Deep Neural Networks: From Flat Minima to Numerically Nonvacuous Generalization Bounds via PAC-Bayes

Daniel M. Roy
University of Toronto; Vector Institute

Joint work with

Gintarė K. Džiugaitė
University of Cambridge

(Almost) 50 Shades of Bayesian Learning: PAC-Bayesian trends and insights
Long Beach, CA. December 2017

How does SGD work?

- Growing body of work arguing that SGD performs implicit regularization
- Problem: No matching generalization bounds that are nonvacuous when applied to real data and networks.
- We focus on **"flat minima"** – weights $\mathbf{w}$ such that training error is "insensitive" to "large" perturbations
- We show the size/flatness/location of minima found by SGD on MNIST imply generalization using PAC-Bayes bounds
- Focusing on MNIST, we show how to *compute* generalization bounds that are *nonvacuous* for stochastic networks with millions of weights.

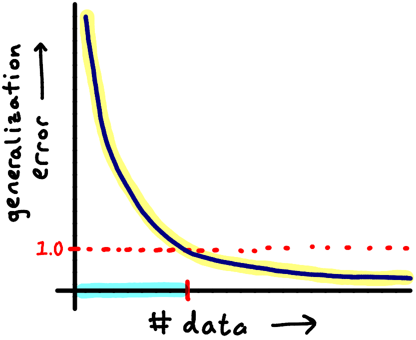
How does SGD work?

- Growing body of work arguing that SGD performs implicit regularization
- Problem: No matching generalization bounds that are nonvacuous when applied to real data and networks.
- We focus on **"flat minima"** – weights $\mathbf{w}$ such that training error is "insensitive" to "large" perturbations
- We show the size/flatness/location of minima found by SGD on MNIST imply generalization using PAC-Bayes bounds
- Focusing on MNIST, we show how to *compute* generalization bounds that are *nonvacuous* for stochastic networks with millions of weights.
- We obtain our (data-dependent, PAC-Bayesian) generalization bounds via a fair bit of computation with SGD. Our approach is a modern take on Langford and Caruana (2002).

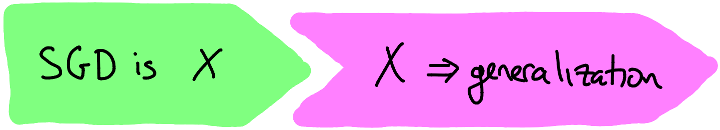
Nonvacuous generalization bounds

$$\begin{aligned} \text{risk:} \quad & L_{\mathcal{D}}(h) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(h(x),y)], \quad \mathcal{D} \text{ unknown} \\ \text{empirical risk:} \quad & L_S(h) = \tfrac{1}{m}\textstyle\sum_{i=1}^m \ell(h(x_i),y_i), \quad S = \{(x_1,y_1),\ldots,(x_m,y_m)\} \\ \text{generalization error:} \quad & L_{\mathcal{D}}(h) - L_S(h) \end{aligned}$$

$$\forall \mathcal{D} \ \mathbb{P}_{S\sim\mathcal{D}^m}\Big(L_{\mathcal{D}}(\hat{h}) - L_S(\hat{h}) < \underbrace{\epsilon(\mathcal{H}, m, \delta, S, \hat{h})}_{\text{generalization err. bound}}\Big) > 1-\delta$$



SGD is X and X implies generalization



"SGD is Empirical Risk Minimization for large enough networks"
"SGD is (Implicit) Regularized Loss Minimization"
"SGD is Approximate Bayesian Inference"
...

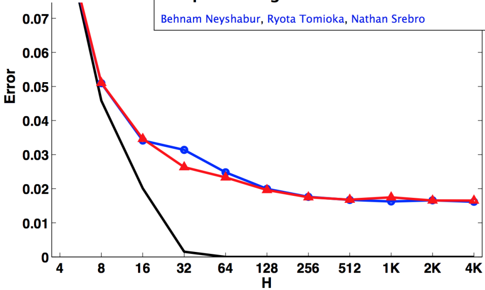
No statement of the form "SGD is X" *explains* generalization in deep learning until we know that X *implies* generalization *under real-world conditions*.

SGD is (not simply) empirical risk minimization

Computer Science > Learning

In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning

Behnam Neyshabur, Ryota Tomioka, Nathan Srebro



Training error of SGD at convergence.

Test error at convergence and for **early stopping** identical.

SGD $\approx$ Empirical Risk Minimization $\quad \arg\min_{\mathbf{w}\in\mathcal{H}} L_S(\mathbf{w})$

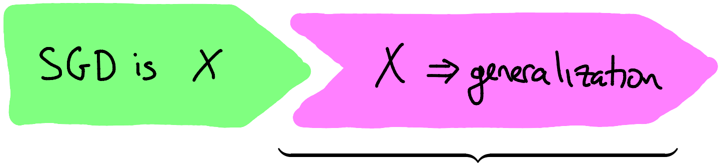
MNIST has 60,000 training data

Two-layer fully connected ReLU network has >1m parameters

$\Longrightarrow$ PAC bounds are vacuous

$\Longrightarrow$ PAC bounds can't explain this curve

Our focus: Statistical Learning Aspect



On MNIST, with realistic networks, . . .

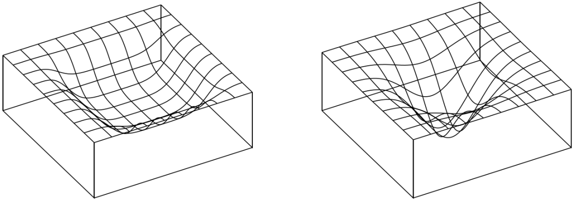
- VC bounds don't imply generalization
- Classic Margin + Norm-bounded Rademacher Complexity Bounds don't imply generalization
- Being "Bayesian" does not necessarily imply generalization (sorry!)

Using **PAC-Bayes bounds**, we show that size/flatness/location of minima, found by SGD on MNIST, imply generalization for MNIST.

Our bounds require a fair bit of computation/optimization to evaluate. Strictly speaking, they bound the error of a random perturbation of the SGD solution.

Flat minima...

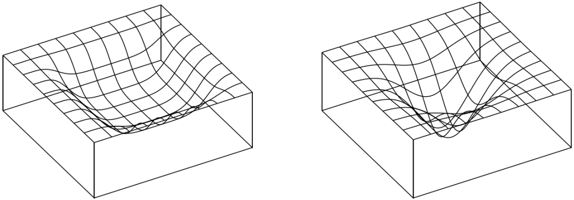
training error in flat minima is “insensitive” to “large” perturbations



(Hochreiter and Schmidhuber, 1997)

Flat minima...

training error in flat minima is “insensitive” to “large” perturbations



(Hochreiter and Schmidhuber, 1997)

... meets the PAC-Bayes theorem (McAllister)

$$\forall \mathcal{D}\ \forall P \ \underset{S\sim\mathcal{D}^m}{\mathbb{P}}\Big[\forall Q\ \Delta\big(L_S(Q), L_{\mathcal{D}}(Q)\big) \le \frac{\mathrm{KL}(Q||P)+\log\frac{\mathcal{I}^{\Delta}(m)}{\delta}}{m}\Big] \ge 1-\delta$$

For any data distribution, $\mathcal{D}$, i.e., no assumptions,
For any “prior” randomized classifier P, even nonsense,
with high probability over m i.i.d. samples $S \sim \mathcal{D}^m$,
For any “posterior” randomized classifier Q, not nec. Bayes rule,

Generalization error of Q bounded approximately by $\frac{1}{m}\mathrm{KL}(Q||P)$

Controlling generalization error of randomized classifiers

Let $\mathcal{H}$ be a hypothesis class of binary classifiers $\mathbb{R}^k \to \{-1, 1\}$.
A randomized classifier is a distribution Q on $\mathcal{H}$. Its risk is

$$L_{\mathcal{D}}(Q) = \mathop{\mathbb{E}}_{w \sim Q}[L_{\mathcal{D}}(h_w)]$$

Among the sharpest generalization bounds for randomized classifiers are PAC-Bayes bounds (McAllester, 1999).

Theorem (PAC-Bayes (Catoni, 2007)). .

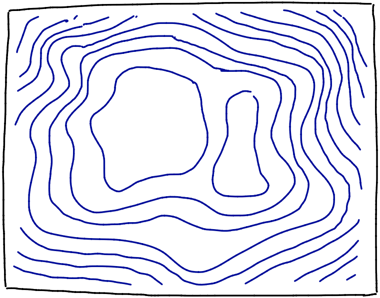
Let $\delta > 0$ and $m \in \mathbb{N}$ and assume $L_{\mathcal{D}}$ is bounded. Then

$$\forall P,\ \forall \mathcal{D},\ \mathop{\mathbb{P}}_{S \sim \mathcal{D}^m}\Big(\forall Q,\ L_{\mathcal{D}}(Q) \leq 2\, L_S(Q) + 2\, \frac{\mathrm{KL}(Q||P) + \log \frac{1}{\delta}}{m}\Big) \geq 1 - \delta$$

Our approach

given m i.i.d. data $S \sim \mathcal{D}^m$

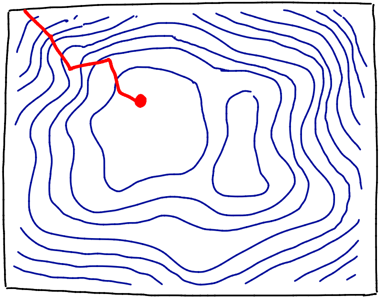
Our approach



given m i.i.d. data $S \sim \mathcal{D}^m$

empirical error surface
$w \mapsto L_S(h_w)$

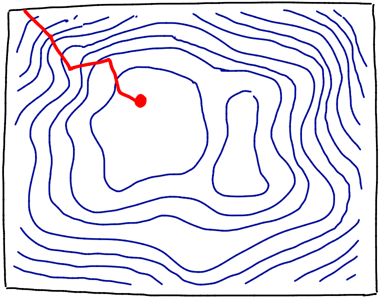
Our approach



given m i.i.d. data $S \sim \mathcal{D}^m$

empirical error surface
$w \mapsto L_S(h_w)$

Our approach

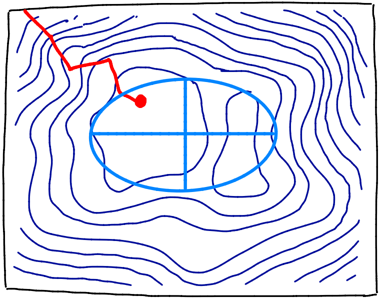


given m i.i.d. data $S \sim \mathcal{D}^m$

empirical error surface
$w \mapsto L_S(h_w)$

• $w_{\mathrm{SGD}} \in \mathbb{R}^{472000}$
weights learned by SGD on MNIST

Our approach

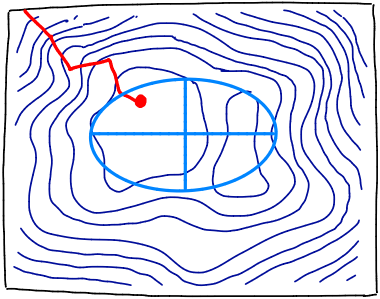


given m i.i.d. data $S \sim \mathcal{D}^m$

empirical error surface
$w \mapsto L_S(h_w)$

- $w_{\mathrm{SGD}} \in \mathbb{R}^{472000}$
weights learned by SGD on MNIST

Our approach



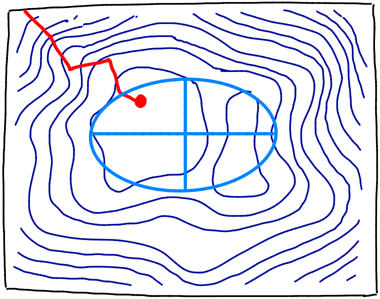
given m i.i.d. data $S \sim \mathcal{D}^m$

empirical error surface
$w \mapsto L_S(h_w)$

• $w_{\mathrm{SGD}} \in \mathbb{R}^{472000}$
weights learned by SGD on MNIST

$\oplus$ $\hat{Q} = \mathcal{N}(w_{\mathrm{SGD}} + w', \Sigma')$
stochastic neural net

Our approach



given m i.i.d. data $S \sim \mathcal{D}^m$

empirical error surface
$w \mapsto L_S(h_w)$

• $w_{\mathrm{SGD}} \in \mathbb{R}^{472000}$
weights learned by SGD on MNIST

$\oplus$ $\hat{Q} = \mathcal{N}(w_{\mathrm{SGD}} + w', \Sigma')$
stochastic neural net

generalization/error bound: $\forall \mathcal{D} \underset{S \sim \mathcal{D}^m}{\mathbb{P}} \left(L_{\mathcal{D}}(\hat{Q}) < 0.17 \right) > 0.95$

Optimizing PAC-Bayes bounds

Given data S, we can find a provably good classifier Q by optimizing the PAC-Bayes bound w.r.t. Q.

For Catoni's PAC-Bayes bound, the optimization problem is of the form

$$\sup_{Q} \; -\tau L_S(Q) - \mathrm{KL}(Q||P).$$

Lemma. Optimal Q satisfies $\underbrace{\frac{\mathrm{d}Q}{\mathrm{d}P}(\mathbf{w}) = \frac{\exp(-\tau L_S(\mathbf{w}))}{\int \exp(-\tau L_S(\mathbf{w}))P(\mathrm{d}\mathbf{w})}}_{\text{generalized Bayes rule}}.$

Observation. Under log loss and $\tau = m$, the term $-\tau L_S(\mathbf{w})$ is the expected log likelihood under Q and the objective is the ELBO.

Lemma. $\log \int \exp(-\tau L_S(\mathbf{w}))P(\mathrm{d}\mathbf{w}) = \sup_{Q} \; -\tau L_S(Q) - \mathrm{KL}(Q||P).$

Observation. Under log loss and $\tau = m$, l.h.s. is log marginal likelihood.

Cf. Zhang 2004, 2006, Alquier et al. 2015, Germain et al. 2016.

PAC-Bayes Bound optimization

PAC-Bayes Bound optimization

$$\inf_{Q} \quad L_S(Q) + \frac{\mathrm{KL}(Q||P) + \log \frac{1}{\delta}}{m}$$

PAC-Bayes Bound optimization

$$\inf_{Q} \quad L_S(Q) + \frac{\mathrm{KL}(Q||P) + \log \frac{1}{\delta}}{m}$$

Let $\tilde{L}_S(Q) \geq L_S(Q)$ with $\tilde{L}_S$ differentiable.

PAC-Bayes Bound optimization

$$\inf_{Q} \quad L_S(Q) + \frac{\mathrm{KL}(Q||P) + \log \frac{1}{\delta}}{m}$$

Let $\tilde{L}_S(Q) \geq L_S(Q)$ with $\tilde{L}_S$ differentiable.

$$\inf_{Q} \quad m\,\tilde{L}_S(Q) + \mathrm{KL}(Q||P)$$

PAC-Bayes Bound optimization

$$\inf_{Q} \quad L_S(Q) + \frac{\mathrm{KL}(Q||P) + \log \frac{1}{\delta}}{m}$$

Let $\tilde{L}_S(Q) \geq L_S(Q)$ with $\tilde{L}_S$ differentiable.

$$\inf_{Q} \quad m\,\tilde{L}_S(Q) + \mathrm{KL}(Q||P)$$

Let $Q_{w,s} = \mathcal{N}(w, \mathrm{diag}(s))$.

PAC-Bayes Bound optimization

$$\inf_{Q} \quad L_S(Q) + \frac{\mathrm{KL}(Q||P) + \log\frac{1}{\delta}}{m}$$

Let $\tilde{L}_S(Q) \geq L_S(Q)$ with $\tilde{L}_S$ differentiable.

$$\inf_{Q} \quad m\,\tilde{L}_S(Q) + \mathrm{KL}(Q||P)$$

Let $Q_{w,s} = \mathcal{N}(w, \mathrm{diag}(s))$.

$$\min_{\substack{w \in \mathbb{R}^d \\ s \in \mathbb{R}^d_+}} \quad m\,\tilde{L}_S(Q_{w,s}) + \mathrm{KL}(Q_{w,s}||P)$$

PAC-Bayes Bound optimization

$$\inf_{Q} \quad L_S(Q) + \frac{\mathrm{KL}(Q||P) + \log\frac{1}{\delta}}{m}$$

Let $\tilde{L}_S(Q) \geq L_S(Q)$ with $\tilde{L}_S$ differentiable.

$$\inf_{Q} \quad m\,\tilde{L}_S(Q) + \mathrm{KL}(Q||P)$$

Let $Q_{w,s} = \mathcal{N}(w, \mathrm{diag}(s))$.

$$\min_{\substack{w\in\mathbb{R}^d \\ s\in\mathbb{R}^d_+}} \quad m\,\tilde{L}_S(Q_{w,s}) + \mathrm{KL}(Q_{w,s}||P)$$

Take $P = \mathcal{N}(w_0, \lambda\, I_d)$ with $\lambda = c\exp\{-j/b\}$.

PAC-Bayes Bound optimization

$$\inf_{Q} \quad L_S(Q) + \frac{\mathrm{KL}(Q||P) + \log \frac{1}{\delta}}{m}$$

Let $\tilde{L}_S(Q) \geq L_S(Q)$ with $\tilde{L}_S$ differentiable.

$$\inf_{Q} \quad m\,\tilde{L}_S(Q) + \mathrm{KL}(Q||P)$$

Let $Q_{w,s} = \mathcal{N}(w, \mathrm{diag}(s))$.

$$\min_{\substack{w \in \mathbb{R}^d \\ s \in \mathbb{R}^d_+}} \quad m\,\tilde{L}_S(Q_{w,s}) + \mathrm{KL}(Q_{w,s}||P)$$

Take $P = \mathcal{N}(w_0, \lambda\, I_d)$ with $\lambda = c \exp\{-j/b\}$.

$$\min_{\substack{w \in \mathbb{R}^d \\ s \in \mathbb{R}^d_+ \\ \lambda \in (0,c)}} \quad m\,\tilde{L}_S(Q_{w,s}) + \underbrace{\mathrm{KL}(Q_{w,s}||\mathcal{N}(w_0, \lambda I))} + 2\log(b \log \frac{c}{\lambda})$$

PAC-Bayes Bound optimization

$$\inf_{Q} \quad L_S(Q) + \frac{\mathrm{KL}(Q||P) + \log \frac{1}{\delta}}{m}$$

Let $\tilde{L}_S(Q) \geq L_S(Q)$ with $\tilde{L}_S$ differentiable.

$$\inf_{Q} \quad m\,\tilde{L}_S(Q) + \mathrm{KL}(Q||P)$$

Let $Q_{w,s} = \mathcal{N}(w, \mathrm{diag}(s))$.

$$\min_{\substack{w \in \mathbb{R}^d \\ s \in \mathbb{R}^d_+}} \quad m\,\tilde{L}_S(Q_{w,s}) + \mathrm{KL}(Q_{w,s}||P)$$

Take $P = \mathcal{N}(w_0, \lambda\, I_d)$ with $\lambda = c \exp\{-j/b\}$.

$$\min_{\substack{w \in \mathbb{R}^d \\ s \in \mathbb{R}^d_+ \\ \lambda \in (0,c)}} \quad m\,\tilde{L}_S(Q_{w,s}) + \underbrace{\mathrm{KL}(Q_{w,s}||\mathcal{N}(w_0, \lambda I))}_{\frac{1}{2}(\frac{1}{\lambda}\|s\|_1 + \frac{1}{\lambda}\|w - w_0\|_2^2 + d \log \lambda - 1_d \cdot \log s - d)} + 2 \log(b \log \frac{c}{\lambda}).$$

Numerical generalization bounds on MNIST

# Hidden Layers	1	2	3	1 (R)
Train error	0.001	0.000	0.000	0.007
Test error	0.018	0.016	0.013	0.508
SNN train error	0.028	0.028	0.027	0.112
SNN test error	0.034	0.033	0.032	0.503
PAC-Bayes bound	0.161	0.186	0.201	1.352
KL divergence	5144	6534	7861	201131
# parameters	472k	832k	1193k	472k
VC dimension	26m	66m	121m	26m

We have shown that type of flat minima found in practice can be turned into a generalization guarantee.

Bounds are loose, but only nonvacuous bounds in this setting.

Actually, SGD is pretty dangerous

UNDERSTANDING DEEP LEARNING REQUIRES RE-THINKING GENERALIZATION

Chiyuan Zhang*
Massachusetts Institute of Technology
chiyuan@mit.edu

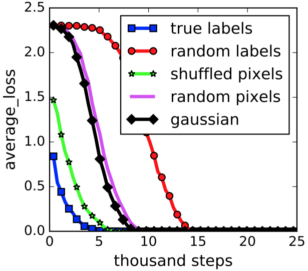
Samy Bengio
Google Brain
bengio@google.com

Moritz Hardt
Google Brain
mrtz@google.com

Benjamin Recht†
University of California, Berkeley
brecht@berkeley.edu

Oriol Vinyals
Google DeepMind
vinyals@google.com

Actually, SGD is pretty dangerous



- SGD achieves zero training error reliably
- Despite no explicit regularization, training and test error very close
- Explicit regularization has minor effect
- SGD can reliably obtain zero training error on *randomized* labels
  - Hence, Rademacher complexity of model class is near maximal w.h.p.

UNDERSTANDING DEEP LEARNING REQUIRES RE-THINKING GENERALIZATION

Chiyuan Zhang*
Massachusetts Institute of Technology
chiyuan@mit.edu

Samy Bengio
Google Brain
bengio@google.com

Moritz Hardt
Google Brain
mrtz@google.com

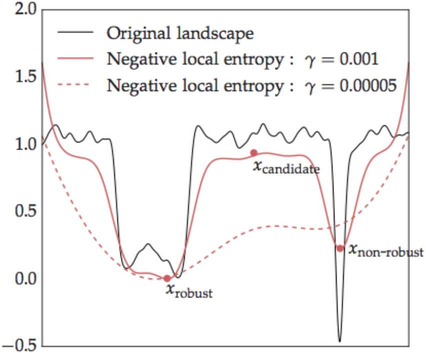
Benjamin Recht†
University of California, Berkeley
brecht@berkeley.edu

Oriol Vinyals
Google DeepMind
vinyals@google.com

Entropy-SGD (Chaudhari et al., 2017)

Entropy-SGD replaces stochastic gradient descent on L_S by stochastic gradient ascent applied to the optimization problem:

$$\arg\max_{\mathbf{w}\in\mathbb{R}^d} F_{\gamma,\tau}(\mathbf{w}; S),$$



where $F_{\gamma,\tau}(\mathbf{w}; S) = \log \int_{\mathbb{R}^p} \exp\left\{-\tau L_S(\mathbf{w}') - \tau\frac{\gamma}{2}\|\mathbf{w}' - \mathbf{w}\|_2^2\right\} \mathrm{d}\mathbf{w}'.$

The local entropy $F_{\gamma,\tau}(\cdot; S)$ emphasizes flat minima of L_S.

Entropy-SGD optimizes PAC-Bayes bound w.r.t. prior

Entropy-SGD optimizes the local entropy

$$F_{\gamma,\tau}(\mathbf{w}; S) = \log \int_{\mathbb{R}^p} \exp\left\{-\tau L_S(\mathbf{w}') - \tau \frac{\gamma}{2} \|\mathbf{w}' - \mathbf{w}\|_2^2\right\} \mathrm{d}\mathbf{w}'.$$

Entropy-SGD optimizes PAC-Bayes bound w.r.t. prior

Entropy-SGD optimizes the local entropy

$$F_{\gamma,\tau}(\mathbf{w};S) = \log \int_{\mathbb{R}^p} \exp\left\{-\tau L_S(\mathbf{w}') - \tau \frac{\gamma}{2} \|\mathbf{w}' - \mathbf{w}\|_2^2\right\} \mathrm{d}\mathbf{w}'.$$

Theorem. Maximizing $F_{\gamma,\tau}(\mathbf{w};S)$ w.r.t. $\mathbf{w}$ corresponds to minimizing PAC-Bayes risk bound w.r.t. prior's mean $\mathbf{w}$.

Entropy-SGD optimizes PAC-Bayes bound w.r.t. prior

Entropy-SGD optimizes the local entropy

$$F_{\gamma,\tau}(\mathbf{w}; S) = \log \int_{\mathbb{R}^p} \exp\left\{-\tau L_S(\mathbf{w}') - \tau \frac{\gamma}{2} \|\mathbf{w}' - \mathbf{w}\|_2^2\right\} \mathrm{d}\mathbf{w}'.$$

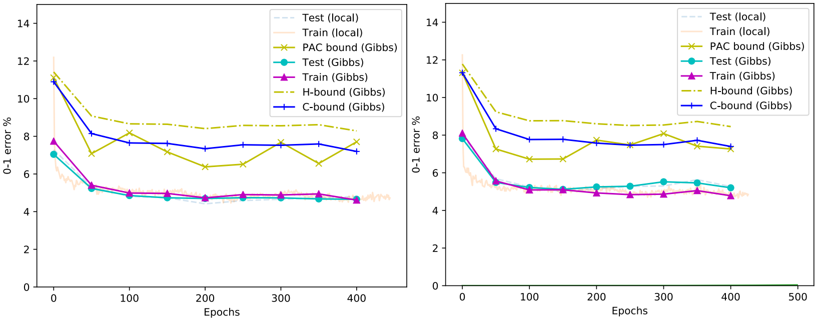
Theorem. Maximizing $F_{\gamma,\tau}(\mathbf{w}; S)$ w.r.t. $\mathbf{w}$ corresponds to minimizing PAC-Bayes risk bound w.r.t. prior's mean $\mathbf{w}$.

Theorem. Let $\mathscr{P}(S)$ be an ϵ-differentially private distribution. Then

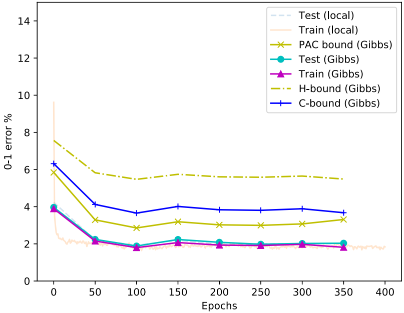
$$\forall \mathcal{D},\ \mathop{\mathbb{P}}_{S \sim \mathcal{D}^m}\left((\forall Q)\ \mathrm{KL}(L_S(Q) || L_{\mathcal{D}}(Q)) \leq \frac{\mathrm{KL}(Q || \mathscr{P}(S)) + \ln 2m + 2 \max\{\ln \frac{3}{\delta},\ m\epsilon^2\}}{m-1} \right) \geq 1 - \delta.$$

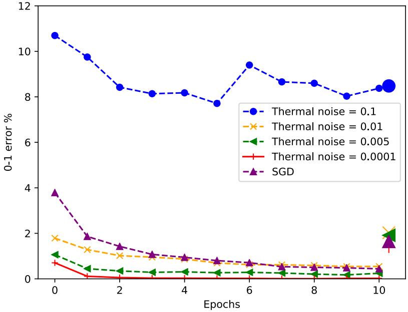
We optimize $F_{\gamma,\tau}(\mathbf{w}; S)$ using SGLD, obtaining (ϵ, δ)-differential privacy.

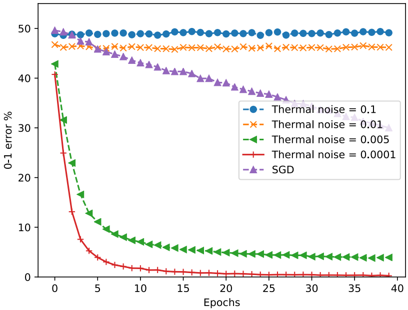
SGLD is known to converge weakly to the ϵ-differentially private exponential mechanism. Our analysis makes a coarse approximation: privacy of SGLD is that of exponential mechanism.

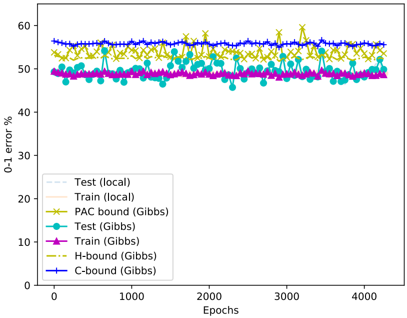


Test (local)
Train (local)
PAC bound (Gibbs)
Test (Gibbs)
Train (Gibbs)
H-bound (Gibbs)
C-bound (Gibbs)
0-1 error %
Epochs
0
2
4
6
8
10
12
14
0
100
200
300
400
500


Test (local)
Train (local)
PAC bound (Gibbs)
Test (Gibbs)
Train (Gibbs)
H-bound (Gibbs)
C-bound (Gibbs)
0-1 error %
Epochs


Thermal noise = 0.1
Thermal noise = 0.01
Thermal noise = 0.005
Thermal noise = 0.0001
SGD
0-1 error %
Epochs


Thermal noise = 0.1
Thermal noise = 0.01
Thermal noise = 0.005
Thermal noise = 0.0001
SGD
0-1 error %
Epochs


Test (local)
Train (local)
PAC bound (Gibbs)
Test (Gibbs)
Train (Gibbs)
H-bound (Gibbs)
C-bound (Gibbs)
0-1 error %
Epochs

Conclusion

- We show that the size/flatness/location of minima (that were found by SGD on MNIST) imply generalization using PAC-Bayes bounds;
- We show Entropy-SGD optimizes the prior in a PAC-Bayes bound, which is not valid;
- We give a differentially private version of PAC-Bayes theorem and modify Entropy-SGD so that prior is privately optimized.

Achille, A. and S. Soatto (2017). "On the Emergence of Invariance and Disentangling in Deep Representations". *arXiv preprint arXiv:1706.01350*.

Catoni, O. (2007). *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*. arXiv: 0712.0248 [stat.ML].

Chaudhari, P., A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, and R. Zecchina (2017). "Entropy-SGD: Biasing Gradient Descent Into Wide Valleys". In: *International Conference on Learning Representations (ICLR)*. arXiv: 1611.01838v4 [cs.LG].

Hinton, G. E. and D. van Camp (1993). "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. Santa Cruz, California, USA: ACM, pp. 5–13.

Hochreiter, S. and J. Schmidhuber (1997). "Flat Minima". *Neural Comput.* 9.1, pp. 1–42.

Keskar, N. S., D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang (2017). "On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima". In: *International Conference on Learning Representations (ICLR)*. arXiv: 1609.04836v2 [cs.LG].

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

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

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