CS 839: Theoretical Foundations of Deep Learning	**Spring 2022**

Lecture 13 Lazy Training

Instructor: Yingyu Liang *Date: Mar 8^{th}, 2022* *Scriber: Yuchen Zeng*

1 Lazy Training of General Neural Networks

In previous lectures, we have shown many good properties of two-layer neural networks using the NTK formulation. In this lecture, we will show the key technical lemma still holds for more general neural networks, and the intuition is still the same. The formulation here is usually called lazy training.

Consider n samples $\{(\mathbf{x}_i, \mathrm{y}_i)\}_{i=1}^n$ and $\mathbf{y} = (\mathrm{y}_1, \dots, \mathrm{y}_n)^\top$. Denote the model of the neural network as function $f(\boldsymbol{x}, \boldsymbol{w})$ for weight $\boldsymbol{w} \in \mathbb{R}^p$, and let

$$\boldsymbol{f}(\boldsymbol{w}) = (f(\mathbf{x}_1, \boldsymbol{w}), \dots, f(\mathbf{x}_n, \boldsymbol{w}))^\top \in \mathbb{R}^n.$$

Note that the loss function

$$L(\alpha \boldsymbol{f}(\boldsymbol{w})) = \frac{1}{2}\|\alpha \boldsymbol{f}(\boldsymbol{w}) - \mathbf{y}\|^2,$$

and denote $L_0 = L(\alpha \boldsymbol{f}(\boldsymbol{w}(0)))$. Besides, we have

$$\frac{\mathrm{d}\boldsymbol{w}(t)}{\mathrm{d}t} = \dot{\boldsymbol{w}}(t) = -\nabla_{\boldsymbol{w}} L(\alpha \boldsymbol{f}(\boldsymbol{w})) = -\alpha \mathbf{J}_t^\top (\nabla L)(\alpha \boldsymbol{f}(\boldsymbol{w}(t))),$$

where $\nabla_{\boldsymbol{w}}$ is w.r.t. $\boldsymbol{w}$, ∇L is w.r.t. z in $L(z)$, and the Jacobian matrix $\mathbf{J}_t$ is equivalent to the Hessian matrix introduced in our previous lectures. To be more specific, we have

$$\mathbf{J}_t = \mathbf{J}_{\boldsymbol{w}(t)} = (\nabla f(\mathbf{x}_1, \boldsymbol{w}(t)), \dots, \nabla f(\mathbf{x}_n, \boldsymbol{w}(t)))^\top \in \mathbb{R}^{n \times p}.$$

We will also introduce another gradient flow, which is on a linear approximation of the network. By the first-order Taylor expansion, we obtain

$$\boldsymbol{f}_0(\boldsymbol{u}) = \boldsymbol{f}(\boldsymbol{w}(0)) + \mathbf{J}_0(\boldsymbol{u} - \boldsymbol{u}(0)), \text{ where } \boldsymbol{u}(0) = \boldsymbol{w}(0).$$

Now we introduce the gradient flow for training the linear function $\boldsymbol{f}_0(\boldsymbol{u})$:

$$\frac{\mathrm{d}\boldsymbol{u}(t)}{\mathrm{d}t} = \dot{\boldsymbol{u}}(t) = -\nabla_{\boldsymbol{u}} L(\alpha \boldsymbol{f}_0(\boldsymbol{u}(t))) = -\alpha \mathbf{J}_0^\top (\nabla L)(\alpha \boldsymbol{f}_0(\boldsymbol{u}(t))).$$

The goal of this lecture is to show that the loss decreases exponentially fast. Remember that in our last lecture, if we do random initialization, we showed that the spectrum of the Hessian matrix is far away from 0, with high probability. To generalize the results to general neural networks, we first introduce some assumption.

Assumption 1.

$$
\begin{aligned}
&\operatorname{rank}(\mathbf{J}_0) = n, p > n,\\
&\sigma_{\min} \triangleq \sigma_{\min}(\mathbf{J}_0) = \sqrt{\lambda_{\min}(\mathbf{J}_0\mathbf{J}_0^\top)} > 0,\\
&\sigma_{\max} \triangleq \sigma_{\max}(\mathbf{J}_0).\\
&\exists \beta > 0, \|\mathbf{J}_{\boldsymbol{w}} - \mathbf{J}_{\boldsymbol{v}}\| \le \beta\|\boldsymbol{w} - \boldsymbol{v}\|.
\end{aligned}
$$

Next, we aim to show the following theorem.

Theorem 2. Under Assumption 1, if the scaling parameter $\alpha \ge \frac{6\beta\sqrt{8\sigma_{\max}^2 L_0}}{\sigma_{\min}^3}$, we have

$$
\begin{aligned}
\max\left\{L(\alpha\boldsymbol{f}(\boldsymbol{w}(t))), L(\alpha\boldsymbol{f}_0(\boldsymbol{w}(t)))\right\} &\le L_0 \exp\left(-t\frac{\alpha^2\sigma_{\min}^2}{2}\right),\\
\max\left\{\|\boldsymbol{w}(t) - \boldsymbol{w}(0)\|, \|\boldsymbol{u}(t) - \boldsymbol{u}(0)\|\right\} &\le \frac{3\sqrt{8\sigma_{\max}^2 L_0}}{\alpha\sigma_{\min}^2}.
\end{aligned}
$$

To prove Theorem 2, we introduce the following lemmas.

Lemma 3.

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\alpha\boldsymbol{f}(\boldsymbol{w}(t)) &= \alpha\mathbf{J}_t\dot{\boldsymbol{w}}(t) = -\alpha^2\mathbf{J}_t\mathbf{J}_t^\top\nabla L(\alpha\boldsymbol{f}(\boldsymbol{w}(t)))\\
&= -\alpha^2\mathbf{J}_t\mathbf{J}_t^\top(\alpha\boldsymbol{f}(\boldsymbol{w}(t)) - \mathbf{y})\\
\frac{\mathrm{d}}{\mathrm{d}t}\alpha\boldsymbol{f}_0(\boldsymbol{u}(t)) &= -\alpha\mathbf{J}_0\mathbf{J}_0^\top(\alpha\boldsymbol{f}_0(\boldsymbol{u}(t)) - \mathbf{y}).
\end{aligned}
$$

We consider some general dynamics as follows. Once we consider this more general setting, we can see the intuition more clearly.

Lemma 4. Suppose $\dot{\boldsymbol{z}}(t) = -Q(t)\nabla L(\boldsymbol{z}(t)), \forall t \in [0, T]$. If $\lambda = \inf_{t\in[0,T]}\lambda_{\min}(Q_t) > 0$, then for $t \in [0, T]$, $L(\boldsymbol{z}(t)) \le L(\boldsymbol{z}(0))\exp(-2\lambda t)$.

Proof.

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}L(\boldsymbol{z}(t)) &= \frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\|\boldsymbol{z}(t) - \mathbf{y}\|^2\\
&= \frac{1}{2}\langle\dot{\boldsymbol{z}}(t), \boldsymbol{z}(t) - \mathbf{y}\rangle\\
&= \langle -Q_t(\boldsymbol{z}(t) - \mathbf{y}), \boldsymbol{z}_t - \mathbf{y}\rangle\\
&\le -\lambda_{\min}(Q_t)\|\boldsymbol{z}(t) - \mathbf{y}\|^2\\
&\le -\lambda\|\boldsymbol{z}(t) - \mathbf{y}\|^2\\
&= 2\lambda L(\boldsymbol{z}(t)).
\end{aligned}
$$

Next, we can use Grönwall's inequality to get the final bound. $\square$

Lemma 5. Suppose $\dot{\boldsymbol{v}}(t) = -\boldsymbol{s}(t)^\top \nabla L(g(\boldsymbol{v}(t)))$ where $\boldsymbol{s}(t) = \mathbf{S}_t$ which is the Jacobian of g. Let $Q_t = \mathbf{S}_t \mathbf{S}_t^\top$ and assume $\forall t \in [0,T], \lambda_i(Q_t) \in [\lambda, \lambda_1]$. Then $\forall t \in [0,T]$,

$$\|\boldsymbol{v}(t) - \boldsymbol{v}(0)\| \le \frac{2\lambda_1 L(g(\boldsymbol{v}(0)))}{\lambda}.$$

Proof.

$$\begin{aligned}
\|\boldsymbol{v}(t) - \boldsymbol{v}(0)\| &= \|\int_0^t \dot{\boldsymbol{v}}(s)\mathrm{d}s\| \\
&\le \int_0^t \|\dot{\boldsymbol{v}}(s)\|\mathrm{d}s \\
&= \int_0^t \|\boldsymbol{s}(s)^\top (g(\boldsymbol{v}(s)) - \mathbf{y})\|\mathrm{d}s \\
&\le \sqrt{\lambda_1} \int_0^t \|g(\boldsymbol{v}(s)) - \mathbf{y}\|\mathrm{d}s \\
&\le \sqrt{\lambda_1} \int_0^t \|g(\boldsymbol{v}(0)) - \mathbf{y}\| \exp(-\lambda s)\mathrm{d}s
\end{aligned}$$

where the last step is by applying Lemma 4 on $\boldsymbol{z}(t) = g(\boldsymbol{v}(t))$ and noting $\dot{\boldsymbol{z}}(t) = -\mathbf{S}_t \mathbf{S}_t^\top \nabla L(\boldsymbol{z}(t))$. □

By the two lemmas above, we can easily get the results for the linear function $\boldsymbol{f}_0(\boldsymbol{u})$:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\alpha \boldsymbol{f}_0(\boldsymbol{u}(t)) &= -\alpha^2 \mathbf{J}_0 \mathbf{J}_0^\top (\alpha \boldsymbol{f}_0(\boldsymbol{u}(t) - \mathbf{y})) \\
L(\alpha \boldsymbol{f}_0(\boldsymbol{u}(t))) &\le L_0 \exp(-2t\alpha^2 \sigma_{\min}^2) \\
\|\boldsymbol{u}(t) - \boldsymbol{u}(0)\| &\le \frac{\sqrt{2\alpha^2 \sigma_{\max}^2 L_0}}{\alpha^2 \sigma_{\min}}.
\end{aligned}$$

To get the results for $\boldsymbol{f}(\boldsymbol{w})$, we will need to show that $\mathbf{J}_t$ is bounded.

Lemma 6. If $\boldsymbol{w}$ satisfies $\|\boldsymbol{w} - \boldsymbol{w}(0)\| \le B \triangleq \sigma_{\min}/(2\beta)$, then

$$\sigma_{\min}(\mathbf{J}_{\boldsymbol{w}}) \ge \frac{\sigma_{\min}}{2}, \quad \sigma_{\max}(\mathbf{J}_{\boldsymbol{w}}) \le \frac{3\sigma_{\max}}{2}.$$

Proof. First consider $\sigma_{\max}(\mathbf{J}_{\boldsymbol{w}}) = \|\mathbf{J}_{\boldsymbol{w}}\|$. Recall that β is the Lipschitz constant for the Jacobian.

$$\begin{aligned}
\|\mathbf{J}_{\boldsymbol{w}}\| &= \|\mathbf{J}_{\boldsymbol{w}} - \mathbf{J}_0 + \mathbf{J}_0\| \\
&\le \|\mathbf{J}_{\boldsymbol{w}} - \mathbf{J}_0\| + \|\mathbf{J}_0\| \\
&= \beta B + \sigma_{\max} \\
&\le \frac{\sigma_{\min}}{2} + \sigma_{\max} \\
&\le \frac{3\sigma_{\max}}{2}.
\end{aligned}$$

Next consider $\sigma_{\min}(\mathbf{J}_{\boldsymbol{w}})$. By definition,

$$\sigma_{\min}(\mathbf{J}_{\boldsymbol{w}}) = \lambda_{\min}(\mathbf{J}_{\boldsymbol{w}}\mathbf{J}_{\boldsymbol{w}}^\top) = \min_{\|\boldsymbol{v}\|=1} \boldsymbol{v}^\top \mathbf{J}_{\boldsymbol{w}}\mathbf{J}_{\boldsymbol{w}}^\top \boldsymbol{v} = \min_{\|\boldsymbol{v}\|=1} \|\mathbf{J}_{\boldsymbol{w}}^\top \boldsymbol{v}\|^2.$$

Let $\mathbf{A}_{\boldsymbol{v}} = (\mathbf{J}_{\boldsymbol{w}} - \mathbf{J}_0)^\top \boldsymbol{v}$, and $\mathbf{B}_{\boldsymbol{v}} = \mathbf{J}_0^\top \boldsymbol{v}$. Then we have

$$\begin{aligned} \|\mathbf{J}_{\boldsymbol{w}}^\top \boldsymbol{v}\|^2 &= \|(\mathbf{J}_{\boldsymbol{w}} - \mathbf{J}_0 + \mathbf{J}_0)^\top \boldsymbol{v}\| \\ &= \|\mathbf{A}_{\boldsymbol{v}} + \mathbf{B}_{\boldsymbol{v}}\| \\ &= \|\mathbf{A}_{\boldsymbol{v}}\|^2 + 2\langle \mathbf{A}_{\boldsymbol{v}}, \mathbf{B}_{\boldsymbol{v}}\rangle + \|\mathbf{B}_{\boldsymbol{v}}\|^2 \\ &\geq \|\mathbf{A}_{\boldsymbol{v}}\|^2 - 2\|\mathbf{A}_{\boldsymbol{v}}\|\|\mathbf{B}_{\boldsymbol{v}}\| + \|\mathbf{B}_{\boldsymbol{v}}\|^2. \end{aligned} \tag{1}$$

Note that

$$\begin{aligned} &\|\mathbf{B}_{\boldsymbol{v}}\| = \|\mathbf{J}_0^\top \boldsymbol{v}\| \geq \sigma_{\min}, \\ &\|\mathbf{A}_{\boldsymbol{v}}\| \leq \|(\mathbf{J}_{\boldsymbol{w}} - \mathbf{J}_0)^\top \boldsymbol{v}\| \leq \|\mathbf{J}_{\boldsymbol{w}} - \mathbf{J}_0\|\|\boldsymbol{v}\| = \|\mathbf{J}_{\boldsymbol{w}} - \mathbf{J}_0\| \leq \beta B = \sigma_{\min}/2. \end{aligned} \tag{2}$$

Combining equation 1 and equation 2 leads to the desired results. □

With this lemma which bounds the spectrum and the other two key lemmas above, we have

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\alpha \boldsymbol{f}(\boldsymbol{w}(t)) &= -\alpha^2 \mathbf{J}_t \mathbf{J}_t^\top (\alpha \boldsymbol{f}(\boldsymbol{w}(t)) - \mathbf{y}) \\ \lambda &= \alpha^2 \frac{\sigma_{\min}^2}{4} \\ L(\alpha \boldsymbol{f}(\boldsymbol{w}(t))) &\leq L_0 \exp(-t\alpha^2 \sigma_{\min}^2/2) \\ \|\boldsymbol{w}(t) - \boldsymbol{w}(0)\| &\leq \frac{\sqrt{\frac{9\sigma_{\min}^2}{2}\alpha^2 L_0}}{\alpha^2 \sigma_{\min}^2/4} = \frac{2\sqrt{8\sigma_{\max}^2 L_0}}{\alpha \sigma_{\min}^2} \triangleq B'. \end{aligned}$$

Now to complete the proof of the Theorem 2, it is sufficient to ensure $B' \leq B$. This is guaranteed by the condition on α in the theorem.