Lecture 3 — September 8, 2020

Prof. Zhao Song *Scribe: Zhao Song*

1 Neural Tangent Kernel

1.1 Problem Formulation

Our problem formulation is the same as [DZPS19, SY19]. We consider a two-layer ReLU activated neural network with m neurons in the hidden layer:

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

where $x\in\mathbb{R}^d$ is the input, $w_1,\cdots,w_m\in\mathbb{R}^d$ are weight vectors in the first layer, $a_1,\cdots,a_m\in\mathbb{R}$ are weights in the second layer. For simplicity, we only optimize W but not optimize a and W at the same time.

Recall that the ReLU function $\phi(x)=\max\{x,0\}$. Therefore for $r\in[m]$, we have

$$\frac{f(W,x,a)}{\partial w_r}=\frac{1}{\sqrt{m}}a_r x\mathbf{1}_{w_r^\top x\geq 0}. \tag{1}$$

We define objective function L as follows

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

We apply the gradient descent to optimize the weight matrix W in the following standard way,

$$W(k+1)=W(k)-\eta\frac{\partial L(W(k))}{\partial W(k)}. \tag{2}$$

We can compute the gradient of L in terms of w_r

$$\frac{\partial L(W)}{\partial w_r}=\frac{1}{\sqrt{m}}\sum_{i=1}^{n}(f(W,x_i,a_r)-y_i)a_r x_i\mathbf{1}_{w_r^\top x_i\geq 0}. \tag{3}$$

We consider the ordinary differential equation defined by

$$\frac{\mathrm{d}w_r(t)}{\mathrm{d}t}=-\frac{\partial L(W)}{\partial w_r}. \tag{4}$$

At time t, let $u(t)=(u_1(t),\cdots,u_n(t))\in\mathbb{R}^n$ be the prediction vector where each $u_i(t)$ is defined as

$$u_i(t)=f(W(t),a,x_i). \tag{5}$$

Algorithm 1 Training neural network using gradient descent.

```
procedure NNTraining({(x_i, y_i)}_{i∈[n]})
    w_r(0) ~ N(0, I_d) for r ∈ [m].
    for t = 1 → T do
        u(t) ← (1/√m) Σ_{r=1}^m a_r σ(w_r(t)^⊤ X)        ▷ u(t) = f(W(t), x, a) ∈ R^n, it takes O(mnd) time
        for r = 1 → m do
            for i = 1 → n do
                Q_{i,:} ← (1/√m) a_r σ'(w_r(t)^⊤ x_i) x_i^⊤    ▷ Q_{i,:} = ∂f(W(t), x_i, a)/∂w_r, it takes O(d) time
            end for
            grad_r ← −Q^⊤ (y − u(t))                       ▷ Q = ∂f/∂w_r ∈ R^{n×d}, it takes O(nd) time
            w_r(t+1) ← w_r(t) − η · grad_r
        end for
    end for
    return W
end procedure
```

1.2 Bounding the difference between continuous and discrete

In this section, we restate a result from [DZPS19], showing that when the width m is sufficiently large, then the continuous version and discrete version of the gram matrix of input data is close in the spectral sense.

Lemma 1 (Lemma 3.1 in [DZPS19]). *We define $H^{\mathrm{cts}}, H^{\mathrm{dis}} \in \mathbb{R}^{n \times n}$ as follows*

$$H^{\mathrm{cts}}_{i,j} = \mathop{\mathbb{E}}_{w \sim \mathcal{N}(0,I)} \left[x_i^\top x_j \mathbf{1}_{w^\top x_i \geq 0, w^\top x_j \geq 0} \right],$$

$$H^{\mathrm{dis}}_{i,j} = \frac{1}{m} \sum_{r=1}^{m} \left[x_i^\top x_j \mathbf{1}_{w_r^\top x_i \geq 0, w_r^\top x_j \geq 0} \right].$$

Let $\lambda = \lambda_{\min}(H^{\mathrm{cts}})$. If $m = \Omega(\lambda^{-2} n^2 \log(n/\delta))$, we have

$$\|H^{\mathrm{dis}} - H^{\mathrm{cts}}\|_F \leq \frac{\lambda}{4}, \text{ and } \lambda_{\min}(H^{\mathrm{dis}}) \geq \frac{3}{4}\lambda.$$

hold with probability at least $1 - \delta$.

The proof can be found in Appendix **??**.

We define the event

$$A_{i,r} = \left\{ \exists u : \|u - \widetilde{w}_r\|_2 \leq R, \mathbf{1}_{x_i^\top \widetilde{w}_r \geq 0} \neq \mathbf{1}_{x_i^\top u \geq 0} \right\}.$$

Note this event happens if and only if $|\widetilde{w}_r^\top x_i| < R$. Recall that $\widetilde{w}_r \sim \mathcal{N}(0, I)$. By anti-concentration inequality of Gaussian (Lemma **??**), we have

$$\Pr[A_{i,r}] = \Pr_{z \sim \mathcal{N}(0,1)}[|z| < R] \leq \frac{2R}{\sqrt{2\pi}}. \tag{6}$$

1.3 Bounding changes of H when w is in a small ball

We improve the Lemma 3.2 in [DZPS19] from the two perspective : one is the probability, and the other is upper bound on spectral norm.

Lemma 2 (perturbed w)**.** *Let $R \in (0,1)$. If $\widetilde{w}_1, \cdots, \widetilde{w}_m$ are i.i.d. generated $\mathcal{N}(0,I)$. For any set of weight vectors $w_1, \cdots, w_m \in \mathbb{R}^d$ that satisfy for any $r \in [m]$, $\|\widetilde{w}_r - w_r\|_2 \leq R$, then the $H : \mathbb{R}^{m \times d} \to \mathbb{R}^{n \times n}$ defined*

$$H(w)_{i,j} = \frac{1}{m} x_i^\top x_j \sum_{r=1}^m \mathbf{1}_{w_r^\top x_i \geq 0, w_r^\top x_j \geq 0}.$$

Then we have

$$\|H(w) - H(\widetilde{w})\|_F < 2nR,$$

holds with probability at least $1 - n^2 \cdot \exp(-mR/10)$.

Proof. The random variable we care is

$$\begin{aligned}
&\sum_{i=1}^n \sum_{j=1}^n |H(\widetilde{w})_{i,j} - H(w)_{i,j}|^2 \\
\leq & \frac{1}{m^2} \sum_{i=1}^n \sum_{j=1}^n \left(\sum_{r=1}^m \mathbf{1}_{\widetilde{w}_r^\top x_i \geq 0, \widetilde{w}_r^\top x_j \geq 0} - \mathbf{1}_{w_r^\top x_i \geq 0, w_r^\top x_j \geq 0} \right)^2 \\
= & \frac{1}{m^2} \sum_{i=1}^n \sum_{j=1}^n \Big(\sum_{r=1}^m s_{r,i,j} \Big)^2,
\end{aligned}$$

where the last step follows from for each r, i, j, we define

$$s_{r,i,j} := \mathbf{1}_{\widetilde{w}_r^\top x_i \geq 0, \widetilde{w}_r^\top x_j \geq 0} - \mathbf{1}_{w_r^\top x_i \geq 0, w_r^\top x_j \geq 0}.$$

We consider i, j are fixed. We simplify $s_{r,i,j}$ to s_r.

Then s_r is a random variable that only depends on $\widetilde{w}_r$. Since $\{\widetilde{w}_r\}_{r=1}^m$ are independent, $\{s_r\}_{r=1}^m$ are also mutually independent.

If $\neg A_{i,r}$ and $\neg A_{j,r}$ happen, then

$$\left| \mathbf{1}_{\widetilde{w}_r^\top x_i \geq 0, \widetilde{w}_r^\top x_j \geq 0} - \mathbf{1}_{w_r^\top x_i \geq 0, w_r^\top x_j \geq 0} \right| = 0.$$

If $A_{i,r}$ or $A_{j,r}$ happen, then

$$\left| \mathbf{1}_{\widetilde{w}_r^\top x_i \geq 0, \widetilde{w}_r^\top x_j \geq 0} - \mathbf{1}_{w_r^\top x_i \geq 0, w_r^\top x_j \geq 0} \right| \leq 1.$$

So we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\widetilde{w}_r}[s_r] \leq \mathop{\mathbb{E}}_{\widetilde{w}_r}\left[\mathbf{1}_{A_{i,r} \vee A_{j,r}}\right] \leq & \Pr[A_{i,r}] + \Pr[A_{j,r}] \\
\leq & \frac{4R}{\sqrt{2\pi}} \\
\leq & 2R,
\end{aligned}$$

and

$$\begin{aligned}
\mathop{\mathbb{E}}_{\widetilde{w}_r}\left[\left(s_r - \mathop{\mathbb{E}}_{\widetilde{w}_r}[s_r]\right)^2\right] &= \mathop{\mathbb{E}}_{\widetilde{w}_r}[s_r^2] - \mathop{\mathbb{E}}_{\widetilde{w}_r}[s_r]^2 \\
&\leq \mathop{\mathbb{E}}_{\widetilde{w}_r}[s_r^2] \\
&\leq \mathop{\mathbb{E}}_{\widetilde{w}_r}\left[\left(\mathbf{1}_{A_{i,r}\vee A_{j,r}}\right)^2\right] \\
&\leq \frac{4R}{\sqrt{2\pi}} \\
&\leq 2R.
\end{aligned}$$

We also have $|s_r| \leq 1$. So we can apply Bernstein inequality (Lemma **??**) to get for all $t > 0$,

$$\begin{aligned}
\Pr\left[\sum_{r=1}^{m} s_r \geq 2mR + mt\right] &\leq \Pr\left[\sum_{r=1}^{m}(s_r - \mathbb{E}[s_r]) \geq mt\right] \\
&\leq \exp\left(-\frac{m^2t^2/2}{2mR + mt/3}\right).
\end{aligned}$$

Choosing $t = R$, we get

$$\begin{aligned}
\Pr\left[\sum_{r=1}^{m} s_r \geq 3mR\right] &\leq \exp\left(-\frac{m^2R^2/2}{2mR + mR/3}\right) \\
&\leq \exp\left(-mR/10\right).
\end{aligned}$$

Thus, we can have

$$\Pr\left[\frac{1}{m}\sum_{r=1}^{m} s_r \geq 3R\right] \leq \exp(-mR/10).$$

Therefore, we complete the proof. $\square$

1.4 Loss is decreasing while weights are not changing much

For simplicity of notation, we provide the following definition.

Definition 3. *For any $s \in [0, t]$, we define matrix $H(s) \in \mathbb{R}^{n\times n}$ as follows*

$$H(s)_{i,j} = \frac{1}{m}\sum_{r=1}^{m} x_i^\top x_j \mathbf{1}_{w_r(s)^\top x_i \geq 0, w_r(s)^\top x_j \geq 0}.$$

With H defined, it becomes more convenient to write the dynamics of predictions (proof can be found in Appendix **??**).

Fact 4. $\frac{\mathrm{d}}{\mathrm{d}t}u(t) = H(t)\cdot(y - u(t))$.

We state two tools from previous work(delayed the proof into Appendix **??**)

Lemma 5 (Lemma 3.3 in [DZPS19]). *Suppose for $0 \le s \le t$, $\lambda_{\min}(H(w(s))) \ge \lambda/2$. Let D_{cts} be defined as $D_{\text{cts}} := \frac{\sqrt{n}\|y-u(0)\|_2}{\sqrt{m}\lambda}$. Then we have*

1. $$\|w_r(t) - w_r(0)\|_2 \le D_{\text{cts}}, \forall r \in [m],$$
2. $$\|y - u(t)\|_2^2 \le \exp(-\lambda t) \cdot \|y - u(0)\|_2^2.$$

Lemma 6 (Lemma 3.4 in [DZPS19]). *If $D_{\text{cts}} < R$. then for all $t \ge 0$, $\lambda_{\min}(H(t)) \ge \frac{1}{2}\lambda$. Moreover,*

1. $$\|w_r(t) - w_r(0)\|_2 \le D_{\text{cts}}, \forall r \in [m],$$
2. $$\|y - u(t)\|_2^2 \le \exp(-\lambda t) \cdot \|y - u(0)\|_2^2.$$

1.5 Convergence

In this section we show that when the neural network is over-parametrized, the training error converges to 0 at linear rate. Our main result is Theorem 7.

Theorem 7 (Main result in [SY19]). *Recall that $\lambda = \lambda_{\min}(H^{\text{cts}}) > 0$. Let $m = \Omega(\lambda^{-4} n^4 \log(n/\delta))$, we i.i.d. initialize $w_r \in \mathcal{N}(0, I)$, a_r sampled from $\{-1, +1\}$ uniformly at random for $r \in [m]$, and we set the step size $\eta = O(\lambda/n^2)$ then with probability at least $1 - \delta$ over the random initialization we have for $k = 0, 1, 2, \cdots$*

$$\|u(k) - y\|_2^2 \le (1 - \eta\lambda/2)^k \cdot \|u(0) - y\|_2^2. \tag{7}$$

Correctness We prove Theorem 7 by induction. The base case is $i = 0$ and it is trivially true. Assume for $i = 0, \cdots, k$ we have proved Eq. (7) to be true. We want to show Eq. (7) holds for $i = k + 1$.

From the induction hypothesis, we have the following Lemma (see proof in Appendix **??**) stating that the weights should not change too much.

Lemma 8 (Corollary 4.1 in [DZPS19]). *If Eq. (7) holds for $i = 0, \cdots, k$, then we have for all $r \in [m]$*

$$\|w_r(k+1) - w_r(0)\|_2 \le \frac{4\sqrt{n}\|y - u(0)\|_2}{\sqrt{m}\lambda} := D.$$

Next, we calculate the different of predictions between two consecutive iterations, analogue to $\frac{\mathrm{d}u_i(t)}{\mathrm{d}t}$ term in Fact 4. For each $i \in [n]$, we have

$$\begin{aligned}
& u_i(k+1) - u_i(k) \\
= & \frac{1}{\sqrt{m}} \sum_{r=1}^m a_r \cdot \left(\phi(w_r(k+1)^\top x_i) - \phi(w_r(k)^\top x_i) \right) \\
= & \frac{1}{\sqrt{m}} \sum_{r=1}^m a_r \cdot \left(\phi\left(\left(w_r(k) - \eta \frac{\partial L(W(k))}{\partial w_r(k)} \right)^\top x_i \right) - \phi(w_r(k)^\top x_i) \right).
\end{aligned}$$

Here we divide the right hand side into two parts. $v_{1,i}$ represents the terms that the pattern does not change and $v_{2,i}$ represents the term that pattern may changes. For each $i \in [n]$, we define the set $S_i \subset [m]$ as

$$\begin{aligned} S_i := \{r \in [m] : \forall w \in \mathbb{R}^d \text{ s.t. } & \|w - w_r(0)\|_2 \le R, \\ & \mathbf{1}_{w_r(0)^\top x_i \ge 0} = \mathbf{1}_{w^\top x_i \ge 0}\}. \end{aligned}$$

Then we define $v_{1,i}$ and $v_{2,i}$ as follows

$$\begin{aligned} v_{1,i} &:= \frac{1}{\sqrt{m}} \sum_{r \in S_i} a_r \left(\phi\left(\left(w_r(k) - \eta \frac{\partial L(W(k))}{\partial w_r(k)} \right)^\top x_i \right) - \phi(w_r(k)^\top x_i) \right), \\ v_{2,i} &:= \frac{1}{\sqrt{m}} \sum_{r \in \overline{S}_i} a_r \left(\phi\left(\left(w_r(k) - \eta \frac{\partial L(W(k))}{\partial w_r(k)} \right)^\top x_i \right) - \phi(w_r(k)^\top x_i) \right). \end{aligned}$$

Define H and $H^\perp \in \mathbb{R}^{n \times n}$ as

$$\begin{aligned} H(k)_{i,j} &= \frac{1}{m} \sum_{r=1}^{m} x_i^\top x_j \mathbf{1}_{w_r(k)^\top x_i \ge 0, w_r(k)^\top x_j \ge 0}, \\ H(k)^\perp_{i,j} &= \frac{1}{m} \sum_{r \in \overline{S}_i} x_i^\top x_j \mathbf{1}_{w_r(k)^\top x_i \ge 0, w_r(k)^\top x_j \ge 0}. \end{aligned}$$

and

$$\begin{aligned} C_1 &= -2\eta (y - u(k))^\top H(k) (y - u(k)), \\ C_2 &= 2\eta (y - u(k))^\top H(k)^\perp (y - u(k)), \\ C_3 &= -2(y - u(k))^\top v_2, \\ C_4 &= \|u(k+1) - u(k)\|_2^2. \end{aligned}$$

Then we have

Claim 9.

$$\|y - u(k+1)\|_2^2 = \|y - u(k)\|_2^2 + C_1 + C_2 + C_3 + C_4.$$

Applying Claim 11, 12, 13 and 14 gives

$$\begin{aligned} \|y - u(k+1)\|_2^2 \le\ & \|y - u(k)\|_2^2 \\ & \cdot (1 - \eta\lambda + 8\eta n R + 8\eta n R + \eta^2 n^2). \end{aligned}$$

Choice of η and R. Next, we want to choose η and R such that

$$(1 - \eta\lambda + 8\eta n R + 8\eta n R + \eta^2 n^2) \le (1 - \eta\lambda/2). \tag{8}$$

If we set $\eta = \frac{\lambda}{4n^2}$ and $R = \frac{\lambda}{64n}$, we have

$$8\eta n R + 8\eta n R = 16\eta n R \le \eta\lambda/4, \quad \text{and} \quad \eta^2 n^2 \le \eta\lambda/4.$$

This implies

$$\|y - u(k+1)\|_2^2 \le \|y - u(k)\|_2^2 \cdot (1 - \eta\lambda/2)$$

holds with probability at least $1 - 3n^2 \exp(-mR/10)$.

Over-parameterization size, lower bound on m. We require

$$D = \frac{4\sqrt{n}\|y - u(0)\|_2}{\sqrt{m}\lambda} < R = \frac{\lambda}{64n}, \text{ and } 3n^2 \exp(-mR/10) \le \delta.$$

By Claim 10, it is sufficient to choose $m = \Omega(\lambda^{-4} n^4 \log(m/\delta) \log^2(n/\delta))$.

1.6 Technical Claims

In this section we only list all the statement and left the proofs as an exercise.

Claim 10. *For $0 < \delta < 1$, we have*

$$\|y - u(0)\|_2^2 = O(n \log(m/\delta) \log^2(n/\delta))$$

holds with probability at least $1 - \delta$.

Claim 11. *Let $C_1 = -2\eta(y - u(k))^\top H(k)(y - u(k))$. We have*

$$C_1 \le -\eta\lambda \cdot \|y - u(k)\|_2^2$$

holds with probability at least $1 - n^2 \cdot \exp(-mR/10)$.

Claim 12. *Let $C_2 = 2\eta(y - u(k))^\top H(k)^\perp (y - u(k))$. We have*

$$C_2 \le 8\eta n R \cdot \|y - u(k)\|_2^2$$

holds with probability $1 - n \cdot \exp(-mR)$.

Claim 13. *Let $C_3 = -2(y - u(k))^\top v_2$. Then we have*

$$C_3 \le 8\eta n R \cdot \|y - u(k)\|_2^2.$$

with probability at least $1 - n \cdot \exp(-mR)$.

Claim 14. *Let $C_4 = \|u(k+1) - u(k)\|_2^2$. Then we have*

$$C_4 \le \eta^2 n^2 \cdot \|y - u(k)\|_2^2.$$

References

[DZPS19] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *ICLR*. https://arxiv.org/pdf/1810.02054, 2019.

[SY19] Zhao Song and Xin Yang. Quadratic suffices for over-parametrization via matrix chernoff bound. In *arXiv preprint*. https://arxiv.org/pdf/1906.03593, 2019.