Gradient Descent for One-Hidden-Layer Neural Networks: Polynomial Convergence and SQ Lower Bounds

Santosh Vempala VEMPALA@GATECH.EDU *Georgia Institute of Technology*

John Wilmes WILMES@BRANDEIS.EDU *Brandeis University*

Editors: Alina Beygelzimer and Daniel Hsu

[1]Abstract

We study the complexity of training neural network models with one hidden nonlinear activation layer and an output weighted sum layer. We analyze Gradient Descent applied to learning a bounded target function on n real-valued inputs. We give an agnostic learning guarantee for GD: starting from a randomly initialized network, it converges in mean squared loss to the minimum error (in 2-norm) of the best approximation of the target function using a polynomial of degree at most k. Moreover, for any k, the size of the network and number of iterations needed are both bounded by $n^{O(k)} \log(1/\varepsilon)$. The core of our analysis is the following existence theorem, which is of independent interest: for any $\varepsilon > 0$, any bounded function that has a degree k polynomial approximation with error ε_0 (in 2-norm), can be approximated to within error $\varepsilon_0 + \varepsilon$ as a linear combination of $n^{O(k)} \cdot \text{poly}(1/\varepsilon)$ *randomly chosen* gates from any class of gates whose corresponding activation function has nonzero coefficients in its harmonic expansion for degrees up to k. In particular, this applies to training networks of unbiased sigmoids and ReLUs. We also rigorously explain the empirical finding that gradient descent discovers lower frequency Fourier components before higher frequency components.

We complement this result with nearly matching lower bounds in the Statistical Query model. GD fits well in the SQ framework since each training step is determined by an expectation over the input distribution. We show that any SQ algorithm that achieves significant improvement over a constant function with queries of tolerance some inverse polynomial in the input dimensionality n must use $n^{\Omega(k)}$ queries even when the target functions are restricted to a set of $n^{O(k)}$ degree-k polynomials, and the input distribution is uniform over the unit sphere; for this class the information-theoretic lower bound is only $\Theta(k \log n)$.

Our approach for both parts is based on spherical harmonics. We view gradient descent as an operator on the space of functions, and study its dynamics. An essential tool is the Funk-Hecke theorem, which explains the eigenfunctions of this operator in the case of the mean squared loss.

1. Introduction

It is well known that artificial neural networks (NNs) can approximate any real-valued function. Fundamental results Hornik et al. (1989); Cybenko (1989); Barron (1993) show that a NN with a *single* hidden layer provides a universal representation up to arbitrary approximation, with the number of hidden units needed depending on the function being approximated and the desired accuracy. In practice, NNs today effectively capture a wide variety of information with remarkably accurate predictions.

[1] Extended Abstract. Full version appears on the arXiv as [https://arxiv.org/abs/1805.02677, v3]

© 2019 S. Vempala & J. Wilmes.

Besides their generality, an important feature of NNs is the ease of training them — gradient descent (GD) is used to minimize the error of the network, measured by a loss function of the current weights. This seems to work across a range of labeled data sets. Yet despite its tremendous success, there is no satisfactory explanation for the efficiency or effectiveness of this generic training algorithm.

In this paper we give nearly matching upper and lower bounds that help explain the phenomena seen in practice when training NNs. The upper bounds are for GD and the lower bounds are for all statistical query algorithms.

We consider NNs with n-dimensional inputs, a single hidden layer with m units having some nonlinear activation $\phi : \mathbb{R} \to \mathbb{R}$, and a single linear output unit. All units are without additive bias terms. We will consider inputs drawn from the uniform distribution on S^{n-1}. We initialize our NNs by choosing the vectors for each hidden-layer unit uniformly and indpenently from S^{n-1} and setting all output-layer weights to 0. The specific GD procedure we consider is as follows: in each iteration, the gradient of the loss function is computed using a finite sample of examples, with the entire sample reused for each iteration. The output-layer weights are then modified by adding a fixed multiple of the estimated gradient, and the hidden-layer weights are kept fixed.

Our algorithmic result is an agnostic upper bound on the approximation error and time and sample complexity of GD with the standard mean squared loss function. Despite training only the output layer weights, our novel proof techniques avoid using any convexity in the problem. Since our analysis does not rely on reaching a global minimum, there is reason to hope the techniques will extend to nonconvex settings where we can in general expect only to find a local minimum. Prior results along this line were either for more complicated algorithms or more restricted settings; the closest is the work of Andoni et al. Andoni et al. (2014) where they assume the target function is a bounded degree polynomial. We illustrating the power of our proof technique by obtaining, as an immediate corollary of our convergence anaalysis, a rigorous proof of the "spectral bias" of gradient descent observed experimentally in Rahaman et al. (2018).

The upper bound shows that to get close to the best possible degree k polynomial approximation of the data, it suffices to run GD on a NN with $n^{O(k)}$ units, using the same number of samples. It suffices to train the output layer weights alone. This is an agnostic guarantee. We prove a matching lower bound for solving this polynomial learning problem over the uniform distribution on the unit sphere, for *any* statistical query algorithm that uses tolerance inversely proportional to $n^{\Omega(k)}$. Thus, for this general agnostic learning problem, GD is as good as it gets.

Acknowlegments

The authors are grateful to Adam Kalai and Le Song for helpful discussions. The authors also thank Joël Bellaïche and the anonymous referees for careful reading and many suggestions that improved the presentation. This work was supported in part by NSF grants CCF-1563838, CCF-1717349 and E2CDA-1640081.

References

Alexandr Andoni, Rina Panigrahy, Gregory Valiant, and Li Zhang. Learning polynomials with neural networks. In *International Conference on Machine Learning*, pages 1908–1916, 2014.

Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)*, 2(4):303–314, 1989.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Nasim Rahaman, Devansh Arpit, Aristide Baratin, Felix Draxler, Min Lin, Fred A Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of deep neural networks. *arXiv preprint arXiv:1806.08734*, 2018.