Mathematical Theory of Neural Network Models

Lecture 6: Two-layer neural network and Barron space

July 12

Lecturer: Lei Wu *Scribe: MuSu Yuan, Pu Yang*

We consider a two-layer neural network with the activation function RelU:

$$f(x;\theta) = \sum_{k=1}^{m} a_k \sigma(b_k^T x)$$

which $x \in [-1,1]^d$, i.e.$||x||_\infty \leqslant 1$. And the activation function RelU: $\sigma(t) = \max(0,t)$.

The model has a special property called **scaling invariance**, which means a system, function, or statistic has scale invariance if changing the scale by a certain amount does not change the system, function, or statistics shape or properties.

Our task is still considering the optimization problem:

$$\hat{\theta}_n = \arg\min[L_n(\theta) + \lambda||\theta||]$$

Definition 0.1 (Path Norm)**.** The **Path Norm** of a vector θ, denoted $||\theta||_P$, is defined as:

$$||\theta||_P = \sum_{k=1}^{m} |a_k|||b_k||_1 = \sum_{k=1}^{m} |a_k|||b_k||_\Omega$$

We have the inequality that: if $\frac{1}{p} + \frac{1}{q} = 1$ then

$$|b^T x| \leqslant ||b||_p||x||_q$$

The following theorem is the most important ones of the whole lecture.

Theorem 0.2 (Approximation)**.**

$$\forall f \in B_2(\Omega)$$

there exists a tow-layer neural network of width m
such that

$$\mathbb{E}_x(f(x) - f(x;\widetilde{\theta}))^2 \leqslant \frac{3||f||_{B_2}^2}{m}$$

and

$$||\widetilde{\theta}||_p \leqslant 2||f||_{B_2}$$

*the $||f||_{B_2}$ is the **Barron Norm** which is defined as*

$$||f||_{B_2}^2 = \inf_{(a,\pi)\in\Omega_f} \int a^2(\omega) d\pi(\omega)$$

which

$$\omega \in S^{d-1} = \{||\omega||_1 = 1 \quad \Big| \quad \omega \in \mathbb{R}^d\}$$

Proof. $\exists(\widetilde{a}, \widetilde{\pi})$, s.t.

$$f(x) = \int \bar{a}(w)\sigma(w^T x) d\bar{\pi}(\omega)$$

$$||f||_{B_2}^2 \leqslant \mathbb{E}_{\omega\sim\bar{\pi}}(\bar{a}^2(\omega)) + \epsilon$$

$$a_k = \bar{a}(\omega_k)/m, \omega_k \sim \bar{\pi}(\cdot)$$

$$f(x, \bar{\theta}) = \frac{1}{m}\sum_{k=1}^{m} \bar{a}(\omega_k)\sigma(\omega_k^T x)$$

$$\begin{aligned}
&\mathbb{E}_{\omega_k}(E_x(\frac{1}{m}\sum_{k=1}^{m} \widetilde{a}(\omega_k)\sigma(\omega_k^T x) - f(x))^2) \\
&= \mathbb{E}_x(E_{\omega_k}(\frac{1}{m}\sum_{k=1}^{m} \widetilde{a}(\omega_k)\sigma(\omega_k^T x) - f(x))^2) \\
&= \mathbb{E}_x\frac{1}{m}(E_{\omega}(\sum_{k=1}^{m} \widetilde{a}(\omega)\sigma(\omega^T x) - f(x))^2) \\
&\leqslant \frac{1}{m}\mathbb{E}_{\omega}(\widetilde{a}^2(\omega) \\
&\leqslant ||f^2||_{B_2}
\end{aligned}$$

From the approximation above we know that:

$$\exists\widetilde{\theta}, s.t. L(\bar{\theta}) \leqslant \frac{1}{m}||f||_{B_2}^2$$

$$\mathbb{E}(||\widetilde{\theta}||_{path}) \leqslant ||f||_{B_2}$$

$$\mathbb{E}(L(\widetilde{\theta})) \leqslant \frac{||f||_{B_2}^2}{m}$$

Now we get these two approximation, but we still don't know whether there exist a situation that both inequalities hold at the same time. We consider

$$\mathbb{E}_1 = \{\widetilde{\theta} | L(\widetilde{\theta}) < \frac{3||f||_{B_2}^2}{m}\}$$

$$\mathbb{E}_2 = \{\widetilde{\theta} | ||\widetilde{\theta}||_p \leqslant 2||f||_{B_2}\}$$

We have

$$P(\mathbb{E}_1^c) = P(L(\widetilde{\theta}) > \frac{3||f||_{B_2}^2}{m}) \leqslant \frac{m\mathbb{E}(L(\widetilde{\theta}))}{3||f||_{B_2}^2} \leqslant \frac{1}{3}$$

$$P(\mathbb{E}_2^c) = P(||\widetilde{\theta}||_p > 2||f||_{B_2}^2) \leqslant \frac{\mathbb{E}(||\widetilde{\theta}||_p)}{2||f||_{B_2}^2} \leqslant \frac{1}{2}$$

So that

$$P(\mathbb{E}_1 \bigcap \mathbb{E}_2) = P(\mathbb{E}_1) + P(\mathbb{E}_2) - 1 \geqslant 1 - \frac{1}{3} + 1 - \frac{1}{2} - 1 > 0$$

That is to say there exists a situation that these two inequalities both hold. $\square$

We want to use model:

$$f(x; a) = \sum_{k=1}^{m} a_k \sigma(\omega_k^T x)$$

$$\omega_k \sim \mathcal{U}(S^{d-1})$$

to approximate

$$f^\star(x) = \sigma(\omega_\star^T x)$$

However, in higher dimensionswith a huge probability that $m < e^d, < \omega_k, \omega_\star \geqslant 0$, in other words, $||f^\star||_{\mathcal{H}_{k\pi_0}} = +\infty$. But for a two-layer neural network, $||f^\star||_{B_2} = 1$, $\pi^\star = \delta(\cdot - \omega_\star)$

Definition 0.3. $\mathcal{F}_Q = \{f(x;\theta) \quad \Big| \quad ||\theta||_P \leqslant Q\}$

Theorem 0.4. $Rad_n(\mathcal{F}_Q) \leqslant 2Q\sqrt{\frac{2\log(2d)}{n}}$

Proof.

$$\begin{aligned}
Rad_n(\mathcal{F}_Q) &= \frac{1}{n}\mathbb{E}_\xi \sup_{||Q||_P \leqslant Q} \sum_{i=1}^{n} \xi_i \sum_{k=1}^{m} a_k ||b_k||_1 \sigma(\hat{b}_k^T x_i) \\
&= \frac{1}{n}\mathbb{E}_\xi \sup_{||Q||_P \leqslant Q} \sum_{k=1}^{m} a_k ||b_k||_1 \sum_{i=1}^{n} \xi_i \sigma(\hat{b}_k^T x_i) \\
&\leqslant \frac{Q}{n}\mathcal{E}_\xi \sup_{||b||_1 \leqslant 1} |\sum_{i}^{n} \xi_i \sigma(b^T x_i)|
\end{aligned}$$

$$T = \{\sigma(b^T x) | ||b||_1 \leqslant 1\}$$

notice that

$$\sup_b |g(b)| \leqslant \sup_b g(b) + \sup_b -g(b)$$

so that

$$Rad_n(\mathcal{F}_Q) \leqslant 2Q\sqrt{\frac{2\log(2d)}{n}}$$

$\square$

Theorem 0.5 (A priory bound)**.**

$$L(\hat{\theta}_n) \leqslant \frac{||f^\star||_{B_2}^2}{m} + \lambda||f^\star||_{B_2} + \frac{1}{\sqrt{n}}[||f^\star||_{B_2} + \sqrt{\log(\frac{n}{\delta})}]$$

where $\lambda \geqslant 4\sqrt{\frac{2\log(2d)}{n}}$

Proof. We have proved in Lecture 5

$$\begin{aligned}
L(\hat{\theta}_n) &\leqslant L_n(\hat{\theta}_n) + 4\sqrt{\frac{2\log(2d)}{n}}||\hat{\theta}_n|| + \sqrt{\frac{\log((1+||\hat{\theta}_n||)^2/\delta)}{n}} \\
&\leqslant L_n(\hat{\theta}_n) + \lambda||\hat{\theta}_n|| + \sqrt{\frac{\log((1+||\hat{\theta}_n||)^2/\delta)}{n}} \\
&\leqslant L_n(\widetilde{\theta}) + \lambda||\widetilde{\theta}|| + \sqrt{\frac{\log((1+||\widetilde{\theta}_n||)^2/\delta)}{n}} \\
&\leqslant L(\widetilde{\theta}) + 4\sqrt{\frac{2\log(2d)}{n}}||\widetilde{\theta}_n|| + \sqrt{\frac{\log((1+||\hat{\theta}_n||)^2/\delta)}{n}} + \lambda||\widetilde{\theta}|| + Q_n
\end{aligned}$$

so that

$$||\hat{\theta}_n||_P \leqslant \frac{1}{\lambda}L_n(\widetilde{\theta}) + ||\widetilde{\theta}||_P \leqslant O(\frac{\sqrt{n}}{m}) + ||\widetilde{\theta}||_P$$

□

Now the question is: how to choose a appropriate function f, which makes $||f||_{B_2}$ small so that we can bound the error. The following theorem shows it.

Theorem 0.6. *Let* $f \in C(X)$, $r(f) = \inf_{\hat{f}} \int_{R^d} ||\omega||_1^2 |\hat{f}(\omega)| d\omega$, $\hat{f}$ *is the Fourier transform of an extension of* f *on* R^d*. We have*

$$||f||_{B_2} \leqslant 2r(f) + 2||f'(0)||_1 + 2|f(0)|$$

Then, we have

$$||f||_{B_2} = ploy(d), x^k = x_{k_1}x_{k_2}\cdots x_{k_{m_0}}$$

$$f(x) = \sum_{|k|\leqslant m_0} a_k f(x_{k_1}, \cdots, x_{k_{m_0}})$$

$$||f||_{B_2} \leqslant \sum_{|k|\leqslant m_0} |a_k| ||x^k||_{B_2} \leqslant A_{m_0} \sum_{|k|\leqslant m_0} |a_k|$$

($||x^k||_{B_2}$is not relevant to d)

so that

$$n = \frac{ploy(d)}{\epsilon}$$

Finally, let's see an example.

Example 0.7. *if $f^{\star}(x) = cos(x)$, then*

$$\int (f(x;\widetilde{\theta}) - f^{\star}(x))^2 d\mu(x) \leqslant \frac{C}{m^{\alpha}}, \quad \alpha > 1$$

In fact utilize multi-grid method in numerical analysis, $\alpha = 2$ can be approached which infers that multi-grid method performs well at low dimension situations, but not so well at high dimension ones.