Improving PAC-Bayes bounds for neural nets

Anirbit Mukherjee*[1], Pushpendre Rastogi*[2], Dan Roy**, Jun Yang**

*[1] Applied Mathematics and Statistics, *[2] Computer Science, Johns Hopkins University

** Statistical Sciences, Vector Institute For Artificial Intelligence, University of Toronto

Introduction

Recent works like Dziugaite and Roy 2017 and Zhou et al. 2018 have shown that one can obtain non-vacuous "computational" PAC-Bayesian bounds on the risk of neural nets. In this work we improve on existing theoretical PAC-Bayesian risk bounds by using data-dependent priors sensitive to the geometrical properties of training and by considering carefully chosen clusters of multiple nets in tandem.

Notation

All losses considered are multi-class margin losses. Training data-sets (denoted as S) will be assumed to be of size m and contained in a ball of radius B, centered at the origin in $\mathbb{R}^n$. Define $f_{\mathrm{A}} : \mathbb{R}^n \to \mathbb{R}^k$ to be a depth-d neural-network with maximum width h, whose ℓ^{th} layer has weight matrix A_ℓ. The first $d-1$ layers of f_{A} use the RELU non-linear activation. A is the vector of parameters formed by concatenating vectorized layer matrices, i.e. A denotes the vector $[\mathrm{vec}(\mathrm{A}_1); \ldots; \mathrm{vec}(\mathrm{A}_d)]$ whose each coordinate is a distinct trainable weight in the net.

Define $\beta(\mathrm{A}) = \left(\prod_{\ell=1}^{d} \|A_\ell\|_2\right)^{1/d}$. We will omit the argument A whenever the neural-network under consideration is clear from the context. *Clearly β^d upperbounds the Lipschitz-constant for f_A.*

Let $\{\mathrm{A}_i \in \mathbb{R}^{\dim(\mathrm{A})} \mid i = 1, .., k_1\}$ denote a set of $k_1 \geq 2$ neural net weight vectors on the underlying architecture of f_{A}. Let $\{p_i \mid i = 1, \ldots, k_1\}$ be a set of non-zero scalars s.t $\sum_i p_i = 1$. Define $\beta_i := \beta(\mathrm{A}_i)$.

Controlled output perturbation with a mixture of Gaussians sampling for the weights

Theorem (1)

Suppose the following condition holds for some $\varepsilon > 0$,

$$\forall i \in \{1 \ldots k_1\}, \mathbf{x} \in S \ \|f_{A_i}(\mathbf{x}) - f_A(\mathbf{x})\| \leq \varepsilon \|f_A(\mathbf{x})\|$$

Then for every $\gamma > \varepsilon \max_{\mathbf{x}\in S} \|f_A(\mathbf{x})\|$, we have,

$$\mathbb{P}_{A' \sim MG(posterior)}\left[\max_{\mathbf{x}\in S} \|f_{A'}(\mathbf{x}) - f_A(\mathbf{x})\| > 2\gamma\right] \leq \delta$$

Where MG(posterior)$(\mathbf{w}) = \sum_i p_i \mathcal{N}_{(A_i,\sigma^2)}(\mathbf{w})$
and $\sigma \leq \frac{1}{\sqrt{2h\log\left(\frac{2dhk_1}{\delta}\right)}} \min_{1\leq i\leq k_1} \min\left\{\frac{\beta_i}{d}, \frac{\gamma}{k_1 edB p_i \beta_i^{d-1}}\right\}$.

Now we will use this above noise resilience theorem about nets to get finer data-dependent PAC-Bayesian risk bounds on neural nets.

Defining "Nice" training dataset

We call a training dataset S as $(\varepsilon, \gamma)-$nice w.r.t neural weight vectors A and $\{\mathrm{A}_i\}_{i=1}^{k_1}$ if it satisfies the following two conditions,

1. $\max_{\mathbf{x}\in S} \|f_{\mathrm{A}_i}(\mathbf{x}) - f_{\mathrm{A}}(\mathbf{x})\| \leq \varepsilon \|f_{\mathrm{A}}(\mathbf{x})\|, \forall 1 \leq i \leq k_1$
2. $\gamma > \varepsilon \max_{\mathbf{x}\in S} \|f_{\mathrm{A}}(\mathbf{x})\|$

A two dimensional grid of priors

Let $\Lambda = \{1, \ldots, 314\}$ and choose a regularization parameter $d_{\min} > 0$. For each $\lambda \in \Lambda$ we choose k_1 distinct neural net weights $\{\mathrm{B}_{\lambda,j}\}_{j=1}^{k_1}$ within a conical half-angle of 0.01λ around the initial neural weight B.

For each λ, we construct a grid $\mathcal{B}_\lambda$, (call it the "beta-grid"), containing at most K_1 points in a fixed compact interval in $\mathbb{R}$. (*Neither the grid points nor the interval depend on λ but they do depend on the parameters $B, d, h, \gamma, m, \delta$ and $d_{\min}$*).

For each $\lambda \in \Lambda$ and $\tilde{\sigma} \in \mathcal{B}_\lambda$ we consider the following mixture of Gaussians $\frac{1}{k_1}\sum_{j=1}^{k_1} \mathcal{N}_{(\mathrm{B}_{\lambda,j}, \tilde{\sigma}^2 I)}$. Thus we have a grid of priors of total size $K := 314K_1$.

A new PAC-Bayesian risk bound on nets

Theorem (2)

Suppose we train using a dataset S to obtain the trained net f_A from an initial neural net f_B. Let $\theta = \arccos \frac{\langle A,B\rangle}{\|A\|,\|B\|}$. Let $\lambda^* = \mathrm{argmin}_{\lambda\in\Lambda} |0.01\lambda - \theta|$.

Let the neural weights $\{A_i\}_{i=1,\ldots,k_1}$ be obtained by training the nets $\{f_{B_{\lambda^*,j}}\}_{j=1,\ldots,k_1}$ on S and suppose S is $(\varepsilon,\gamma)-$nice w.r.t $\{A, A_{i=1,\ldots,k_1}\}$

Now for each such i define $d_{i,*} = \min_{j=1,\ldots,k_1} \|A_i - B_{\lambda^*,j}\|^2$ and $\tilde{\beta}_i = \mathrm{argmin}_{x\in\mathcal{B}_{\lambda^*}} |x - \beta(A_i)|$ Then it follows that for all $\varepsilon > 0$ and $\delta \in (0, \frac{1}{K})$,

$$\mathbb{P}_S\left[\exists \tilde{\mu}_A \text{ s.t. } , \forall i \text{ s.t. } d_{i,*} \geq d_{\min}, \mathbb{E}_{A+\tilde{\mathbf{u}}\sim\tilde{\mu}_A}[L_0(f_{A+\tilde{\mathbf{u}}})] \leq \hat{L}_{\frac{\gamma}{2}}(f_A) + \sqrt{\frac{1}{m-1}}\sqrt{-\log\left(\frac{1}{k_1}\sum_{j=1}^{k_1} \exp\left(-\frac{1}{2\tilde{\sigma}_i^2}\|A_i - B_{\lambda^*,j}\|^2\right)\right) + \log\frac{3m}{\delta}}\right] \geq 1 - K\delta \quad (1)$$

where $\tilde{\sigma}_i^2 := \frac{1}{2h\log\left(\frac{2dh}{\delta}\right)}\left(\min\left\{\frac{\tilde{\beta}_i}{de^{\frac{1}{d-1}}}, \frac{\frac{\gamma}{8}}{e^2 dB\tilde{\beta}_i^{d-1}}\right\}\right)^2$

Some salient points about the (main) Theorem (2)

1. The distribution $\tilde{\mu}_{\mathrm{A}}$ is explicitly constructed such that when the noisy weights above $\mathrm{A} + \tilde{\mathbf{u}}$ are sampled from this distribution, it is ensured that $\max_{\mathbf{x}\in S} \|f_{\mathrm{A}+\tilde{\mathbf{u}}}(\mathbf{x}) - f_{\mathrm{A}}(\mathbf{x})\|_\infty < \frac{\gamma}{4}$. Also note that as defined above, λ^* is the closest angle in the set $\{0.01, 0.02, \ldots, 3.14\}$ to the angle between the initial net's weight vector B and the trained net's weight vector A
2. In our experiments the nets $\{\mathrm{A}_i\}_{i=1,\ldots,k_1}$ were obtained by training the nets $\{f_{\mathrm{B}_{\lambda^*,j}}\}_{j=1,\ldots,k_1}$ on S by the same algorithm by which we obtained f_{A} from f_{B}. A direction to explore further is if (maybe data-dependent) compression techniques can be used to obtain particularly good clusters of $\{B_{\lambda^*,j}\}$ from B and $\{\mathrm{A}_i\}$ from A.

Comparison with Neyshabur et al. 2017 on CIFAR-10

In the following figure we see examples of how OUR bounds do better than NBS (Neyshabur et al. 2017) . We plot OUR bound for the i^{th} point in the final cluster (as defined in equation 1) that achieves the lowest bound. Note the log-scale in the $y-$axis in this figure and hence the relative advantage of our bound is a significant multiplicative factor which is *increasing* with depth. *And at large widths our bound seems to essentially flatten out in its depth dependence.*

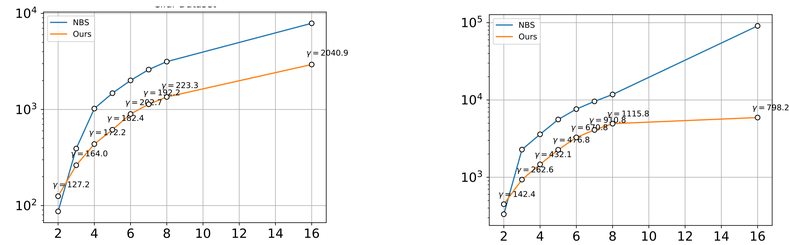
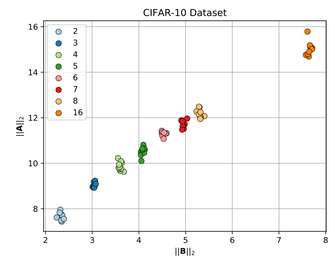
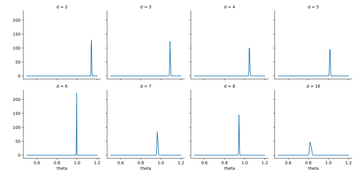


Figure: Net width is 100 on left and 400 on the right.

Demonstration of stability of the relevant geometrical properties of training

In the *figure to the left* we plot the initial parameter norm $\|\mathrm{B}\|_2$ vs the final parameter norm $\|\mathrm{A}\|_2$ for training nets of different depths at width 100 on the CIFAR-10 dataset. Thus its demonstrated that the multiplicative factor with which the norm increases during training is fairly stable to architectural changes.





In the *figure to the right* is displayed the Gaussian kernel density estimate on 10 trials of the experiment (at different depths and 100 width) measuring the angular deviation of the weight vector under training. This deflection angle only decreases slightly with depth and showed negligible dependence over the widths tried.

References

Dziugaite, Gintare Karolina and Daniel M Roy (2017). "Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data". In: *arXiv preprint arXiv:1703.11008*.

Neyshabur, Behnam et al. (2017). "A pac-bayesian approach to spectrally-normalized margin bounds for neural networks". In: *arXiv preprint arXiv:1707.09564*.

Zhou, Wenda et al. (2018). "Non-vacuous generalization bounds at the imagenet scale: a PAC-bayesian compression approach". In: *arXiv preprint arXiv:1804.05862*.