Mathematics Theory of Neural Network Models **Summer 2019**

Lecture 7: Explicit & implicit regularization for two-layer neural networks

Lecturer: Chao Ma *Scribe: Zhongtian Zheng, Minjie Yu*

1 Two-layer network using gradient descent

7.1 Two-layer network

Definition 7.1 (Two-layer network) *Let $m \in \mathbb{N}$,$a_k(0) = \pm 1$,$w_k(0) \sim \pi$.Then we write*

$$f(x; a, w) = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} a_k \sigma(w_k^T x)$$

where π is a distribution on $\mathbb{R}^d$,$x \in \mathbb{R}^d$.

Let

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

then we have gradient flow:

$$\frac{d}{dt} w_k = -\frac{\partial L(w)}{\partial w_k} = -\frac{1}{\sqrt{m}} \sum_{i=1}^{n} (f(x_i; a, w) - y_i) a_k \sigma'(w_k^T x_i) x_i$$

Let

$$u_i(t) = f(x_i; a(0), w(t))$$

then we have:

$$\begin{aligned}
\frac{d}{dt} u_i(t) &= \frac{1}{\sqrt{m}} \sum_{k=1}^{m} a_k \sigma'(w_k^T x_i) \frac{d}{dt} w_k^T(t) x_i \\
&= \frac{1}{\sqrt{m}} \sum_{k=1}^{m} a_k \sigma'(w_k^T x_i) \left[-\frac{1}{\sqrt{m}} \sum_{j=1}^{n} (f(x_j; a, w) - y_j) a_k \sigma'(w_k^T x_i) x_j \right] x_i \\
&= \sum_{j=1}^{n} (y_i - u_i(t)) \frac{1}{m} \sum_{k=1}^{m} \sigma'(w_k^T(t) x_i) \sigma'(w_k^T(t) x_j) x_i^T x_j
\end{aligned}$$

Let

$$H(t) : H_{i,j} = \frac{1}{m} \sum_{k=1}^{m} \sigma'(w_k^T(t) x_i) \sigma'(w_k^T(t) x_j) x_i^T x_j$$

$$L(t) = \frac{1}{2} ||y - u(t)||^2$$

then we have:

$$\frac{d}{dt}L(t) = -(y-u(t))^T H(t)(y-u(t))$$

As a result, if $\lambda(H(t)) > \lambda_0 > 0$, $\frac{d}{dt}L(t) \to 0$. Then we study $H(t)$

7.2 H(t)

Notice:

- $H(t) = \psi(t)^T \psi(t) \geq 0$
- At initialization, $H(0) \approx \mathbb{E}_w \sigma'(w^T x_i)\sigma'(w^T x_j) x_i^T x_j = H^{\infty}_{i,j}$.
- Strategy for convergence:
  - $H(0) \succ \lambda I$
  - $H(t) \approx H(0)$,when m suffiently large.

First assume $\lambda_{min}(H^{\infty}) \stackrel{\text{def}}{=} \lambda_0 > 0$(Holds true if $x_i \not\parallel x_j$)

Lemma 7.2 *Let* $m = \Omega(\frac{n^2}{\lambda_0^2}\log(\frac{n}{\delta}))$,*then* $w.p \geq 1-\delta, ||H(0)-H^{\infty}||_2 \leq \frac{\lambda_0}{4}$, *which means* $\lambda_{min}(H(0)) \geq \frac{3}{4}\lambda_0$.

Proof: By Hoeffding, $|H_{i,j}(0) - H^{\infty}_{i,j}| \leq \frac{2\sqrt{log\frac{1}{\delta'}}}{\sqrt{m}}, w.p \geq 1-\delta'$
Let $\delta = n^2\delta', w.p \geq 1-\delta$, for $\forall i,j$ we have:

$$|H_{i,j}(0) - H^{\infty}_{i,j}| \leq \frac{4\sqrt{log\frac{n}{\delta}}}{\sqrt{m}}$$

$$\begin{aligned} \Rightarrow ||H(0)-H^{\infty}||_2^2 &\leq ||H(0)-H^{\infty}||_F^2 \\ &\leq \sum_{i,j} |H_{i,j}(0) - H^{\infty}_{i,j}|^2 \\ &\leq \frac{16n^2 log\frac{n}{\delta}}{m} \end{aligned}$$

if$m = \Omega(\frac{n^2}{\lambda_0^2}\log(\frac{n}{\delta}))$,then $w.p \geq 1-\delta, ||H(0)-H^{\infty}||_2 \leq \frac{\lambda_0}{4}$ ■

Lemma 7.3 $w_1(0), w_2(0), \ldots, w_n(0) \sim^{i.i.d} \mathcal{N}(0,1)$, *w.p. at least 1-*$\delta$*, for* $w_1, \ldots, w_m$ *satisfying:*$\|w_k - w_k(0)\|_2 \leqslant \frac{c\delta\lambda_0}{n^2}$*,then:*$\|H(w) - H\left(w(0)\right)\|_2 < \frac{\lambda_0}{4}$.

Proof: Let $A_{ik} = \{\exists w : \|w - w_k(0)\|_2 \leqslant \frac{c\delta\lambda_0}{n^2}, L\{x_i^T w \geqslant 0\} \neq L\{x_i^T w_k(0) \geqslant 0\}\}$

$$\begin{aligned}
E\left|H_{ij}(w) - H_{ij}(w(0))\right| &= \frac{1}{m} E\left|x_i^T x_j \sum_{k=1}^m [L(w_k^T(0)x_i \geqslant 0, w_k^T(0)x_j \geqslant 0) - L(w_k^T x_i \geqslant 0, w_k^T x_j \geqslant 0)]\right| \\
&\leqslant \frac{1}{m}\left|x_i^T x_j\right| \sum_{k=1}^m E\left|L(w_k^T(0)x_i \geqslant 0, w_k^T(0)x_j \geqslant 0) - L(w_k^T x_i \geqslant 0, w_k^T x_j \geqslant 0)\right| \\
&\leqslant \frac{1}{m}\sum_{k=1}^m P(A_{ik} \cup A_{jk}) \\
A_{ik} \subset \{|w_k^T(0)x_i| \leqslant \frac{c\delta\lambda_0}{n^2}\} & \\
&\Rightarrow P(A_{ik}) \leqslant \frac{2}{\sqrt{2\pi}}\frac{c\delta\lambda_0}{n^2} \\
&\Rightarrow E|H_{ij}(w) - H_{ij}(w_0)| \leqslant \frac{4}{\sqrt{2\pi}}\frac{c\delta\lambda_0}{n^2} \\
&\Rightarrow E\sum_{ij}|H_{ij}(w) - H_{ij}(w_0)| \leqslant \frac{4c\delta\lambda_0}{\sqrt{2\pi}}
\end{aligned}$$

$$w.p \geq 1-\delta, \sum_{ij}|H_{ij}(w) - H_{ij}(w_0)| \leqslant \frac{4c\delta\lambda_0}{\sqrt{2\pi}}$$

$$||H(w) - H(w(0))||_2 \leqslant \frac{\lambda_0}{4}$$

■

With lemmas above,We have$\lambda_{min}(H(w)) \geqslant \frac{3}{4}\lambda_0 - \frac{1}{4}\lambda_0 = \frac{1}{2}\lambda_0$.

$$\begin{aligned}
if for s \in [0,t], \lambda_{min}(H(w(s))) \geqslant \frac{1}{2}\lambda_0, then: \|y - u(t)\|_2^2 \leqslant e^{-\lambda_0 t}\|y - u(0)\|_2^2 \\
\|W_k(t) - w_k(0)\|_2 \leqslant \left\|\int_0^t |\frac{dw_k(s)}{dt}|ds\right\|_2 \\
\leqslant \frac{\sqrt{n}\|y - u(0)\|_2}{\sqrt{m}\lambda_0}
\end{aligned}$$

7.3 Mean-field

$$f(x; a, w) = \frac{1}{m}\sum_{k=1}^m a_k \sigma(w_k^T x)$$

$a \in \mathbb{R}, w \in \mathbb{R}^d, (a, w) \in \mathbb{R}^{d+1}$,Thus,$\phi$:Distribution on$\mathbb{R}^{d+1}$

$$\phi = \frac{1}{m}\sum_{k=1}^m \delta_{(a_k, w_k)} = \mathbb{E}_{(a,w)\sim\phi} a\sigma(w^T x)$$

$$f(x; a, w) = \int a\sigma(w^T x)\phi(da, dw)$$

References