Gradient Descent Provably Optimizes Over-parametrized Neural Networks – Paper Review

William McCloskey

April 6, 2022

1 Problem Statement

This review covers the October 2018 paper Gradient Descent Provably Optimizes Over-parametrized Neural Networks by Du et al. The paper fits into a body of research investigating first order methods for neural networks. Empirically, neural networks train well using gradient descent. From a theoretical perspective, however, the loss of a neural network is a complicated non-convex (and potentially non-smooth) function, and results for using gradient descent to optimize this loss have been unsatisfactory.

The authors provide a number of related results in the literature. Several authors have proven that the loss satisfies nice landscape properties such as all local minima are global. The most related result is by Li and Liang, who show that a two-layer over-parametrized neural network with ReLU activation and cross-entropy loss can achieve accuracy ϵ with $m = \text{poly}(1/\epsilon)$ hidden nodes. This result does not achieve zero training loss unless $m \to \infty$. Indeed, no prior paper was able to justify directly the empirical observation by Zhang et al. that a neural network can achieve zero training loss on randomly labeled (non-degenerate) data via first-order optimization methods.

Du et al. seek to provide such a justification. Like Li and Liang, they consider a two-layer neural network with ReLU activation

$$f(W, a, x) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \sigma(w_r^\top x).$$

However, instead of the cross-entropy loss they use quadratic loss, so their training loss is

$$L(W) = \sum_{i=1}^{n} \frac{1}{2} (f(W, a, x_i) - y_i)^2.$$

They want to prove that gradient descent converges to a global minimum for sufficiently large m (i.e. for a sufficiently overparametrized network).

Such a result is important for a number of reasons: it provides a theoretical justification for a simple empirical observation; it is potentially a stepping stone to studying gradient descent for deeper networks; and it may lead to insights for other similar first order optimization algorithms, such as gradient descent with variable learning rate or accelerated

algorithms like RMSProp or Adam. However, the objective function is non-smooth and non-convex and thus quite challenging to optimize.

2 Main Result

The main result provides convergence of gradient descent to a global minimum with high probability. We restate it here. Assume the following:

Assumptions

- For all $i \in [n]$, $\|x_i\|_2 = 1, |y_i| \leq C$ for some constant C.
- The matrix $H^\infty \in \mathbb{R}^{n \times n}$ with $H^\infty_{i,j} = \mathbb{E}_{w \sim N(0,I)}(x_i^\top x_j \mathbb{1}\{w^\top x_i \geq 0, w^\top x_j \geq 0\})$ satisfies $\lambda_{\min}(H^\infty) \triangleq \lambda_0 > 0$.

Next, suppose the network and optimization is set up as follows:

Setup

- For initialization, we have $w_r \sim N(0, I)$ and $a_r \sim \text{unif}[\{-1, 1\}]$.
- The network has $m = \Omega\left(\frac{n^6}{\lambda_0^2}\right)$ hidden nodes.
- The gradient descent step size is $\eta = O\left(\frac{\lambda_0}{n^2}\right)$.

Then with high probability over the initialization,

$$\|u(k) - y\|_2^2 \leq (1 - \frac{\eta\lambda_0}{2})^k \|u(0) - y\|_2^2$$

for all k, where $u(k)$ is the prediction vector after k iterations.

The conclusion is strong. It shows not only that $u(k)$ converges to y as $k \to \infty$, but it also gives a convergence rate that is linear in k. Notably, the result does not depend on the distribution of input data $x_1, \ldots, x_n$.

The assumptions are much weaker than they appear. The first assumption is just to simplify the proof, and the authors claim that they can modify the proof for other cases by scaling the initialization. The second assumption looks harder to guarantee. But Xie et al. and Tsuchida et al. studied the matrix H^∞, and in their work it is shown that $\lambda_{\min}(H^\infty) > 0$ as long as the data is nondegenerate.

Nonetheless, all the bounds heavily depend on $\lambda_0 = \lambda_{\min}(H^\infty)$, and all get worse as λ_0 gets smaller. Indeed, as $\lambda_0 \to 0$, we have $m \to \infty, \eta \to 0$, and the decreasing coefficient for the convergence bound $(1 - \frac{\eta\lambda_0}{2})^k \to 1$.

All of the same behavior occurs as the number of training examples n increases. However, we expect to have to try harder with more data, and the authors believe that the dependency on n is not tight. It is not clear whether λ_0 can become infinitely small in practice even with a small dataset, so the dependence on λ_0 is potentially more serious.

3 Examples And Counter-Examples

First we discuss the intuition behind some of the assumptions and setup and indicate why they are necessary. Essentially, the proof shows that for a sufficiently over-parametrized network, the dynamics of gradient descent do not change much after initialization. Thus, gradient descent converges well as long as the initialization is regular enough.

The assumptions and setup provide this regularity. Specifically, the convergence rate of gradient descent on iteration k depends on the smallest eigenvalue of a matrix $H(k)$, which is at least $\frac{\lambda_0}{2}$ with high probability as long as $w(k)$ does not deviate too much from $w(0)$, which in turn holds true for large enough m.

$$H_{ij}(k) \triangleq \frac{1}{m}\sum_{r=1}^{m} x_i^\top x_j \mathbb{1}\{w_r(k)^\top x_i \geq 0, w_r(k)^\top x_j \geq 0\}$$

To prove this, the authors leverage the fact that $\lambda_0 > 0$, and since $\mathbb{E}(H(0)) = H^\infty$ they can use Hoeffding's inequality to show that after initialization $\lambda_{\min}(H(0)) \geq \frac{3\lambda_0}{4}$.

For this step in the proof, it is essential that $a_r^2 = 1$. Indeed, the authors need the dynamics of gradient descent to depend on $H(k)$, which by construction does not depend on a. If instead we had $a_r \sim \text{unif}[-1, 1]$, then the dynamics would depend on a_r and the theorem would become much more difficult to prove, perhaps false. Interestingly, the initialization of w_r as Gaussian is not used to control the initialization $H(0)$. Indeed, the matrices $H(0)$ and H^∞ only depend on the direction of w and not its magnitude. So thus far the proof would still hold if the vectors w_r were distributed uniformly on the unit sphere.

The reason w_r is chosen to be Guassian is in fact that the quantities $w_1(0)^\top x_i, \ldots, w_m(0)^\top x_i$ are very spread out, so that the vectors $w_r(k)$ do not need to deviate much from their initialization. The authors are then able to show that $H(k)$ does not deviate much from $H(0)$, so that $\lambda_{\min}(H(k)) \geq \lambda_0/2$. As a result, they get a nice convergence rate for gradient descent.

Next, we discuss the nebulous quantity $\lambda_0 = \lambda_{\min}(H^\infty)$. Most of the abstraction and difficulty in the statement of the main result comes from λ_0, so we give a few examples as to how λ_0 behaves depending on the data $x_1, \ldots, x_n$. To unpack λ_0, let us first rewrite

$$H_{i,j}^\infty = (x_i^\top x_j) P_{w\sim N(0,I)}(w^\top x_i \geq 0, w^\top x_j \geq 0).$$

Now $\|x_i\|_2^2 = 1$ and $P_{w\sim N(0,I)}(w^\top x_i \geq 0, w^\top x_j \geq 0) \leq \frac{1}{2}$, with equality if and only if $i = j$ (assuming the data is nondegenerate). So

$$|H_{i,j}^\infty| \leq \frac{1}{2},$$

with equality if and only if $i = j$.

One family of data $x_1, \ldots, x_n$ for which λ_0 is easy to compute is when $x_1, \ldots, x_n$ are orthogonal. Then $H_{i,j}^\infty = \mathbb{1}[i = j] P_{w\sim N(0,I)}(w^\top x_i \geq 0) = \frac{1}{2}\mathbb{1}[i = j]$. In other words, $H^\infty = \frac{1}{2}I$, so that $\lambda_0 = \frac{1}{2}$.

Following our discussion in the previous section, this suggests that the network and gradient descent fare better the more spread out the data is. This fact corresponds with the intuition that sparse data should be easier to fit.

A bit more generally, suppose that for each $i \in [n]$

$$\sum_{j \neq i} |x_i^\top x_j| P_{w \sim N(0,I)}(w^\top x_i \geq 0, w^\top x_j \geq 0) \leq \frac{1}{2} - \delta$$

Then $H^\infty_{i,i} - \sum_{j \neq i} |H^\infty_{i,j}| \geq \delta$. By the Gershgorin circle theorem, it follows that $\lambda_0 \geq \delta$. For this inequality to hold, the data again needs to be fairly sparse.

Conversely, we consider a case where the data is concentrated. Send $x_j \to x_i$ for some $i \neq j$ and fix all data points besides x_j. Let H^∞_ϵ be the matrix H^∞ where $\|x_j - x_i\| = \epsilon$, and let $\lambda_{0,\epsilon}$ be the smallest eigenvalue of H^∞_ϵ. We show that $\lambda_{0,\epsilon} \to 0$. Observe that the matrix $\lim_{\epsilon \to 0} H^\infty_\epsilon$ has two identical columns and therefore has an eigenvalue of zero. Then the characteristic polynomial $\chi_\epsilon(\lambda)$ of H^∞_ϵ approaches a polynomial which has a zero root. A continuity argument using the argument principle from complex analysis then shows that $\chi_\epsilon(\lambda)$ has a root approaching zero, so $\lambda_{0,\epsilon} \to 0$. Thus the quantities in the theorem can become arbitrarily bad as the data becomes degenerate, even for a small dataset.

4 Limitations And Future Work

The authors present a strong theoretical result: that gradient descent can converge to a global minimum with high probability for an over-parametrized network. However, from a practical perspective, their result does not offer very much. When training a neural network, one hopes to generalize the network's performance on the training set to unseen data, not perfectly fit the training set. Achieving zero training loss often means the network is overfitting. (That being said, it is encouraging that adding more nodes to the network eventually leads to arbitrarily small training loss.)

And from a theoretical perspective, it is not clear how the quantity λ_0 behaves under different datasets $x_1, \ldots, x_n$. Future work is needed here. One might ask, for example, what is $\mathbb{E}(\lambda_0)$ when the x_i are drawn uniformly from the unit sphere? And how does this quantity depend on the dimension of the data? This can be estimated empirically. Some work in this direction has already been done in Xie et al., where they suggest that the distribution of w_r should match that of x_i. How much does this choice of distribution for w_r help? Further, suppose we have a guarantee of sparseness, for example that $\|x_i - x_j\| \geq \delta$ for $i \neq j$. (Perhaps after cleaning the dataset.) How small can λ_0 be in terms of δ?

Other ideas for future work involve improving and generalizing the paper to other settings. The authors suggest that with more careful analysis, the bounds on the quantities in the main result could be improved. Also, since the proof techniques were relatively standard, the argument may generalize to deeper networks, different loss functions, or other first-order optimization algorithms.

As a starting point, one might empirically study how the vectors $w_r^{(i)}$ change under gradient descent when training an over-parametrized deeper network. That $w_r(k)$ does not deviate too much from the initialized value $w_r(0)$ is a key step in the proof. If an analogous argument works for deeper networks, one may expect similar behavior in the parameters $w_r^{(i)}$ as the network trains.

References

[1] Russell Tsuchida, Farbod Roosta-Khorasani, and Marcus Gallagher. Invariance of weight distributions in rectified mlps. *CoRR*, abs/1711.09090, 2017.

[2] Bo Xie, Yingyu Liang, and Le Song. Diversity leads to generalization in neural networks. *CoRR*, abs/1611.03131, 2016.

[3] Barnabas Poczos Simon Du, Xiyu Zhai and Aarti Singh. Gradient descent provably optimizes over-parametrized neural networks. 2018.

[4] Yuanzhi Li and Yingyu Liang. Gradient descent provably optimizes over-parametrized neural networks. 2018.

[5] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *CoRR*, abs/1611.03530, 2016.