Lecture 6: NTK continued.

$$f(x;W) := \frac{1}{\sqrt{m}} \sum_{j=1}^{m} a_j \sigma(w_j^T x)$$

(a_j: ± 1; w_j: Gaussian)

$$f_0(x;W) := f(x;W_0) + \langle \nabla f(x;W_0), W - W_0 \rangle$$

NTK story:

near initialization, with large width, $f \approx f_0$, f_0 still universal approximator.

"kernel"

* In ML, whenever we have "linear" (affine) predictor over some feature mapping $x \mapsto \Phi(x)$,

we consider "kernel" $k(x,x') := \langle \Phi(x), \Phi(x') \rangle$

(for us, $\Phi_m(x) \overset{\sim}{:=} \nabla f(x;W_0)$.)

$\Phi_m(x)$: r.v.

$$k_m(x,x') := \langle \nabla f(x;W_0), \nabla f(x';W_0) \rangle$$

$m \to \infty$ ⟶ ♯ problem 3 in hwk

$f \approx f_0$ near init for large width

Prop. $|\sigma''| \le \beta$, $\|x\| \le 1$, $|a_j| \le 1$, $\forall W$

$$|f(x;W) - f_0(x;W)| \le \frac{\beta}{2\sqrt{m}} \|W - V\|_F^2.$$

Lemma. Let $B \ge 0$ be given, & fixed $\|x\| \le 1$.
With probability $\ge 1-\delta$ (over W_0)
for any $W \in \mathbb{R}^d$ with $\|W - W_0\|_2 \le B$,

$$|f(x;W) - f_0(x;W)| \le \frac{2B^{4/3} + B\ln(1/\delta)^{1/4}}{m^{1/6}}.$$

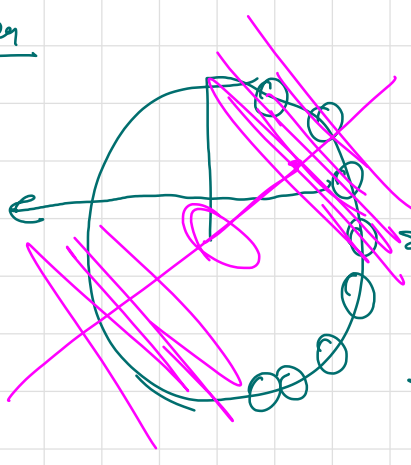
Remark. Brute force: (ReLU)

$\sigma(z) = z\,\sigma'(z)$

$$|f(x;W) - f_0(x;W)|$$
$$= \left| \frac{1}{\sqrt{m}} \sum_j a_j \left(\sigma(w_j^T x) - \sigma(w_{0,j}^T x) - \sigma'(w_{0,j}^T x) x^T (w_j - w_{0,j}) \right) \right|$$
$$= \left| \frac{1}{\sqrt{m}} \sum_j a_j \left(\sigma'(w_j^T x) - \sigma'(w_{0,j}^T x) \right) w_j^T x \right|$$
$$\le \frac{1}{\sqrt{m}} \sum_j |a_j| \cdot |\sigma'(w_j^T x) - \sigma'(w_{0,j}^T x)| \cdot |w_j^T x|$$
$$\le \frac{1}{\sqrt{m}} \sum_j \|w_j\| \le \frac{1}{\sqrt{m}} \sqrt{m} \cdot \sqrt{\sum_j \|w_j\|^2} = \|W\|_F^2.$$

Remark. First such bound was
Allen-Zhu / Li / Song '18
(multi-layer; this version one of mine.)

Proof idea



Keep $\|W - W_0\|_F$ fixed, & increase m:
$\operatorname{sgn}(w_j^T x) \approx \operatorname{sgn}(w_{0,j}^T x)$ for "most" j

Lemma. For any $\tau > 0$, & any $x \ne 0$, with probability $\ge 1-\delta$,

$$\sum_j \mathbb{1}\left[|w_{0,j}^T x| \le \tau \|x\| \right] \le \tau m + \sqrt{\frac{m}{2} \ln \frac{1}{\delta}}$$

Proof: Define indicator r.v. $P_j := \mathbb{1}\left[|w_{0,j}^T x| \le \tau \|x\| \right]$ (↓ $x \ne 0$)

By rotational invariance of $w_{0,j}$, $P_j := \mathbb{1}[|g| \le \tau]$ ← univariate gaussian

$$\# P_j = \int_{-\tau}^{\tau} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \le \int_{-\tau}^{\tau} \frac{1}{\sqrt{2\pi}}\, dx = \frac{2\tau}{\sqrt{2\pi}} \le \tau.$$

By Hoeffding's inequality, w/ $\ge 1-\delta$

$$\sum_j P_j \le m \cdot \# P_1 + \sqrt{\frac{m}{2} \ln \frac{1}{\delta}} \le m\tau + \sqrt{\frac{m}{2} \ln \frac{1}{\delta}}.$$

Proof of $f \approx f_0$ for ReLU.

If $x = 0$, $f(x;W) = 0 = f_0(x;W)$ $\forall W$,
proof is complete; henceforth, $\|x\| > 0$.

Let W be given with $\|W - W_0\|_2 \le B$,

$$S_1 := \{ j \in \{1, \dots, m\} : |w_{0,j}^T x| \le r \cdot \|x\| \},$$
$$S_2 := \{ j \in \{1, \dots, m\} : \|w_j - w_{0,j}\| \ge r \},$$
$$S := S_1 \cup S_2,$$

where r is a free parameter optimized shortly.

By the Hoeffding proof above,

$$|S_1| \le rm + \sqrt{m \ln \frac{1}{\delta}}.$$

For $|S_2|$,

$$B^2 \ge \|W - W_0\|_2^2 = \sum_{j=1}^m \|w_j - w_{0,j}\|^2 \ge \sum_{j \in S_2} \|w_j - w_{0,j}\|^2$$
$$\ge r^2 \cdot |S_2|,$$
$$\Rightarrow |S_2| \le \frac{B^2}{r^2}.$$

For any $j \notin S$, if $w_{0,j}^T x > 0$ ($w_{0,j}^T x < 0$ is analogous)

$$w_j^T x = w_{0,j}^T x + (w_j - w_{0,j})^T x$$
$$\ge r \cdot \|x\| - \|w_j - w_{0,j}\| \cdot \|x\| > \|x\| (r - r) = 0,$$

$$\Rightarrow \operatorname{sgn}(w_{0,j}^T x) = \operatorname{sgn}(w_j^T x).$$

Summarizing, S satisfies

$$|S| \le rm + \sqrt{m \ln \tfrac{1}{\delta}} + \frac{B^2}{r^2} \quad \text{and} \quad j \notin S \Rightarrow \begin{cases} \operatorname{sgn}(w_j^T x) \\ = \operatorname{sgn}(w_{0,j}^T x). \end{cases}$$

$r := \frac{B^{2/3}}{m^{1/3}}$: $= (Bm)^{2/3} + \sqrt{m \ln \frac{1}{\delta}} + (Bm)^{2/3}$
$= 2(Bm)^{2/3} + \sqrt{m \ln \frac{1}{\delta}}.$

$$|f(x;W) - f_0(x;W)|$$
$$\le \frac{1}{\sqrt{m}} \sum_j |a_j| \cdot |\sigma'(w_j^T x) - \sigma'(w_{0,j}^T x)| \cdot |w_j^T x|$$
$$\le \frac{1}{\sqrt{m}} \sum_{j \in S} |w_j^T x| \mathbb{1}\left[\operatorname{sgn}(w_j^T x) \ne \operatorname{sgn}(w_{0,j}^T x) \right]$$
$$\le \frac{1}{\sqrt{m}} \sum_{j \in S} |w_j^T x - w_{0,j}^T x| \cdot \left| \mathbb{1}\left[\operatorname{sgn}(w_j^T x) \ne \operatorname{sgn}(w_{0,j}^T x) \right] \right| \le 1$$
$$\le \frac{1}{\sqrt{m}} \sum_{j \in S} \|w_j - w_{0,j}\| \cdot \|x\| \le 1$$
$$\le \frac{1}{\sqrt{m}} \sqrt{|S|} \cdot \|W - W_0\|_F^2$$

Offline hours

if $\operatorname{sgn}(w_j^T x) \neq \operatorname{sgn}(w_{0,j}^T x)$

$\Rightarrow |w_j^T x - w_{0,i}^T x| = |w_j^T x| + |w_{0,i}^T x|$

$\geq |w_j^T x|$

$\Rightarrow \sum_{j \in S} |w_j^T x| \cdot \mathbb{1}\{\operatorname{sgn}(w_i^T x) \neq \operatorname{sgn}(w_{0,j}^T x)\}$

$\leq \sum_{j \in S} |w_j^T x - w_{0,j}^T x| \cdot \mathbb{1}$ ⟶

$\sum_{j \in S} \|w_j\| \cdot \|x\|$

$\leq \sqrt{|S|} \cdot \|W\|_F^2$

$\mathbb{E}\|W_0\|_F^2 = \sum_{j=1}^{m} \sum_{i=1}^{d} w_{j,i}^2 = m \cdot d$

$\frac{1}{\sqrt{n}} \cdot \sqrt{m^{2/3}} \cdot md$

$\frac{1}{n^{1/6}}$