**CS : Deep Learning** Fall 2020

Lecture 3 — September 8, 2020

*Prof. Zhao Song* *Scribe: Zhao Song*

# 1 Neural Tangent Kernel

Neural Tangent Kernel is introduced by Jacob, Gabriel and Hongler [JGH18] and played a crucial role in showing the convergence of neural network training. In this lecture notes, we will use one-hidden layer neural network as an example.

## 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,

$$W(t+1) = W(t) - \eta \frac{\partial L(W)}{\partial W} \Big|_{W = W(t)}. \tag{2}$$

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. This algorithm takes $Tmnd$ time without using fast matrix multiplication.

**procedure** NNTraining($\{(x_i, y_i)\}_{i\in[n]}$)
  $w_r(0) \sim \mathcal{N}(0, I_d)$ for $r \in [m]$.
  **for** $t = 1 \to T$ **do**
    $u(t) \leftarrow \frac{1}{\sqrt{m}} \sum_{r=1}^m a_r \sigma(w_r(t)^\top X)$  ▷ $u(t) = f(W(t), x, a) \in \mathbb{R}^n$, it takes $O(mnd)$ time
    **for** $r = 1 \to m$ **do**
      **for** $i = 1 \to n$ **do**
        $Q_{i,:} \leftarrow \frac{1}{\sqrt{m}} a_r \sigma'(w_r(t)^\top x_i) x_i^\top$  ▷ $Q_{i,:} = \frac{\partial f(W(t), x_i, a)}{\partial w_r}$, it takes $O(d)$ time
      **end for**
      $\mathrm{grad}_r \leftarrow -Q^\top (y - u(t))$  ▷ $Q = \frac{\partial f}{\partial w_r} \in \mathbb{R}^{n\times d}$, it takes $O(nd)$ time
      $w_r(t+1) \leftarrow w_r(t) - \eta \cdot \mathrm{grad}_r$
    **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 $1-\delta$.*

*Proof.* For every fixed pair $(i,j)$, $H^{\rm dis}_{i,j}$ is an average of independent random variables, i.e.

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

Then the expectation of $H^{\rm dis}_{i,j}$ is

$$\begin{aligned}
\mathbb{E}[H^{\rm dis}_{i,j}] &= \frac{1}{m}\sum_{r=1}^m \mathop{\mathbb{E}}_{w_r\sim\mathcal{N}(0,I_d)}\left[x_i^\top x_j \mathbf{1}_{w_r^\top x_i\geq 0, w_r^\top x_j \geq 0}\right] \\
&= \mathop{\mathbb{E}}_{w\sim\mathcal{N}(0,I_d)}\left[x_i^\top x_j \mathbf{1}_{w^\top x_i\geq 0, w^\top x_j \geq 0}\right] \\
&= H^{\rm cts}_{i,j}.
\end{aligned}$$

For $r\in[m]$, let $z_r = \frac{1}{m}x_i^\top x_j \mathbf{1}_{w_r^\top x_i\geq 0, w_r^\top x_j \geq 0}$. Then $z_r$ is a random function of $w_r$, hence $\{z_r\}_{r\in[m]}$ are mutually independent. Moreover, $-\frac{1}{m}\leq z_r \leq \frac{1}{m}$. So by Hoeffding inequality(Lemma 15) we have for all $t>0$,

$$\begin{aligned}
\Pr\left[|H^{\rm dis}_{i,j} - H^{\rm cts}_{i,j}| \geq t\right] &\leq 2\exp\Big(-\frac{2t^2}{4/m}\Big) \\
&= 2\exp(-mt^2/2).
\end{aligned}$$

Setting $t = (\frac{1}{m}2\log(2n^2/\delta))^{1/2}$, we can apply union bound on all pairs $(i,j)$ to get with probability at least $1-\delta$, for all $i,j\in[n]$,

$$|H^{\rm dis}_{i,j} - H^{\rm cts}_{i,j}| \leq \Big(\frac{2}{m}\log(2n^2/\delta)\Big)^{1/2} \leq 4\Big(\frac{\log(n/\delta)}{m}\Big)^{1/2}.$$

Thus we have

$$\begin{aligned}
\|H^{\rm dis} - H^{\rm cts}\|^2 &\leq \|H^{\rm dis} - H^{\rm cts}\|_F^2 \\
&= \sum_{i=1}^n\sum_{j=1}^n |H^{\rm dis}_{i,j} - H^{\rm cts}_{i,j}|^2 \\
&\leq \frac{1}{m}16n^2\log(n/\delta).
\end{aligned}$$

Hence if $m=\Omega(\lambda^{-2}n^2\log(n/\delta))$ we have the desired result. □

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

[DZPS19] proved a weaker version of Lemma 2. Later, [SY19] improve it from the two perspective : one is the probability, and the other is upper bound on spectral norm.

**Lemma 2** (perturbed $w$, [SY19])**.** *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.* 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 17), we have

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

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 16) 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 [SY19])

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

We state two tools from [DZPS19] without provide a proof.

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

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

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

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

### 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 $\lambda = \lambda_{\min}(H^{\rm cts}) > 0$. 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 $k = 0, 1, 2, \cdots$*

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

**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 $r \in [m]$*

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

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

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

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

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

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

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

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

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

**Choice of $\eta$ and $R$.** Next, we want to choose $\eta$ and $R$ such that

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

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

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

This implies

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

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. The proof details of the following Claims can be found in [SY19].

**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*

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

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

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

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

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

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

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

### 1.7 Probability tools

**Lemma 15** (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*

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

**Lemma 16** (Bernstein inequality [Ber24])**.** *Let $X_1, \cdots, X_n$ be independent zero-mean random variables. Suppose that $|X_i| \le M$ almost surely, for all $i$. Then, for all positive $t$,*

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

**Lemma 17** (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*

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

## 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.

[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.

[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.

[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.