A PAC-Bayesian Approach to Spectrally-Normalized Margin Bounds for Neural Networks

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, Nathan Srebro
Toyota Technological Institute at Chicago
{bneyshabur, srinadh, mcallester, nati}@ttic.edu

Abstract

We present a generalization bound for feedforward neural networks in terms of the product of the spectral norm of the layers and the Frobenius norm of the weights. The generalization bound is derived using a PAC-Bayes analysis.

1 Introduction

In this note we present and prove a margin based generalization bound for feedforward neural networks, that depends on the product of the spectral norm of the weights in each layer, as well as the Frobenius norm of the weights.

Our generalization bound shares much similarity with a margin based generalization bound recently presented by Bartlett et al. [1]. Both bounds depend similarly on the product of the spectral norms of each layer, multiplied by a factor that is additive across layers. In addition, Bartlett et al.'s [1] bound depends on the elementwise ℓ_1-norm of the weights in each layer, while our bound depends on the Frobenius (elementwise ℓ_2) norm of the weights in each layer, with an additional multiplicative dependence on the "width". The two bounds are thus not directly comparable, and each one dominates in a different regime, roughly depending on the sparsity of the weights.

More importantly, our proof technique is entirely different, and arguably simpler, than that of Bartlett et al. [1]. We derive our bound using PAC-Bayes analysis, and more specifically a generic PAC-Bayes margin bound (Lemma 1). The main ingredient is a perturbation bound (Lemma 2), bounding the changes in the output of a network when the weights are perturbed, in terms of the product of the spectral norm of the layers. This is an entirely different analysis approach from Bartlett et al.'s [1] covering number analysis. We hope our analysis can give more direct intuition into the different ingredients in the bound and will allow modifying the analysis, e.g. by using different prior and perturbation distributions in the PAC-Bayes bound, to obtain tighter bounds, perhaps with dependence on different layer-wise norms.

We note that prior bounds in terms of elementwise or unit-wise norms (such as the Frobenius norm and elementwise ℓ_1 norms of layers), without a spectral norm dependence, all have a multiplicative dependence across layers or exponential dependence on depth Bartlett and Mendelson [3], Neyshabur et al. [11], or are for constant depth networks Bartlett [2]. Here only the spectral norm is multiplied across layers, and thus if the spectral norms are close to one, the exponential dependence on depth can be avoided.

1.1 Preliminaries

Consider the classification task with input domain $\mathcal{X}_{B,n} = \left\{\mathbf{x} \in \mathbb{R}^n \mid |\sum_{i=1}^n x_i^2 \leq B^2\right\}$ and output domain $\mathbb{R}^k$ where the output of the model is a score for each class and the class with the maximum score will be selected as the predicted label. Let $f_{\mathbf{w}}(\mathbf{x}) : \mathcal{X}_{B,n} \to \mathbb{R}^k$ be the function computed

by a d layer feed-forward network for the classification task with parameters $\mathbf{w} = \text{vec}\left(\{W_i\}_{i=1}^d\right)$, $f_{\mathbf{w}}(\mathbf{x}) = W_d\,\phi(W_{d-1}\,\phi(....\phi(W_1\mathbf{x})))$, here ϕ is the ReLU activation function. Let $f^i_{\mathbf{w}}(\mathbf{x})$ denote the output of layer i before activation and h be an upper bound on the number of output units in each layer. We can then define fully connected feedforward networks recursively: $f^1_{\mathbf{w}}(\mathbf{x}) = W_1\mathbf{x}$ and $f^i_{\mathbf{w}}(\mathbf{x}) = W_i\phi(f^{i-1}_{\mathbf{w}}(\mathbf{x}))$. Let $\|.\|_F$, $\|.\|_1$ and $\|.\|_2$ denote the Frobenius norm, the element-wise ℓ_1 norm and the spectral norm respectively. We further denote the ℓ_p norm of a vector by $|.|_p$.

For any distribution $\mathcal{D}$ and margin $\gamma > 0$, we define the expected margin loss as follows:

$$L_\gamma(f_{\mathbf{w}}) = \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left[f_{\mathbf{w}}(\mathbf{x})[y] \leq \gamma + \max_{j\neq y} f_{\mathbf{w}}(\mathbf{x})[j]\right] \tag{1}$$

Let $\widehat{L}_\gamma(f_{\mathbf{w}})$ be the empirical estimate of the above expected margin loss. Since setting $\gamma = 0$ corresponds to the classification loss, we will use $L_0(f_{\mathbf{w}})$ and $\widehat{L}_0(f_{\mathbf{w}})$ to refer to the expected risk and the training error. The loss L_γ defined this way is bounded between 0 and 1.

1.2 PAC-Bayesian framework

The PAC-Bayesian framework [9, 10] provides generalization guarantees for randomized predictors, drawn form a learned distribution Q (as opposed to a learned single predictor) that depends on the training data. In particular, let $f_{\mathbf{w}}$ be any predictor (not necessarily a neural network) learned from the training data and parametrized by $\mathbf{w}$. We consider the distribution Q over predictors of the form $f_{\mathbf{w}+\mathbf{u}}$, where $\mathbf{u}$ is a random variable whose distribution may also depend on the training data. Given a "prior" distribution P over the set of predictors that is independent of the training data, the PAC-Bayes theorem states that with probability at least $1-\delta$ over the draw of the training data, the expected error of $f_{\mathbf{w}+\mathbf{u}}$ can be bounded as follows [8]:

$$\mathbb{E}_{\mathbf{u}}[L_0(f_{\mathbf{w}+\mathbf{u}})] \leq \mathbb{E}_{\mathbf{u}}[\widehat{L}_0(f_{\mathbf{w}+\mathbf{u}})] + 2\sqrt{\frac{2\left(KL\left(\mathbf{w}+\mathbf{u}\|P\right) + \ln\frac{2m}{\delta}\right)}{m-1}}. \tag{2}$$

To get a bound on the expected risk $L_0(f_{\mathbf{w}})$ for a single predictor $f_{\mathbf{w}}$, we need to relate the expected perturbed loss, $\mathbb{E}_{\mathbf{u}}[L_0(f_{\mathbf{w}+\mathbf{u}})]$ in the above equation with $L_0(f_{\mathbf{w}})$. Toward this we use the following lemma that gives a margin-based generalization bound derived from the PAC-Bayesian bound (2):

Lemma 1. *Let $f_{\mathbf{w}}(\mathbf{x}) : \mathcal{X} \to \mathbb{R}^k$ be any predictor (not necessarily a neural network) with parameters $\mathbf{w}$, and P be any distribution on the parameters that is independent of the training data. Then, for any $\gamma, \delta > 0$, with probability $\geq 1-\delta$ over the training set of size m, for any $\mathbf{w}$, and any random perturbation $\mathbf{u}$ s.t. $\mathbb{P}_{\mathbf{u}}\left[\max_{\mathbf{x}\in\mathcal{X}} |f_{\mathbf{w}+\mathbf{u}}(\mathbf{x}) - f_{\mathbf{w}}(\mathbf{x})|_\infty < \frac{\gamma}{4}\right] \geq \frac{1}{2}$, we have:*

$$L_0(f_{\mathbf{w}}) \leq \widehat{L}_\gamma(f_{\mathbf{w}}) + 4\sqrt{\frac{KL\left(\mathbf{w}+\mathbf{u}\|P\right) + \ln\frac{6m}{\delta}}{m-1}}.$$

In the above expression the KL is evaluated for a fixed $\mathbf{w}$ and only $\mathbf{u}$ is random, i.e. the distribution of $\mathbf{w}+\mathbf{u}$ is the distribution of $\mathbf{u}$ shifted by $\mathbf{w}$. The lemma is analogous to similar analysis of Langford and Shawe-Taylor [7] and McAllester [8] obtaining PAC-Bayes margin bounds for linear predictors. As we state the lemma, it is not specific to linear separators, nor neural networks, and holds generally for any real-valued predictor.

We next show how to utilize the above general PAC-Bayes bound to prove generalization guarantees for feedforward networks based on the spectral norm of its layers.

2 Generalization Bound

In this section we present our generalization bound for feedfoward networks with ReLU activations, derived using the PAC-Bayesian framework. Langford and Caruana [6], and more recently Dziugaite and Roy [4] and Neyshabur et al. [12], used PAC-Bayes bounds to analyze generalization behavior in neural networks, evaluating the KL-divergence, "perturbation error" $L[f_{\mathbf{w}+\mathbf{u}}] - L[f_{\mathbf{w}}]$, or the entire bound numerically. Here, we use the PAC-Bayes framework as a tool to analytically derive a margin-based bound in terms of norms of the weights. As we saw in Lemma 1, the key to doing so is bounding the change in the output of the network when the weights are perturbed. In the following lemma, we bound this change in terms of the spectral norm of the layers:

Lemma 2 (Perturbation Bound). *For any $B, d > 0$, let $f_{\mathbf{w}} : \mathcal{X}_{B,n} \to \mathbb{R}^k$ be a d-layer network. Then for any $\mathbf{w}$, and $\mathbf{x} \in \mathcal{X}_{B,n}$, and any perturbation $\mathbf{u} = vec\left(\{U_i\}_{i=1}^d\right)$ such that $\|U_i\|_2 \leq \frac{1}{d}\|W_i\|_2$, the change in the output of the network can be bounded as follows:*

$$|f_{\mathbf{w}+\mathbf{u}}(\mathbf{x}) - f_{\mathbf{w}}(\mathbf{x})|_2 \leq eB\left(\prod_{i=1}^d \|W_i\|_2\right)\sum_{i=1}^d \frac{\|U_i\|_2}{\|W_i\|_2}.$$

Next we use the above perturbation bound and the PAC-Bayes result (Lemma 1) to derive the following generalization guarantee.

Theorem 1 (Generalization Bound). *For any $B, d, h > 0$, let $f_{\mathbf{w}} : \mathcal{X}_{B,n} \to \mathbb{R}^k$ be a d-layer feedforward network with ReLU activations. Then, for any $\delta, \gamma > 0$, with probability $\geq 1-\delta$ over a training set of size m, for any $\mathbf{w}$, we have:*

$$L_0(f_{\mathbf{w}}) \leq \widehat{L}_\gamma(f_{\mathbf{w}}) + \mathcal{O}\left(\sqrt{\frac{B^2 d^2 h \ln(dh)\Pi_{i=1}^d \|W_i\|_2^2 \sum_{i=1}^d \frac{\|W_i\|_F^2}{\|W_i\|_2^2} + \ln\frac{dm}{\delta}}{\gamma^2 m}}\right).$$

Comparing the above result to Bartlett et al.'s [1] boils down to comparing $\sqrt{h}\|W_i\|_F$ with $\|W_i\|_1$. Recalling that W_i is an $h \times h$ matrix, we have that $\|W_i\|_F \leq \|W_i\|_1 \leq h\|W_i\|_F$. When the weights are fairly dense and are of uniform magnitude, the second inequality will be tight, and we will have $\sqrt{h}\|W_i\|_F \ll \|W_i\|_1$, and Theorem 1 will dominate. When the weights are sparse with roughly a constant number of significant weights per unit (i.e. weight matrix with sparsity $\Theta(h)$), the bounds will be similar. Bartlett et al.'s [1] bound will dominate when the weights are extremely sparse, with much fewer significant weights than units, i.e. when most units do not have any incoming or outgoing weights of significant magnitude.

Proof of Theorem 1. The proof involves mainly two steps. In the first step we calculate what is the maximum allowed perturbation of parameters to satisfy a given margin condition γ, using Lemma 2. In the second step we calculate the KL term in the PAC-Bayes bound in Lemma 1, for this value of the perturbation.

Let $\beta = \left(\prod_{i=1}^d \|W_i\|_2\right)^{1/d}$ and consider a network with the normalized weights $\widetilde{W_i} = \frac{\beta}{\|W_i\|_2}W_i$. Due to the homogeneity of the ReLU, we have that for feedforward networks with ReLU activations $f_{\widetilde{\mathbf{w}}} = f_{\mathbf{w}}$, and so the (empirical and expected) loss (including margin loss) is the same for $\mathbf{w}$ and $\widetilde{\mathbf{w}}$. We can also verify that $\left(\prod_{i=1}^d \|W_i\|_2\right) = \left(\prod_{i=1}^d \left\|\widetilde{W_i}\right\|_2\right)$ and $\frac{\|W_i\|_F}{\|W_i\|_2} = \frac{\left\|\tilde{W}_i\right\|_F}{\left\|\tilde{W}_i\right\|_2}$, and so the excess error in the Theorem statement is also invariant to this transformation. It is therefore sufficient to prove the Theorem only for the normalized weights $\tilde{\mathbf{w}}$, and hence we assume w.l.o.g. that the spectral norm is equal across layers, i.e. for any layer i, $\|W_i\|_2 = \beta$.

Choose the distribution of the prior P to be $\mathcal{N}(0, \sigma^2 I)$, and consider the random perturbation $\mathbf{u} \sim \mathcal{N}(0, \sigma^2 I)$, with the same σ, which we will set later according to β. More precisely, since the prior cannot depend on the learned predictor $\mathbf{w}$ or its norm, we will set σ based on an approximation $\tilde{\beta}$. For each value of $\tilde{\beta}$ on a pre-determined grid, we will compute the PAC-Bayes bound, establishing the generalization guarantee for all $\mathbf{w}$ for which $|\beta - \tilde{\beta}| \leq \frac{1}{d}\beta$, and ensuring that each relevant value of β is covered by some $\tilde{\beta}$ on the grid. We will then take a union bound over all $\tilde{\beta}$ on the grid. For now, we will consider a fixed $\tilde{\beta}$ and the $\mathbf{w}$ for which $|\beta - \tilde{\beta}| \leq \frac{1}{d}\beta$, and hence $\frac{1}{e}\beta^{d-1} \leq \tilde{\beta}^{d-1} \leq e\beta^{d-1}$.

Since $\mathbf{u} \sim \mathcal{N}(0, \sigma^2 I)$, we get the following bound for the spectral norm of U_i [**?**]:

$$\mathbb{P}_{U_i \sim N(0,\sigma^2 I)}\left[\|U_i\|_2 > t\right] \leq 2he^{-t^2/2h\sigma^2}.$$

Taking a union bond over the layers, we get that, with probability $\geq \frac{1}{2}$, the spectral norm of the perturbation U_i in each layer is bounded by $\sigma\sqrt{2h\ln(4dh)}$. Plugging this spectral norm bound into Lemma 2 we have that with probability at least $\frac{1}{2}$,

$$\max_{\mathbf{x}\in\mathcal{X}_{B,n}} |f_{\mathbf{w}+\mathbf{u}}(\mathbf{x}) - f_{\mathbf{w}}(\mathbf{x})| \leq eB\beta^d \sum_i \frac{\|U_i\|_2}{\beta} = eB\beta^{d-1}\sum_i \|U_i\|_2 \leq e^2 dB\tilde{\beta}^{d-1}\sigma\sqrt{2h\ln(4dh)} \leq \frac{\gamma}{4},$$

where we choose $\sigma = \frac{\gamma}{42dB\tilde{\beta}^{d-1}\sqrt{h\ln(4hd)}}$ to get the last inequality. Hence, the perturbation $\mathbf{u}$ with the above value of σ satisfies the assumptions of the Lemma 1.

We now calculate the KL-term in Lemma 1 with the chosen distributions for P and $\mathbf{u}$, for the above value of σ.

$$KL(\mathbf{w}+\mathbf{u}||P) \leq \frac{|\mathbf{w}|^2}{2\sigma^2} \leq \mathcal{O}\left(B^2d^2h\ln(dh)\frac{\Pi_{i=1}^d \|W_i\|_2^2}{\gamma^2}\sum_{i=1}^d \frac{\|W_i\|_F^2}{\|W_i\|_2^2}\right).$$

Hence, for any $\tilde{\beta}$, with probability $\geq 1-\delta$ and for all $\mathbf{w}$ such that, $|\beta - \tilde{\beta}| \leq \frac{1}{d}\beta$, we have:

$$L_0(f_{\mathbf{w}}) \leq \widehat{L}_\gamma(f_{\mathbf{w}}) + \mathcal{O}\left(\sqrt{\frac{B^2d^2h\ln(dh)\Pi_{i=1}^d \|W_i\|_2^2 \sum_{i=1}^d \frac{\|W_i\|_F^2}{\|W_i\|_2^2} + \ln\frac{m}{\delta}}{\gamma^2 m}}\right). \tag{3}$$

Finally we need to take a union bound over different choices of $\tilde{\beta}$. Let us see how many choices of $\tilde{\beta}$ we need to ensure we always have $\tilde{\beta}$ in the grid s.t. $|\tilde{\beta}-\beta| \leq \frac{1}{d}\beta$. We only need to consider values of β in the range $\left(\frac{\gamma}{2B}\right)^{1/d} \leq \beta \leq \left(\frac{\gamma\sqrt{m}}{2B}\right)^{1/d}$. For β outside this range the theorem statement holds trivially: Recall that the LHS of the theorem statement, $L_0(f_{\mathbf{w}})$ is always bounded by 1. If $\beta^d < \frac{\gamma}{2B}$, then for any $\mathbf{x}$, $|f_{\mathbf{w}}(\mathbf{x})| \leq \beta^d B \leq \gamma/2$ and therefore $L_\gamma = 1$. Alternately, if $\beta^d > \frac{\gamma\sqrt{m}}{2B}$, then the second term in equation 2 is greater than one. Hence, we only need to consider values of β in the range discussed above. Since we need $\tilde{\beta}$ to satisfy $|\tilde{\beta}-\beta| \leq \frac{1}{d}\beta \leq \frac{1}{d}\left(\frac{\gamma}{2B}\right)^{1/d}$, the size of the cover we need to consider is bounded by $dm^{\frac{1}{2d}}$. Taking a union bound over the choices of $\tilde{\beta}$ in this cover and using the bound in equation (3) gives us the theorem statement. □

Proof of Lemma 2. Let $\Delta_i = \left|f^i_{\mathbf{w}+\mathbf{u}}(\mathbf{x}) - f^i_{\mathbf{w}}(\mathbf{x})\right|_2$. We will prove using induction that for any $i \geq 0$:

$$\Delta_i \leq \left(1+\frac{1}{d}\right)^i \left(\prod_{j=1}^i \|W_j\|_2\right)|\mathbf{x}|_2 \sum_{j=1}^i \frac{\|U_j\|_2}{\|W_j\|_2}.$$

The above inequality together with $\left(1+\frac{1}{d}\right)^d \leq e$ proves the lemma statement. The induction base clearly holds since $\Delta_0 = |\mathbf{x}-\mathbf{x}|_2 = 0$. For any $i \geq 1$, we have the following:

$$\begin{aligned}
\Delta_{i+1} &= \left|(W_{i+1}+U_{i+1})\,\phi_i(f^i_{\mathbf{w}+\mathbf{u}}(\mathbf{x})) - W_{i+1}\phi_i(f^i_{\mathbf{w}}(\mathbf{x}))\right|_2 \\
&= \left|(W_{i+1}+U_{i+1})\left(\phi_i(f^i_{\mathbf{w}+\mathbf{u}}(\mathbf{x})) - \phi_i(f^i_{\mathbf{w}}(\mathbf{x}))\right) + U_{i+1}\phi_i(f^i_{\mathbf{w}}(\mathbf{x}))\right|_2 \\
&\leq (\|W_{i+1}\|_2 + \|U_{i+1}\|_2)\left|\phi_i(f^i_{\mathbf{w}+\mathbf{u}}(\mathbf{x})) - \phi_i(f^i_{\mathbf{w}}(\mathbf{x}))\right|_2 + \|U_{i+1}\|_2\left|\phi_i(f^i_{\mathbf{w}}(\mathbf{x}))\right|_2 \\
&\leq (\|W_{i+1}\|_2 + \|U_{i+1}\|_2)\left|f^i_{\mathbf{w}+\mathbf{u}}(\mathbf{x}) - f^i_{\mathbf{w}}(\mathbf{x})\right|_2 + \|U_{i+1}\|_2\left|f^i_{\mathbf{w}}(\mathbf{x})\right|_2 \\
&= \Delta_i(\|W_{i+1}\|_2 + \|U_{i+1}\|_2) + \|U_{i+1}\|_2\left|f^i_{\mathbf{w}}(\mathbf{x})\right|_2,
\end{aligned}$$

where the last inequality is by the Lipschitz property of the activation function and using $\phi(0)=0$. The ℓ_2 norm of outputs of layer i is bounded by $|\mathbf{x}|_2\,\Pi_{j=1}^i \|W_j\|_2$ and by the lemma assumption we have $\|U_{i+1}\|_2 \leq \frac{1}{d}\|W_{i+1}\|_2$. Therefore, using the induction step, we get the following bound:

$$\begin{aligned}
\Delta_{i+1} &\leq \Delta_i\left(1+\frac{1}{d}\right)\|W_{i+1}\|_2 + \|U_{i+1}\|_2\,|\mathbf{x}|_2\prod_{j=1}^i\|W_j\|_2 \\
&\leq \left(1+\frac{1}{d}\right)^{i+1}\left(\prod_{j=1}^{i+1}\|W_j\|_2\right)|\mathbf{x}|_2\sum_{j=1}^i\frac{\|U_j\|_2}{\|W_j\|_2} + \frac{\|U_{i+1}\|_2}{\|W_{i+1}\|_2}|\mathbf{x}|_2\prod_{j=1}^{i+1}\|W_i\|_2 \\
&\leq \left(1+\frac{1}{d}\right)^{i+1}\left(\prod_{j=1}^{i+1}\|W_j\|_2\right)|\mathbf{x}|_2\sum_{j=1}^{i+1}\frac{\|U_j\|_2}{\|W_j\|_2}.
\end{aligned}$$

□

References

[1] P. Bartlett, D. J. Foster, and M. Telgarsky. Spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1706.08498*, 2017.

[2] P. L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE transactions on Information Theory*, 44(2):525–536, 1998.

[3] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[4] G. K. Dziugaite and D. M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

[5] N. Harvey, C. Liaw, and A. Mehrabian. Nearly-tight vc-dimension bounds for piecewise linear neural networks. *arXiv preprint arXiv:1703.02930*, 2017.

[6] J. Langford and R. Caruana. (not) bounding the true error. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, pages 809–816. MIT Press, 2001.

[7] J. Langford and J. Shawe-Taylor. Pac-bayes & margins. In *Advances in neural information processing systems*, pages 439–446, 2003.

[8] D. McAllester. Simplified pac-bayesian margin bounds. *Lecture notes in computer science*, pages 203–215, 2003.

[9] D. A. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 230–234. ACM, 1998.

[10] D. A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 164–170. ACM, 1999.

[11] B. Neyshabur, R. Tomioka, and N. Srebro. Norm-based capacity control in neural networks. In *Proceeding of the 28th Conference on Learning Theory (COLT)*, 2015.

[12] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro. Exploring generalization in deep learning. *arXiv preprint arXiv:1706.08947*, 2017.