CS : Deep Learning Fall 2020

Lecture 4 — September 8, 2020

Prof. Zhao Song *Scribe: Zhao Song*

1 Data separation implies eigenvalue bound on NTK

Data-separation is a reasonable assumption in deep learning theory. In this notes, we will show a connection between separation parameter δ and the smallest eigenvalue of NTK. Similar as previous lecture note, we will use one-hidden layer neural network with ReLU activation function as an example.

1.1 Neural Tangent Kernel

Definition 1 (Neural Tangent Kernel [JGH18])**.** *Given a set of points* $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\} \subset \mathbb{R}^d \times \mathbb{R}$*. We define Neural Tangent Kernel* $H^{\mathrm{cts}} \in \mathbb{R}^{n \times n}$

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

1.2 Main result

Lemma 2 (Lemma I.1 in [OS20])**.** *Let* $x_1, \cdots, x_n$ *be points in* $\mathbb{R}^d$ *with* $\|x_i\|_2 = 1, \forall i \in [n]$*. Let* $X = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top \in \mathbb{R}^{n \times d}$*. Let* $\delta > 0$ *be the parameter such that*

$$\min_{i \neq j} \{ \|x_i - x_j\|_2, \|x + x_j\|_2 \} \geq \delta$$

Then we have

$$\mathop{\mathbb{E}}_{w \sim \mathcal{N}(0,I_d)} [\phi'(Xw)\phi'(Xw)^\top] \succeq \delta/(100n^2)$$

Lemma 3 (Lemma I.1 in [OS20])**.** *Let* $x_1, \cdots, x_n$ *be points in* $\mathbb{R}^d$ *with* $\|x_i\|_2 = 1, \forall i \in [n]$*. Let* $X = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top \in \mathbb{R}^{n \times d}$*. Let* $\delta > 0$ *be the parameter such that*

$$\min_{i \neq j} \{ \|x_i - x_j\|_2, \|x + x_j\|_2 \} \geq \delta$$

Then we have

$$\mathop{\mathbb{E}}_{w \sim \mathcal{N}(0,I_d)} [(\phi'(Xw)\phi'(Xw)^\top) \circ (XX^\top)] \succeq \delta/(100n^2)$$

1.3 Probability tools

Lemma 4 (Hoeffding bound [Hoe63])**.** *Let* $X_1, \cdots, X_n$ *denote* n *independent bounded variables in* $[a_i, b_i]$*. Let* $X = \sum_{i=1}^n X_i$*, then we have*

$$\Pr[|X - \mathbb{E}[X]| \geq t] \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Lemma 5 (Bernstein inequality [Ber24])**.** *Let* $X_1, \cdots, X_n$ *be independent zero-mean random variables. Suppose that* $|X_i| \leq M$ *almost surely, for all* i*. Then, for all positive* t*,*

$$\Pr\left[\sum_{i=1}^n X_i > t\right] \leq \exp\left(-\frac{t^2/2}{\sum_{j=1}^n \mathbb{E}[X_j^2] + Mt/3}\right).$$

Lemma 6 (Anti-concentration of Gaussian distribution)**.** *Let* $X \sim \mathcal{N}(0, \sigma^2)$*, that is, the probability density function of* X *is given by* $\phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$*. Then*

$$\Pr[|X| \leq t] \in \left(\frac{2}{3}\frac{t}{\sigma}, \frac{4}{5}\frac{t}{\sigma}\right).$$

References

[Ber24] Sergei Bernstein. On a modification of chebyshev's inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, 1(4):38–49, 1924.

[Hoe63] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[OS20] Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. In *IEEE Journal on Selected Areas in Information Theory*. https://arxiv.org/pdf/1902.04674.pdf, 2020.