Proceedings of Machine Learning Research vol 75:1–3, 2018 31st Annual Conference on Learning Theory

Size-Independent Sample Complexity of Neural Networks (Extended Abstract)

Noah Golowich NGOLOWICH@COLLEGE.HARVARD.EDU
Harvard University

Alexander Rakhlin RAKHLIN@MIT.EDU
MIT

Ohad Shamir OHAD.SHAMIR@WEIZMANN.AC.IL
Weizmann Institute of Science
and Microsoft Research

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

We study the sample complexity of learning neural networks, by providing new bounds on their Rademacher complexity assuming norm constraints on the parameter matrix of each layer. Compared to previous work, these complexity bounds have improved dependence on the network depth, and under some additional assumptions, are fully independent of the network size (both depth and width). These results are derived using some novel techniques, which may be of independent interest[1].

Keywords: Neural Networks, Deep Learning, Sample Complexity, Rademacher Complexity

One of the major challenges involving neural networks is explaining their ability to generalize well, even if they are very large and have the potential to overfit the training data (Neyshabur et al., 2014; Zhang et al., 2016). Learning theory teaches us that this must be due to some inductive bias, which constrains one to learn networks of specific configurations (either explicitly, e.g., via regularization, or implicitly, via the algorithm used to train them). However, understanding the nature of this inductive bias is still largely an open problem.

In our work, we consider whether it is possible to prove sample complexity bounds for neural networks, which are not strongly dependent on the network size, under suitable norm constraints. Such bounds exist for linear predictors (or equivalently, one-layer networks), but for networks with more layers, existing results strongly depend on the number of layers, sometimes exponentially, regardless of the norms of the parameter matrices (e.g. Anthony and Bartlett (2009); Neyshabur et al. (2015); Bartlett et al. (2017); Neyshabur et al. (2017)). We make the following contributions:

- We show that the exponential depth dependence in Rademacher complexity-based analysis (e.g. Neyshabur et al. (2015)) can be avoided by applying contraction to a slightly different object than what has become standard since the work of Bartlett and Mendelson (2002). For example, for networks with d layers, where each layer j has a parameter matrix with Frobenius norm at most $M_F(j)$, and m i.i.d. training examples, one can prove a generalization bound

1. This paper is an extended abstract. The full version appears as arXiv preprint 1712.06541 v3

of $\mathcal{O}\left(\sqrt{d}\left(\prod_{j=1}^{d} M_F(j)\right)/\sqrt{m}\right)$. The technique can also be applied to other types of norm constraints. For example, if we consider networks where the 1-norm of each row of the j-th parameter matrix is at most $M(j)$, we attain a bound of $\mathcal{O}\left(\sqrt{d+\log(n)}\left(\prod_{j=1}^{d} M(j)\right)/\sqrt{m}\right)$, where n is the input dimension. Again, the dependence on d is polynomial and quite mild.

- We develop a generic technique to convert depth-dependent bounds to depth-independent bounds, assuming some control over any Schatten norm of the parameter matrices (which includes, for instance, the Frobenius norm and the trace norm as special cases). The key observation we utilize is that the prediction function computed by such networks can be approximated by the composition of a shallow network and univariate Lipschitz functions. For example, again assuming that the Frobenius norms of the layers are bounded by $M_F(1), \ldots, M_F(d)$, we can further improve the result above to

$$\tilde{\mathcal{O}}\left(\left(\prod_{j=1}^{d} M_F(j)\right) \cdot \min\left\{\sqrt{\frac{\log\left(\frac{1}{\Gamma}\prod_{j=1}^{d} M_F(j)\right)}{\sqrt{m}}}, \sqrt{\frac{d}{m}}\right\}\right), \tag{1}$$

  where Γ is a lower bound on the product of the *spectral* norms of the parameter matrices (note that $\Gamma \leq \prod_j M_F(j)$ always). Assuming that $\prod_j M_F(j) \leq R$ for some R, this can be upper bounded by $\tilde{\mathcal{O}}(R\sqrt{\log(R/\Gamma)/\sqrt{m}})$, which to the best of our knowledge, is the first explicit bound for standard neural networks which is fully size-independent, assuming only suitable norm constraints. We also apply this technique to get a depth-independent version of the bound in (Bartlett et al., 2017): Specifically, if we assume that the spectral norms satisfy $\|W_j\| \leq M(j)$ for all j, and $\max_j \frac{\|W_j^T\|_{2,1}}{\|W_j\|} \leq L$, then the bound in provided by Bartlett et al. (2017) becomes $\tilde{\mathcal{O}}\left(BL\prod_{j=1}^{d} M(j) \cdot \sqrt{d^3/m}\right)$. In contrast, we show the following bound for any $p \geq 1$ (ignoring some lower-order logarithmic factors):

$$\tilde{\mathcal{O}}\left(BL\prod_{j=1}^{d} M(j) \cdot \min\left\{\frac{\log\left(\frac{1}{\Gamma}\prod_{j=1}^{d} M_p(j)\right)^{\frac{1}{2}+p}}{m^{\frac{1}{2+3p}}}, \sqrt{\frac{d^3}{m}}\right\}\right),$$

  where $M_p(j)$ is an upper bound on the Schatten p-norm of W_j, and Γ is a lower bound on $\prod_{j=1}^{d}\|W_j\|$. Again, by upper bounding the $\min$ by its first argument, we get a bound independent of the depth d, assuming the norms are suitably constrained.

- We provide a lower bound, showing that for any p, the class of depth-d, width-h neural networks, where each parameter matrix W_j has Schatten p-norm at most $M_p(j)$, can have Rademacher complexity of at least

$$\Omega\left(\frac{B\prod_{j=1}^{d} M_p(j) \cdot h^{\max\left\{0,\frac{1}{2}-\frac{1}{p}\right\}}}{\sqrt{m}}\right).$$

  This somewhat improves on Bartlett et al. (2017, Theorem 3.6), which only showed such a result for $p = \infty$ (i.e. with spectral norm control), and without the h term. For $p = 2$, it matches the upper bound in Eq. (1) in terms of the norm dependencies and B. Moreover, it establishes that

controlling the spectral norm alone (and indeed, any Schatten p-norm control with $p > 2$) cannot lead to bounds independent of the size of the network.

References

Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

Peter Bartlett, Dylan J Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1706.08498*, 2017.

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

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401, 2015.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.