1 Gradient Descent Finds Global Minima of Deep Neural Networks

Definitions

- m: Width of each layer of the neural network.
- n: number of samples.
- d: dimension of training data.
- H: number of layers of the neural network.
- η: learning rate for gradient descent.
- θ: parameters of the neural network.
- $\theta(k)$: parameters of the neural network after k iterations of training with gradient descent. $\theta(0)$ are the parameters at initialization (iid $N(0,1)$).
- σ: Activation function. It is Lipschitz, smooth, analytical and not a polynomial.
- $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}, 1 \leq i \leq n$: training data and corresponding labels. In this work, it is assumed that no two input points are parallel, i.e. $x_i \nparallel x_j$ for $i \neq j$.
- $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$: vector of labels.
- $\mathbf{W}^{(1)} \in \mathbb{R}^{m \times d}, \mathbf{W}^{(h)} \in \mathbb{R}^{m \times m} 2 \leq h \leq H, \mathbf{a} \in \mathbb{R}^m$ are, respectively, the first layer, the h layer and the output layer of the neural network respectively. We also use $\mathbf{W}^{(h)}(k)$, $\mathbf{a}(k)$ to denote the layers after k iterations of training with GD.
- $c_\sigma = \left(\mathbb{E}_{x \sim N(0,1)}\left[\sigma(x)^2\right]\right)^{-1}$ is a scaling factor to normalize the input in the initialization phase of the neural network.
- **Fully-connected neural network (NN)**. Let $\mathbf{x}^{(0)}$ be an input of the NN. Then the fully-connected neural network function f is defined recursively in the following way:

$$\mathbf{x}^{(h)} = \sqrt{\frac{c_\sigma}{m}}\sigma\left(\mathbf{W}^{(h)}\mathbf{x}^{(h-1)}\right), 1 \leq h \leq H$$
$$f(\mathbf{x}, \theta) = \mathbf{a}^\top \mathbf{x}^{(H)}.$$

  where $c_\sigma = \left(\mathbb{E}_{x \sim N(0,1)}\left[\sigma(x)^2\right]\right)^{-1}$ is the scaling defined above.
- **Loss function (ℓ_2)**. $L(\theta) = \frac{1}{2}\sum_{i=1}^n (f(\theta, \mathbf{x}_i) - y_i)^2$.
- $u_i(k) = f(\theta(k), \mathbf{x}_i)$. Output of the NN for sample i after k iterations of GD.

- $\mathbf{u}(k) = (u_1(k), \ldots, u_n(k))^\top \in \mathbb{R}^n$.
- $\mathbf{G}^{(h)}(k) \in \mathbb{R}^{n\times n}$, $1 \le h \le H+1$ defined as $\mathbf{G}^{(h)}_{ij}(k) = \left\langle \frac{\partial u_i(k)}{\partial \mathbf{W}^{(h)}(k)}, \frac{\partial u_j(k)}{\partial \mathbf{W}^{(h)}(k)} \right\rangle$ for $h = 1, \ldots, H$ and $\mathbf{G}^{(H+1)}_{ij}(k) = \left\langle \frac{\partial u_i(k)}{\partial \mathbf{a}(k)}, \frac{\partial u_j(k)}{\partial \mathbf{a}(k)} \right\rangle$. So that the following definition can be used to express the dynamics of the NN.
- $\mathbf{G}(k)$ defined as $\mathbf{G}_{ij}(k) = \sum_{h=1}^{H+1} \mathbf{G}^{(h)}_{ij}(k)$. Note that for the infinite NTK the function behaves as its linearization and it holds

  $$\mathbf{y} - \mathbf{u}(k+1) = (\mathbf{I} - \eta K)(\mathbf{y} - \mathbf{u}(k)),$$

  We want to argue that

  $$\mathbf{y} - \mathbf{u}(k+1) \approx (\mathbf{I} - \eta \mathbf{G}(k))(\mathbf{y} - \mathbf{u}(k)),$$

  in a precise way. Note the gradient descent update is

  $$\mathbf{W}^{(h)}(k) = \mathbf{W}^{(h)}(k-1) - \eta \frac{\partial L(\theta(k-1))}{\partial \mathbf{W}^{(h)}(k-1)},$$
  $$\mathbf{a}(k) = \mathbf{a}(k-1) - \eta \frac{\partial L(\theta(k-1))}{\partial \mathbf{a}(k-1)}.$$

  Remark 1.1. Each entry of $\mathbf{G}^{(h)}(k)$ is an inner product and thus $\mathbf{G}^{(h)}(k)$ is a PSD matrix. Furthermore, if there exists one $h \in [H]$ such that $\mathbf{G}^{(h)}(k)$ is strictly positive definite, then if one chooses the step size η to be sufficiently small, the loss decreases at the $k-th$ iteration according the analysis of power method, which presents linear convergence rate. In the paper they focus on $\mathbf{G}^{(H)}(k)$ only.
- $\mathbf{K}^{(h)}$ is a fixed matrix which depends on the input data, neural network architecture (including the activation function but does not depend on the parameters θ. It will be shown that $\mathbf{G}^{(H)}(0)$ at initialization is close to $\mathbf{K}^{(H)}$, that $\mathbf{G}^{(H)}(k)$ is close to $\mathbf{G}^{(H)}(0)$ and that $\mathbf{K}^{(H)}$ is positive semidefinite. These three things imply linear convergence of gradient descent by proving that the minimum eigenvalue of $\mathbf{G}^{(H)}(k)$ is bounded below by a constant independent of k. The definition of these matrices for the fully neural network connected the following:

  $$\begin{aligned}
  \mathbf{K}^{(0)}_{ij} &= \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\
  \mathbf{A}^{(h)}_{ij} &= \begin{pmatrix} \mathbf{K}^{(h-1)}_{ii} & \mathbf{K}^{(h-1)}_{ij} \\ \mathbf{K}^{(h-1)}_{ji} & \mathbf{K}^{(h-1)}_{jj} \end{pmatrix} \\
  \mathbf{K}^{(h)}_{ij} &= c_\sigma \mathbb{E}_{(u,v)^\top \sim N\left(\mathbf{0}, \mathbf{A}^{(h)}_{ij}\right)} [\sigma(u)\sigma(v)] \\
  \mathbf{K}^{(H)}_{ij} &= c_\sigma \mathbf{K}^{(H-1)}_{ij} \mathbb{E}_{(u,v)^\top \sim N\left(\mathbf{0}, \mathbf{A}^{(H-1)}_{ij}\right)} \left[\sigma'(u)\sigma'(v)\right]
  \end{aligned} \tag{1}$$

- $u_i'(\theta) = \frac{\partial u_i}{\partial \theta}, u_i^{(h)}(\theta) = \frac{\partial u_i}{\partial \mathbf{W}^{(h)}}, u_i^{(a)}(\theta) = \frac{\partial u_i}{\partial \mathbf{a}}, L'(\theta) = \frac{\partial L(\theta)}{\partial \theta}, L^{(h)}(\mathbf{W}^{(h)}) = \frac{\partial L(\theta)}{\partial \mathbf{W}^{(h)}}, L^{(a)}(\theta) = \frac{\partial L}{\partial \mathbf{a}}$.

Results

The paper proves linear global convergence, i.e. to zero training error of some deep networks architectures with high probability with respect to the initialization assuming the networks are sufficiently overparametrized and that ℓ_2 loss is used. Note the learning rate has to be quite small, much more that what would be used in practice. Another caveat is that overparametrization depends on λ_0 the minimum eigenvalue of $K^(H)$ which is proved to be positive but it is not provided any kind of guarantee for λ_0 not being arbitrarily small in some cases.

The results of the paper are for fully-connected NNs, which needs exponential over-parametrization with depth, for ResNets, in which this dependence with depth drops to a polynomial, and convolutional ResNets. In these notes we focus on the fully-connected architecture for simplicity. The arguments are quite similar across architectures.

Theorem 1.2 (Convergence Rate of Gradient Descent for Deep Fully-connected Neural Networks)**.** *Assume for all $i \in [n]$, $\|\mathbf{x}_i\|_2 = 1$, $|y_i| = O(1)$ and the number of hidden nodes per layer*

$$m = \Omega\left(2^{O(H)} \max\left\{\frac{n^4}{\lambda_{\min}^4\left(\mathbf{K}^{(H)}\right)}, \frac{n}{\delta}, \frac{n^2\log(\frac{Hn}{\delta})}{\lambda_{\min}^2\left(\mathbf{K}^{(H)}\right)}\right\}\right)$$

If we set the step size

$$\eta = O\left(\frac{\lambda_{\min}\left(\mathbf{K}^{(H)}\right)}{n^2 2^{O(H)}}\right),$$

then with probability at least $1-\delta$ over the random initialization, for $k = 1, 2, \ldots$, the loss at each iteration satisfies

$$L(\theta(k)) \leq \left(1 - \frac{\eta\lambda_{\min}\left(\mathbf{K}^{(H)}\right)}{2}\right)^k L(\theta(0)).$$

In order to prove the theorem, we introduce a few lemmas. First, we state the condition of the theorem we want to prove for all k with high probability, where λ_0 is the minimum eigenvalue of $\mathbf{K}^{(H)}$.

Condition 1.3. *At the k-th iteration, we have*

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

Lemma 1.4 (Initialization norm)**.** *If $\sigma(\cdot)$ is L-Lipschitz and $m = \Omega\left(\frac{nHg_c(H)^2}{\delta}\right)$ with $C = c_\sigma L(2|\sigma(0)|\sqrt{\frac{2}{\pi}} + 2L)$, then with probability at least $1-\delta$ over random initialization, for every $h \in [H]$ and $i \in [n]$ we have*

$$\frac{1}{c_{x,0}} \leq \left\|\mathbf{x}_i^{(h)}(0)\right\|_2 \leq c_{x,0},$$

where $c_{x,0} = 2$.

A similar lemma can be proven for different architectures with a different value of $c_{x,0}$. This lemma is needed in the proofs of Lemmas 1.6 and 1.7.

Lemma 1.5 (Least Eigenvalue at the Initialization)**.** *If* $m = \Omega\left(\frac{n^2 \log(Hn/\delta) 2^{O(H)}}{\lambda_0^2}\right)$ *we have*

$$\lambda_{\min}(\mathbf{G}^{(H)}(0)) \geq \frac{3}{4}\lambda_0.$$

Lemma 1.6 (Least Eigenvalue at the Initialization)**.** *Suppose for every* $h \in [H]$, $\left\|\mathbf{W}^{(h)}(0)\right\|_2 \leq c_{w,0}\sqrt{m}$, $\left\|\mathbf{x}^{(h)}(0)\right\|_2 \leq c_{x,0}$ *and* $\left\|\mathbf{W}^{(h)}(k) - \mathbf{W}^{(h)}(0)\right\|_F \leq \sqrt{m}R$ *for some constant* $c_{w,0}, c_{x,0} > 0$ *and* $R \leq c_{w,0}$*. If* $\sigma(\cdot)$ *is* $L-$*Lipschitz, we have*

$$\left\|\mathbf{x}^{(h)}(k) - \mathbf{x}^{(h)}(0)\right\|_2 \leq \sqrt{c_\sigma} L c_{x,0} g_{c_x}(h) R$$

where $c_x = 2\sqrt{c_\sigma} L c_{w,0}$.

Lemma 1.7. *Suppose* $\sigma(\cdot)$ *is* $L-$*Lipschitz and* $\beta-$*smooth. Suppose for* $h \in [H]$, $\left\|\mathbf{W}^{(h)}(0)\right\|_2 \leq c_{w,0}\sqrt{m}$, $\|\mathbf{a}(0)\|_2 \leq a_{2,0}\sqrt{m}$, $\|\mathbf{a}(0)\|_4 \leq a_{4,0}m^{1/4}$, $\frac{1}{c_{x,0}} \leq \left\|\mathbf{x}^{(h)}(0)\right\|_2 \leq c_{x,0}$*, if* $\left\|\mathbf{W}^{(h)}(k) - \mathbf{W}^{(h)}(0)\right\|_F$, $\|\mathbf{a}(k) - \mathbf{a}(0)\|_2 \leq \sqrt{m}R$ *where* $R \leq c g_{c_x}(H)^{-1}\lambda_0 n^{-1}$ *and* $R \leq c g_{c_x}(H)^{-1}$ *for some small constant* c *and* $c_x = 2\sqrt{c_\sigma} L c_{w,0}$*, we have*

$$\left\|\mathbf{G}^{(H)}(k) - \mathbf{G}^{(H)}(0)\right\|_2 \leq \frac{\lambda_0}{4}.$$

The assumption $\left\|W^{(h)}(0)\right\|_2 \leq c_{w,0}\sqrt{m}$ is a well know fact of gaussian initialized matrices and the bounds on $\|a(0)\|_2$ and $\|a(0)\|_4$ can be proved using standard concentration inequalities. $a_{2,0}$ and $a_{4,0}$ are universal constants.

Lemma 1.8. *If Condition 1.3 holds for* $k' = 1, \ldots, k$*, we have for any* $s = 1, \ldots, k+1$

$$\left\|\mathbf{W}^{(h)}(s) - \mathbf{W}^{(h)}(0)\right\|_F, \|\mathbf{a}(s) - \mathbf{a}(0)\|_2 \leq R'\sqrt{m}$$
$$\left\|\mathbf{W}^{(h)}(s) - \mathbf{W}^{(h)}(s-1)\right\|_F, \|\mathbf{a}(s) - \mathbf{a}(s-1)\|_2 \leq \eta Q'(s-1)$$

where $R' = \frac{16 c_{x,0} a_{2,0} (c_x)^H \sqrt{n} \|\mathbf{y} - \mathbf{u}(0)\|_2}{\lambda_0 \sqrt{m}} \leq c g_{c_x}(H)^{-1}$ *for some small constant* c *with* $c_x = \max\{2\sqrt{c_\sigma} L c_{w,0}, 1\}$ *and* $Q'(s) = 4 c_{x,0} a_{2,0} (c_x)^H \sqrt{n} \|\mathbf{y} - \mathbf{u}(s)\|_2$

Lemma 1.9. *Let*

$$I_2^i(k) = \int_{s=0}^{\eta} \left\langle L'(\theta(k)), u_i'(\theta(k)) - u_i'\left(\theta(k) - sL'(\theta(k))\right)\right\rangle ds$$

and $\mathbf{I}_2(k) = (I_2^1(k), \ldots, I_2^n(k))^\top$*. If Condition 1.3 holds for* $k' = 1, \ldots, k$*, suppose* $\eta \leq c\lambda_0 \left(n^2 H^2 (c_x)^{3H} g_{2c_x}(H)\right)^{-1}$ *for some small constant* c*, we have*

$$\|\mathbf{I}_2(k)\|_2 \leq \frac{1}{8}\eta\lambda_0 \|\mathbf{y} - \mathbf{u}(k)\|_2 .$$

Lemma 1.10. *If Condition 1.3 holds for* $k' = 1, \ldots, k$*, suppose* $\eta \leq c\lambda_0 \left(n^2 H^2 (c_x)^{2H} g_{2c_x}(H)\right)^{-1}$ *for some small constant* c*, then we have* $\|\mathbf{u}(k+1) - \mathbf{u}(k)\|_2^2 \leq \frac{1}{8}\eta\lambda_0 \|\mathbf{y} - \mathbf{u}(k)\|_2^2$.

Proof of Theorem 1.2. We want to prove Condition 1.3 for all k. We proceed by induction. Note that

$$\begin{aligned}
&\|\mathbf{y}-\mathbf{u}(k+1)\|_2^2 \\
=&\|\mathbf{y}-\mathbf{u}(k)-(\mathbf{u}(k+1)-\mathbf{u}(k))\|_2^2 \\
=&\|\mathbf{y}-\mathbf{u}(k)\|_2^2-2(\mathbf{y}-\mathbf{u}(k))^\top(\mathbf{u}(k+1)-\mathbf{u}(k))+\|\mathbf{u}(k+1)-\mathbf{u}(k)\|_2^2
\end{aligned} \tag{2}$$

We need the second summand to be greater in absolute value than the third one for the loss to decrease. Intuitively this is true because by a Taylor expansion of $\mathbf{u}(k+1)-\mathbf{u}(k)$ with respect to η we have that the second summand is of order η plus second order terms and the third summand is of order η^2, so for η small enough we can proof that the loss decreases. Then we have to prove that the first order term in η is proportional to the constant of 1.3. Expanding one coordinate of $\mathbf{u}(k+1)-\mathbf{u}(k)$ by Taylor we obtain

$$\mathbf{u}_i(k+1)-\mathbf{u}_i(k)=\left(-\eta\left\langle L'(\theta(k)),u_i'(\theta(k))\right\rangle\right)+I_2^i(k)$$

where, following the notation of the paper we denote $I_2^i(k)$ the second order term on η. It is equal to

$$I_2^i(k)=\int_{s=0}^{\eta}\left\langle L'(\theta(k)),u_i'(\theta(k))-u_i'\left(\theta(k)-sL'(\theta(k))\right)\right\rangle \mathrm{d}s.$$

But let's focus on the first term, which we denote $I_1^i(k)$, and let $\mathbf{I}_1(k)=(I_1^1(k)),\dots,(I_1^n(k))^\top$ and $\mathbf{I}_2(k)=(I_2^1(k)),\dots,(I_2^n(k))^\top$. We have

$$\begin{aligned}
I_1^i&=-\eta\left\langle L'(\theta(k)),u_i'(\theta(k))\right\rangle \\
&=-\eta\sum_{j=1}^{n}(u_j-y_j)\left\langle u_j'(\theta(k)),u_i'(\theta(k))\right\rangle \\
&\triangleq-\eta\sum_{j=1}^{n}(u_j-y_j)\sum_{h=1}^{H+1}\mathbf{G}_{ij}^{(h)}(k)
\end{aligned}$$

or in matricial form

$$\mathbf{I}_1(k)=-\eta\mathbf{G}(k)(\mathbf{u}(k)-\mathbf{y})$$

Now observe that

$$\begin{aligned}
(\mathbf{y}-\mathbf{u}(k))^\top\mathbf{I}_1(k)&=\eta(\mathbf{y}-\mathbf{u}(k))^\top\mathbf{G}(k)(\mathbf{y}-\mathbf{u}(k)) \\
&\geq\lambda_{\min}(\mathbf{G}(k))\|\mathbf{y}-\mathbf{u}(k)\|_2^2 \\
&\geq\lambda_{\min}\left(\mathbf{G}^{(H)}(k)\right)\|\mathbf{y}-\mathbf{u}(k)\|_2^2
\end{aligned} \tag{3}$$

We will only need to look at $\mathbf{G}^{(H)}$ which has the following form

$$\mathbf{G}_{i,j}^{(H)}(k)=\left(\mathbf{x}_i^{(H-1)}(k)\right)^\top\mathbf{x}_j^{(H-1)}(k)\cdot\frac{c_\sigma}{m}\sum_{r=1}^{m}a_r^2\sigma'\left(\left(\theta_r^{(H)}(k)\right)^\top\mathbf{x}_i^{(H-1)}(k)\right)\sigma'\left(\left(\theta_r^{(H)}(k)\right)^\top\mathbf{x}_j^{(H-1)}(k)\right)$$

In principle one could look at $\mathbf{G}(k)$ but in the paper they do not do that. The analysis becomes simple if only $\mathbf{G}^{(H)}$ is used.

So putting all together we have

$$\begin{aligned}
&\|\mathbf{y} - \mathbf{u}(k+1)\|_2^2 \\
&\overset{①}{\leq} \left(1 - \eta\lambda_{\min}\left(\mathbf{G}^{(H)}(k)\right)\right)\|\mathbf{y} - \mathbf{u}(k)\|_2^2 - 2(\mathbf{y} - \mathbf{u}(k))^\top \mathbf{I}_2(k) + \|\mathbf{u}(k+1) - \mathbf{u}(k)\|_2^2 \\
&\overset{②}{\leq} (1 - \eta\lambda_0)\|\mathbf{y} - \mathbf{u}(k)\|_2^2 - 2(\mathbf{y} - \mathbf{u}(k))^\top \mathbf{I}_2 + \|\mathbf{u}(k+1) - \mathbf{u}(k)\|_2^2 \\
&\overset{③}{\leq} \left(1 - \frac{\eta\lambda_0}{2}\right)\|\mathbf{y} - \mathbf{u}(k)\|_2^2.
\end{aligned}$$

① uses Equation (2) and inequality (3). ③ uses Lemmas 1.9 and 1.10. For ②, by induction hypothesis, using Lemma 1.8 we obtain

$$\begin{aligned}
\left\|\mathbf{W}^{(h)}(k) - \mathbf{W}^{(h)}(0)\right\|_F &\leq R'\sqrt{m} \\
&\leq R\sqrt{m}
\end{aligned}$$

for the choice of m in the theorem. By Lemma 1.7 we get $\lambda_{\min}\left(\mathbf{G}^{(H)}(k)\right) \geq \frac{\lambda_0}{2}$.

□