Generalization Bounds for Rank-Sparse Neural Networks.

Antoine Ledent[1], Rodrigo Alves[2] and Yunwen Lei[3]

[1] Singapore Management University (SMU)
[2] Czech Technical University in Prague (CTU)
[3] The University of Hong Kong (HKU)

Fully-connected Neural Networks: Definitions I

- Consider fully connected neural networks of the form

$$x \to F_A(x) := A_L \sigma_L(A^{L-1} \sigma_{L-1}(\dots \sigma_1(A^1 x) \dots)). \tag{1}$$

- Here σ_ℓ denotes the (elementwise) activation function at layer ℓ (e.g. Relu), and $F_A(x) \in \mathbb{R}^{\mathcal{C}}$ are scores for each of $\mathcal{C}$ classes.
- Given an i.i.d. training set $(x_1, y_1), \dots, (x_N, y_N)$, we are interested in high-probability bounds on the generalization gap

$$\mathbb{E}_{x,y}(\mathrm{l}(F_A(X))) - \frac{1}{N} \sum_{i=1}^{N} \mathrm{l}(F_A(x_i), y_i), \tag{2}$$

where l is a margin-based loss function, e.g.,

$$\mathrm{l}(\hat{y}, y) = \begin{cases} 1, & \text{if } \arg\max_i \hat{y}_i \neq y, \\ 1 - \frac{\hat{y}_y - \max\limits_{i \neq y} \hat{y}_i}{\gamma}, & \text{if } 0 \leq \hat{y}_y - \max\limits_{i \neq y} \hat{y}_i \leq \gamma, \\ 0, & \text{if } \hat{y}_y \geq \max\limits_{i \neq y} \hat{y}_i + \gamma. \end{cases} \tag{3}$$

DNNs and CNNs: Previous Works I

One of the most recognizable generalization bound for fully connected neural networks is that of Bartlett et al. 2017 [SPEC17]

$$\widetilde{O}\left(\ell \frac{1}{\sqrt{N}} \prod_{\ell=1}^{L} \|A_\ell\| \left(\sum_{\ell=1}^{L} \frac{\|(A_\ell - M_\ell)^\top\|_{2,1}^{\frac{2}{3}}}{\|A_\ell\|^{\frac{2}{3}}}\right)^{\frac{3}{2}}\right), \tag{4}$$

where M_ℓ are fixed reference matrices (e.g. initialization).
Slightly weaker result which was independently discovered in Neyshabur et al. 2018 [Ney18]:

$$\widetilde{O}\left(\ell \frac{L\sqrt{W}}{\sqrt{N}} \left(\prod_{\ell=1}^{L} \|A_\ell\|\right) \left(\sum_{\ell=1}^{L} \frac{\|A_\ell - M_\ell\|_{\mathrm{Fr}}^{2}}{\|A_\ell\|_{\sigma}^{2}}\right)^{\frac{1}{2}}\right), \tag{5}$$

where W denotes the width of the network.

DNNs and CNNs: Previous Works II

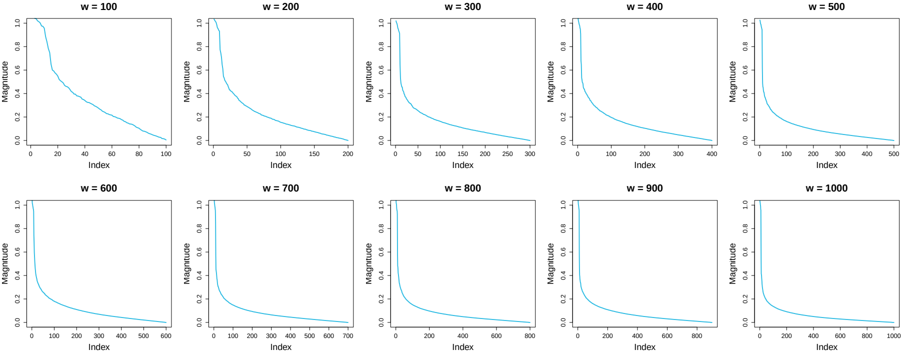
For CNNs (and also DNNs in particular), the following parameter-counting bound (see Long and Sedghi 2020 [LSed20] and Graf et al. 2022 [Graf22]):

$$\widetilde{O}\left(\mathcal{B}\sqrt{\frac{\mathcal{W}\, L}{N}} + \mathcal{B}\sqrt{\frac{\log(1/\delta)}{N}}\right), \tag{6}$$

where $\mathcal{W}$ denotes the total number of parameters.
In [L21c], I have proved the following (norm-based) bounds for CNNs:

$$\widetilde{O}\left(\frac{\ell\sqrt{L}\prod_{\ell=1}^{L}\|\operatorname{op}(A_\ell)\|}{\sqrt{N}}\left[\sum_{\ell=1}^{L}\left(\frac{U_\ell w_\ell \|[A_\ell - M_\ell]^\top\|_{2,1}}{\|\operatorname{op}(A_\ell)\|}\right)^{\frac{2}{3}}\right]^{\frac{3}{2}}\right). \tag{7}$$

Spectral Decay in Trained Neural Networks



→ The effective (soft) rank converges to a “bottleneck rank” determined by the complexity of the data. See [JAC23,JAC24].

→ How can we capture the implications in terms of sample complexity?

[JAC23] Arthur Jacot, ’Implicit Bias of Large Depth Networks: a Notion of Rank for Nonlinear Functions’, NeurIPS 2023.

[JAC4] Zihan Wang, Arthur Jacot, ‘Implicit bias of SGD in L2-regularized linear DNNs: One-way jumps from high to low rank’

Linear Networks

We begin with Linear Networks (without activations) as a toy example.

- Consider linear maps $Z \in \mathbb{R}^{\mathcal{C}\times d}$ in a $\mathcal{C}$-output problem with an L^∞ loss function $\mathrm{l} : \mathbb{R}^{\mathcal{C}} \times \mathcal{Y} \to \mathbb{R}^+$. We are given N training samples $(x_1, y_1), \ldots, (x_N, y_N)$.
- Theorem (Dai, Liu and Srebro 2021, [DS21]) for any matrix A,

$$\begin{aligned} &\min_{B_1,\ldots,B_L} \sum_{k=1}^{L} \|B_k\|_{\mathrm{Fr}}^2 \\ &\text{subject to } A = B_L B_{L-1} \ldots B_1, \\ &= L\|A\|_{\mathrm{sc},2/L}^{2/L}. \end{aligned} \tag{8}$$

- Hence, imposing a weight decay regularizer on $B_1, \ldots, B_L$ is equivalent to imposing a Schatten quasi norm constraint on the full predictor $A = B_L B_{L-1} \ldots B_1$.

Complexity of Linear Classification: Existing Works I

- Consider linear maps $A \in \mathbb{R}^{\mathcal{C} \times d}$ in a $\mathcal{C}$-class classification problem with constraint $\|A\|_{\mathrm{Fr}}^2 \leq a^2$, and an L^∞-Lipschitz loss.
- It is known from norm-based results [Lei et al. TIT'19] that the generalization error can be bounded with high probability by

$$\mathrm{GAP} \leq \widetilde{O}\left(\sqrt{\frac{a^2}{N}}\right) \tag{9}$$

- With a parameter counting involving a low rank representation of Z with rank r, we can obtain the following novel bound which assumes there is low-rank structure over the classes:

$$\mathrm{GAP} \leq \widetilde{O}\left(\sqrt{\frac{[\mathcal{C}+d]r}{N}}\right). \tag{10}$$

- In this work, we *interpolate* between the regime in equations (10) and (9) to obtain bounds which capture the approximate low-rank structure in low Schatten quasi norm matrices.

Generalization Bounds for Linear Networks

Theorem

With probability greater than $1-\delta$ *over the draw of the training set, set* $\mathfrak{B} = \sup_{i=1}^{N} \|x_i\|$, *we have* $\mathrm{GAP} - O\left(\mathcal{B}\sqrt{\frac{\log(1/\delta)}{N}}\right) \leq$

$$\widetilde{O}\left(\sqrt{\frac{[\ell\,\mathfrak{B}]^{\frac{2p}{2+p}} \|A\|_{\mathrm{sc},p}^{\frac{2p}{2+p}} \min(\mathcal{C},d)^{\frac{p}{p+2}} \max(\mathcal{C},d)^{\frac{2}{p+2}}}{N}}\right)$$

$$\leq \widetilde{O}\left(\ell^{\frac{1}{L+1}} \sqrt{\mathfrak{B}^{\frac{2}{L+1}} \frac{\left[\sum_{\ell=1}^{L} \|B_\ell\|_{\mathrm{Fr}}^2\right]^{\frac{L}{L+1}} [\mathcal{C}+d]}{N L^{\frac{L}{L+1}}}}\right)$$

- As $p \to 0$, we expect $\|A\|^p \to r$, thus, the sample complexity scales as $\max(\mathcal{C},d)r$, as expected from a parameter counting argument.

Fully-connected Neural Networks: Results

Theorem

Fix reference matrices $M_1, \ldots, M_L$. The following holds w.h.p. simultaneously over all values of $p_\ell \in [0,2]$ for $\ell = 1, \ldots, L$:

$$\mathrm{GAP} \leq \widetilde{O}\left(\mathcal{B}\sqrt{\frac{\log(1/\delta)}{N}} + \mathcal{B}\sqrt{\frac{L^3}{N}} + \mathcal{B}\left[L\,\ell\right]^{\frac{p}{2+p}}\sqrt{\frac{L}{N}}\mathcal{R}_{F_A}\right),$$

where

$$\mathcal{R}_{F_A} := \left[\sum_{\ell=1}^{L}\left[\mathfrak{B}\prod_{i=1}^{L}\rho_i\|A_i\|\right]^{\frac{2p_\ell}{p_\ell+2}} \times \left[\frac{\|A_\ell - M_\ell\|_{\mathrm{sc},p_\ell}{}^{p_\ell}}{\|A_\ell\|^{p_\ell}}\right]^{\frac{2}{p_\ell+2}}\left[w_\ell + w_{\ell-1}\right]^{1+\frac{p_\ell}{p_\ell+2}}\right]^{\frac{1}{2}} \tag{11}$$

and $p := \max_{\ell=1}^{L} p_\ell$ denotes the maximum of all the indices p_ℓ.

Comments

- We rely on the parametric interpolation strategy for each individual layer. This strategy was originally introduced in the context of Matrix Completion in our earlier work.
- The bounds are posthoc with respect to the choice of p_ℓ.
- The bounds are a hybrid between norm-based and parameter-counting bounds. Hence the non-trivial dependence on margin and scaling parameters.
- The results can be extended to convolutional neural networks, taking weight sharing into account.
- We can also replace the course estimate $\prod_{i=1}^{\ell} \|\operatorname{op}(A_i)\| \|x\|$ by an empirical estimate of the maximum norm of the activations at layer ℓ by using Lipschitz augmentation.

Generalization Bounds with Loss Augmentation (CNNs)

Theorem

The generalization gap is bounded by:

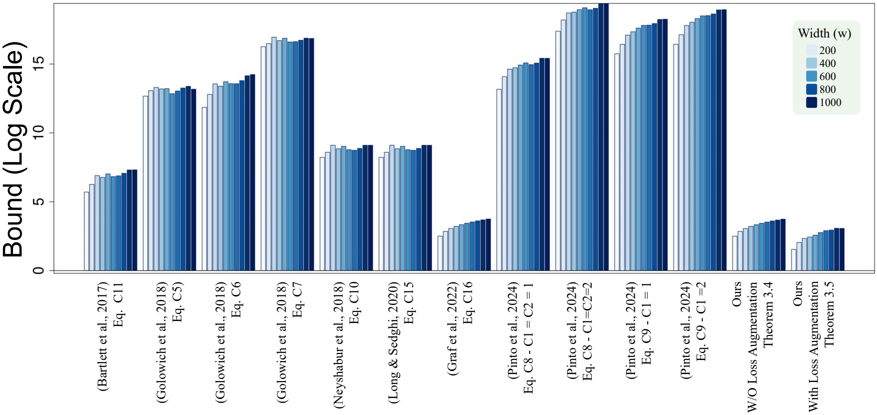
$$\widetilde{O}\left(\mathcal{B}\sqrt{\frac{\log(1/\delta)}{N}} + \mathcal{B}\sqrt{\frac{L^3}{N}} + \mathcal{B}\,[L\,\ell]^{\frac{p}{2+p}}\sqrt{\frac{L}{N}}\mathcal{R}_{F_A}^{emp,\mathfrak{C}}\right),$$

where $\mathcal{R}_{F_A}^{emp,\mathfrak{C}} :=$

$$\left[\sum_{\ell=1}^{L}\left[\mathfrak{B}_{\ell-1,A}^{\mathfrak{C}}\prod_{i=\ell}^{L}\rho_i\|\operatorname{op}(A_i)\|\right]^{\frac{2p_\ell}{p_\ell+2}} \times \left[\frac{\|A_\ell - M_\ell\|_{\mathrm{sc},p_\ell}{}^{p_\ell}}{\|\operatorname{op}(A_\ell)\|^{p_\ell}}\right]^{\frac{2}{p_\ell+2}}[U_\ell + d_{\ell-1}]W_\ell^{\frac{p_\ell}{p_\ell+2}}\right]^{\frac{1}{2}}, \tag{12}$$

where $\mathfrak{B}_{\ell-1,A}^{\mathfrak{C}} := \max\left(\max_{i\leq N,o}\|[F_{A^1,\ldots,A^{\ell-1}}(x_i)]_{S_{\ell-1,o}}\|, 1\right)$.
Here, $W_\ell := U_\ell \times w_\ell$ *(for* $\ell \neq L$*) and* $W_L := 1$.

Experimental Behavior on Real Data



- We observe much more moderate growth with overparametrization compared to the literature.
- → the bounds are able to capture unused capacity due to **overparametrization**.

References

[LSed20] Long, P. M., and Sedghi, H. Generalization bounds for deep convolutional neural networks. ICLR 2020.
[PISIER80] Pisier, G. Remarques sur un résultat non publié de B. Maurey. Séminaire Analyse Fonctionnelle 1980.
[SPEC17] Bartlett, P. L., Foster, D. J., and Telgarsky, M. Spectrally-normalized margin bounds for neural networks. NeurIPS 2017.
[ZHANG02] Zhang, T. Covering number bounds of certain regularized linear function classes. JMLR 2002.
[L21a] Ledent, A., Lei, Y., Wu, L., and Kloft, M. Fine-grained generalization analysis of inductive matrix completion. NeurIPS 2021.
[DS21] Dai, H., Liu, H., and Srebro, N. Representation costs of linear neural networks. NeurIPS 2021.
[SRE11] Srebro, N., Rennie, J., and Jaakkola, T. Rank, max norm, trace norm. COLT 2011.
[L24] Ledent, A., and Alves, G. Generalization analysis of deep nonlinear matrix completion. ICML 2024.
[JAC23] Jacot, A., Bocquet, O., and Rotskoff, G. Implicit bias of large depth networks: A notion of rank for nonlinear functions. NeurIPS 2023.
[JAC24] Wang, Z., and Jacot, A. Implicit bias of SGD in L2-regularized linear DNNs: One-way jumps from high to low rank. Preprint 2024.
[YL21] Lei, Y., Dogan, U., Zhou, D.-X., and Kloft, M. Data-dependent generalization bounds for multi-class classification. TIT 2021.
[W21] Wu, L., Ledent, A., Lei, Y., and Kloft, M. Fine-grained generalization analysis of vector-valued learning. AAAI 2021.
[L21b] Ledent, A., Mustafa, W., Lei, Y., and Kloft, M. Norm-based generalization bounds for multi-class convolutional neural networks. AAAI 2021.
[Ney18] Neyshabur, B., Bhojanapalli, S., and Srebro, N. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. ICLR 2018.
[Graf22] Graf, F., Zeng, S., Rieck, B., Niethammer, M., and Kwitt, R. On measuring excess capacity in neural networks. NeurIPS 2022.
[NK19] Nagarajan, V., and Kolter, J. Z. Uniform convergence may be unable to explain generalization in deep learning. NeurIPS 2019.
[WM19] Wei, C., and Ma, T. Data-dependent sample complexity of deep neural networks via Lipschitz augmentation. NeurIPS 2019.