau}^{\mathbf{w},S})$, i.e., the mean training and test error of the local Gibbs distribution, viewed as a randomized classifier ("Gibbs")

and, using differentially privacy, compute

(iii) a PAC-Bayes bound on $R_{\mathscr{D}}(P_{\gamma,\tau}^{\mathbf{w},S})$ using Theorem 1.1 ("PAC-bound");

(iv) the mean of a Hoeffding-style ("H") bound on $R_{\mathscr{D}}(\mathbf{w}')$, where $\mathbf{w}' \sim P_{\gamma,\tau}^{\mathbf{w},S}$, using Oneto, Ridella, and Anguita (2017, Lem. 2), see also (Dwork et al., 2015b, Thm. 9);

(v) an upper bound on the mean of a Chernoff-style ("C") bound on $R_{\mathscr{D}}(\mathbf{w}')$, where $\mathbf{w}' \sim P_{\gamma,\tau}^{\mathbf{w},S}$, using (Oneto, Ridella, and Anguita, 2017, Lem. 3).

Note that $R_{\mathscr{D}}(P_{\gamma,\tau}^{\mathbf{w},S}) = \mathop{\mathbb{E}}_{\mathbf{w}' \sim P_{\gamma,\tau}^{\mathbf{w},S}}(R_{\mathscr{D}}(\mathbf{w}'))$, 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.

2.1 Results

Key results on FC600 and CONV appear in Fig. 1.

On the true label dataset, Entropy-SGLD with our choice of differential privacy parameter ε achieves a lower training accuracy than vanilla SGD or Entropy-SGD. However, both the local and Gibbs classifiers found by the algorithm have essentially zero generalization error. Performance and bounds for FC600 and FC1200 are nearly identical, despite FC1200 having three times as many parameters. Training the CONV network produces the lowest training/test errors. On random labels, vanilla SGD and Entropy-SGD on FC600 overfit, while Entropy-SGLD maintain essentially zero generalization error.

We find that PAC-Bayes bound is comparable or tighter than H- and C-bounds . All bounds are nonvacuous for the choice of the algorithmic parameters, though still loose. We discuss this gap later. The error bounds computed here are tighter than the ordinary PAC-Bayes bounds reported by Dziugaite and Roy (2017). On the other hand, there are several unrealistic assumptions that have gone into the analysis, which affects the validity of the bounds. Foremost, running SGLD for a finite number of iterations does not deliver pure differential privacy. Despite the unrealistic assumptions in the analysis, no bound is ever violated. SGLD converges weakly to the exponential release mechanism. We use this fact to justify assuming that the privacy of SGLD does not in fact grow with the number of iterations, at least asymptotically. Alternative analysis of per step differential privacy would require the use of composition and sequencing of differential privacy of the algorithm, which in turn very quickly lead to vacuous bounds.

2.1.1 Comparison to SGLD

For comparison, we evaluate SGLD performance under the (even more unrealistic) assumption that SGLD gives us a perfect sample from the Gibbs posterior. We train the FC600 network with SGLD minimizing (clipped) binary cross entropy loss using different τ values. Similarly as for Entropy-SGLD and Entropy-SGD algorithms, the larger the τ value, the smaller the training error can be achieved. On the random label dataset, this means that larger τ values results in overfitting, since the training error drops well below 50%. On the true labelling of binarized MNIST dataset, we compare SGLD performance to Entropy-SGLD using the same τ. SGLD achieved lower accuracy on the train and test sets ($\sim 1-2\%$ lower than Entropy-SGLD). The C-bound on the test error

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

evaluated on SGLD network was above 10%, which is around 2% higher than for Entropy-SGLD trained network.

3 Discussion

Given how the training and test error track each other for Entropy-SGLD, it seems possible that our differential privacy bounds are very loose. Indeed, given the similarity between Entropy-SGD and vanilla SGLD, and the fact that SGLD approximates a sample from a Gibbs distribution, it seems possible that the gap in our analysis is substantial.

On the other hand, it also seems conceivable that there is a tradeoff between the speed of learning, the achievable error, and the ability to produce a certificate of one's generalization error (e.g., via a DP bound). EntropySGD learns much faster in its original configuration, but its performance on random labels implies it has poor differential privacy.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang (2016). "Deep Learning with Differential Privacy". *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. CCS '16. Vienna, Austria: ACM, pp. 308–318. DOI: 10.1145/2976749.2978318.

Alessandro Achille and Stefano Soatto (2017). "On the Emergence of Invariance and Disentangling in Deep Representations". *CoRR* abs/1706.01350. arXiv: 1706.01350.

Carlo Baldassi, Alessandro Ingrosso, Carlo Lucibello, Luca Saglietti, and Riccardo Zecchina (2015). "Subdominant Dense Clusters Allow for Simple Learning and High Computational Performance in Neural Networks with Discrete Synapses". *Phys. Rev. Lett.* 115 (12), p. 128101. DOI: 10.1103/PhysRevLett.115.128101.

Carlo Baldassi, Christian Borgs, Jennifer T. Chayes, Alessandro Ingrosso, Carlo Lucibello, Luca Saglietti, and Riccardo Zecchina (2016). "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. DOI: 10.1073/pnas.1608103113. eprint: http://www.pnas.org/content/113/48/E7655.full.pdf.

Peter Bartlett, Dylan J. Foster, and Matus Telgarsky (2017). "Spectrally-normalized margin bounds for neural networks". *CoRR* abs/1706.08498. arXiv: 1706.08498.

Peter L. Bartlett and Shahar Mendelson (2003). "Rademacher and Gaussian Complexities: Risk Bounds and Structural Results". *J. Mach. Learn. Res.* 3, pp. 463–482.

Raef Bassily, Adam Smith, and Abhradeep Thakurta (2014). "Differentially private empirical risk minimization: Efficient algorithms and tight error bounds". *arXiv preprint arXiv:1405.7085*.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman (2016). "Algorithmic stability for adaptive data analysis". *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, pp. 1046–1059.

Olivier Catoni (2007). "PAC-Bayesian supervised classification: the thermodynamics of statistical learning". *arXiv preprint arXiv:0712.0248*.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina (2017). "Entropy-SGD: Biasing Gradient Descent Into Wide Valleys". *International Conference on Learning Representations (ICLR)*. arXiv: 1611.01838v4 [cs.LG].

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate (2011). "Differentially private empirical risk minimization". *Journal of Machine Learning Research* 12.Mar, pp. 1069–1109.

Changyou Chen, Nan Ding, and Lawrence Carin (2015). "On the convergence of stochastic gradient MCMC algorithms with high-order integrators". *Advances in Neural Information Processing Systems*, pp. 2278–2286.

Christos Dimitrakakis, Blaine Nelson, Aikaterini Mitrokotsa, and Benjamin IP Rubinstein (2014). "Robust and private Bayesian inference". *International Conference on Algorithmic Learning Theory*. Springer, pp. 291–305.

Cynthia Dwork (2006). "Differential Privacy". *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10–14, 2006, Proceedings, Part II*. Ed. by Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–12. DOI: 10.1007/11787006_1.

– (2008). "Differential privacy: A survey of results". *International Conference on Theory and Applications of Models of Computation*. Springer, pp. 1–19.

Cynthia Dwork, Aaron Roth, et al. (2014). "The algorithmic foundations of differential privacy". *Foundations and Trends in Theoretical Computer Science* 9.3–4, pp. 211–407.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth (2015a). "Generalization in adaptive data analysis and holdout reuse". *Advances in Neural Information Processing Systems*, pp. 2350–2358.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth (2015b). "Preserving statistical validity in adaptive data analysis". *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*. ACM, pp. 117–126.

Gintare Karolina Dziugaite and Daniel M. Roy (2017). "Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data". *arXiv preprint arXiv:1703.11008*.

Pascal Germain, Francis Bach, Alexandre Lacoste, and Simon Lacoste-Julien (2016). "PAC-Bayesian Theory Meets Bayesian Inference". *Advances in Neural Information Processing Systems*, pp. 1884–1892.

Moritz Hardt, Benjamin Recht, and Yoram Singer (2015). "Train faster, generalize better: Stability of stochastic gradient descent". *CoRR* abs/1509.01240.

Geoffrey E. Hinton and Drew van Camp (1993). "Keeping the Neural Networks Simple by Minimizing the Description Length of the Weights". *Proceedings of the Sixth Annual Conference on Computational Learning Theory*. COLT '93. Santa Cruz, California, USA: ACM, pp. 5–13. DOI: 10.1145/168304.168306.

Sepp Hochreiter and Jürgen Schmidhuber (1997). "Flat Minima". *Neural Comput.* 9.1, pp. 1–42. DOI: 10.1162/neco.1997.9.1.1.

Daniel Kifer, Adam Smith, and Abhradeep Thakurta (2012). "Private convex empirical risk minimization and high-dimensional regression". *Journal of Machine Learning Research* 1.41, pp. 1–40.

Diederik P Kingma, Tim Salimans, and Max Welling (2015). "Variational Dropout and the Local Reparameterization Trick". *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett. Curran Associates, Inc., pp. 2575–2583.

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

John Langford and Rich Caruana (2002). "(Not) Bounding the True Error". *Advances in Neural Information Processing Systems 14*. Ed. by T. G. Dietterich, S. Becker, and Z. Ghahramani. MIT Press, pp. 809–816.

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 (2010). *MNIST handwritten digit database*. http://yann.lecun.com/exdb/mnist/.

Guy Lever, François Laviolette, and John Shawe-Taylor (2013). "Tighter PAC-Bayes bounds through distribution-dependent priors". *Theoretical Computer Science* 473, pp. 4–28. DOI: http://dx.doi.org/10.1016/j.tcs.2012.10.013.

David A. McAllester (1999). "PAC-Bayesian Model Averaging". *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*. COLT '99. Santa Cruz, California, USA: ACM, pp. 164–170. DOI: 10.1145/307400.307435.

– (2013). "A PAC-Bayesian Tutorial with A Dropout Bound". *CoRR* abs/1307.2118.

Kentaro Minami, Hitomi Arai, Issei Sato, and Hiroshi Nakagawa (2016). "Differential Privacy without Sensitivity". *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett. Curran Associates, Inc., pp. 956–964.

Darakhshan J Mir (2013). "Differential privacy: an exploration of the privacy-utility landscape". PhD thesis. Rutgers University.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro (2014). *In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning*. Workshop track poster at ICLR 2015. arXiv: 1412.6614v4 [cs.LG].

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro (2017a). "A PAC-Bayesian Approach to Spectrally-Normalized Margin Bounds for Neural Networks". *CoRR* abs/1707.09564. arXiv: 1707.09564.

– (2017b). "Exploring Generalization in Deep Learning". *CoRR* abs/1706.08947. arXiv: 1706.08947.

Luca Oneto, Sandro Ridella, and Davide Anguita (2017). "Differential privacy and generalization: Sharper bounds with applications". *Pattern Recognition Letters* 89, pp. 31–38. DOI: https://doi.org/10.1016/j.patrec.2017.02.006.

Yu-Xiang Wang, Stephen E. Fienberg, and Alexander J. Smola (2015). "Privacy for Free: Posterior Sampling and Stochastic Gradient Monte Carlo". *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. ICML'15. Lille, France: JMLR.org, pp. 2493–2502.

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

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals (2017). "Understanding deep learning requires rethinking generalization". *International Conference on Representation Learning (ICLR)*. arXiv: 1611.03530v2 [cs.LG].

A Entropy SGD algorithm

Algorithm 1 One ESGD step along the local entropy gradient

Input:

$\mathbf{w} \in \mathbb{R}^p$ ▷ Current weight

$S \in Z^m$ ▷ Data

$\ell : \mathbb{R}^p \times Z \to \mathbb{R}$ ▷ Loss

$\tau, \gamma, \eta, \eta', L, K$ ▷ Parameters

Output: Weight vector $\mathbf{w}$ moved along stochastic gradient

procedure ENTROPY-SGD-STEP($\tau, \gamma, \eta, \eta', L, K, \mathbf{w}, S$)

  $\mathbf{w}', \mu \leftarrow \mathbf{w}$

  for $i \in \{1, \ldots, L\}$ **do** ▷ Run SGLD for L iterations.

    $\eta'_i \leftarrow \eta'/i$

    $(z_{j_1}, \ldots, z_{j_K}) \leftarrow$ sample size K minibatch from S

    $\mathrm{d}\mathbf{w}' \leftarrow \frac{1}{K}\sum_{i=1}^{K} \nabla_{\mathbf{w}'} \ell(\mathbf{w}', z_{j_i}) - \gamma(\mathbf{w}' - \mathbf{w})$

    $\mathbf{w}' \leftarrow \mathbf{w}' - \eta'_i \mathrm{d}\mathbf{w}' + \sqrt{\eta'_i}\frac{1}{\sqrt{\tau}} N(0, I)$

    $\mu \leftarrow (1-\alpha)\mu + \alpha\mathbf{w}'$

  $\mathbf{w} \leftarrow \mathbf{w} - \eta\tau\gamma(\mathbf{w} - \mu)$ ▷ Step along stochastic local entropy ∇

  return $\mathbf{w}$

B Related work

A key aspect of our analysis relies on the stability of a data-dependent prior. Stability has long been understood to relate to generalization. (See recent work by Hardt, Recht, and Singer (2015) and references therein.) This work was also inspired in part by observations of Zhang et al. (2017), who studied empirical properties of SGD. They show that, without regularization, SGD can achieve zero training error on MNIST and CIFAR, even if the labels are chosen uniformly at random. At the same time, SGD obtains weights with very small generalization error with the original labels. The first observation is strong evidence that the set of classifier accessible to SGD within a reasonable number of iterations is extremely rich. Indeed, with probability almost indistinguishable from one, fitting random labels on a large data set implies that the Rademacher complexity of the hypothesis class is essentially the maximum possible (Bartlett and Mendelson, 2003, Thm. 11).

Similar observations were made by Neyshabur, Tomioka, and Srebro (2014), who argue that implicit regularization underlies the ability of SGD to generalize. Recent work also connects the curvature (or local volume) of the empirical risk surface to generalization (Bartlett, Foster, and Telgarsky, 2017; Neyshabur et al., 2017b; Neyshabur et al., 2017a; Dziugaite and Roy, 2017)

These ideas connect to early work by Hinton and Camp (1993); Hochreiter and Schmidhuber (1997) which introduced regularization schemes based on information theoretic ideas. These ideas, now referred to as "flat minima", can be related to minimizing PAC-Bayes bounds, although these bounds are minimized with respect to the posterior, not the prior, as is done by Entropy-SGD (Dziugaite and Roy, 2017). Achille and Soatto (2017) provides additional information-theoretic arguments for a regularization scheme similar to that of Hinton and Camp. Their objective takes the form of regularized empirical cross entropy

$$\hat{R}_{S_m}(Q) + \beta \mathrm{KL}(Q||P), \tag{2}$$

where Q and P are the prior and posterior on the weights, respectively. For an appropriate range of β, linear PAC-Bayes bounds are exactly of this form. In Achille and Soatto (2017) they empirically observe that varying β correlates with a degree of overfitting on a random label dataset. Their experimental insights agree with our privacy analysis as β directly affects the differential privacy, and thus controls an upper bound on generalization error. In addition, Achille and Soatto (2017) also note the connections with variational inference (Kingma, Salimans, and Welling, 2015).

This work also relates to renewed interest in nonvacuous generalization bounds (Langford, 2002; Langford and Caruana, 2002), i.e., bounds on the numerical difference between the unknown classification error and the training error that are (much) tighter than the tautological upper bound of one. Recently, Dziugaite and Roy (2017) demonstrated nonvacuous generalization bounds for random perturbations of SGD solutions using PAC-Bayes bounds for networks with millions of weights. Their work builds on the core insight demonstrated nearly 15 years ago by Langford and Caruana (2002), who computed nonvacuous bounds for neural networks five orders of magnitude smaller.

The analysis of Entropy-SGLD rests on results in differential privacy (see (Dwork, 2008) for a survey) and its connection to generalization (Dwork et al., 2015b; Dwork et al., 2015a; Bassily et al., 2016; Oneto, Ridella, and Anguita, 2017). Entropy-SGLD can be seen as an instance of differentially private empirical risk minimization, which is well studied, both in the abstract (Chaudhuri, Monteleoni, and Sarwate, 2011; Kifer, Smith, and Thakurta, 2012; Bassily, Smith, and Thakurta, 2014) and in the particular setting of private training via SGD (Bassily, Smith, and Thakurta, 2014; Abadi et al., 2016). Our analysis also rests on the differential privacy of Bayesian and Gibbs posteriors, and approximate sampling algorithms (Mir, 2013; Bassily, Smith, and Thakurta, 2014; Dimitrakakis et al., 2014; Wang, Fienberg, and Smola, 2015; Minami et al., 2016).

Our differentially private PAC-Bayes bound rely on data-distribution-dependent priors. Such bounds were first studied by Catoni (2007) and further studied by Lever, Laviolette, and Shawe-Taylor (2013).