Lecture 5:

Ann.

* TA OH
* hw1 Q?
* typed notes

Lec 1-4 constructive apx
Lec 5-8 near initialization
?-9 / over parameterization

Lec 1-4: care about # nodes to apx.

Remark: training one ReLU has negative results & many papers.

GD seems to care about two function classes:

① Near initialization ② low norm.

Contradictory: ① is close (large) Gaussian initialization
② measured wrt origin.

Next few lectures: ① (aka NTK) in these steps:

(A) Near initialization ⇒ near Taylor expansion around initialization

[critical point: scaling of error with m.]

(B) Characterize Taylor expansion as $m \to \infty$.
Remark: This will reveal a "signal/noise" separation; this lets us choose our scaling constants

(C) kernel perspective ("neural tangent kernel").

[(D) Refined complexity estimates.]

Remark: On one hand, Taylor expansion not a good approximator of network in practice;
on the other hand, this perspective has good predictive power.

Definition (NTK lectures.)

minimum norm element of Clarke differential

$$F_0(x; w) := F(x; w_0) + \langle \boxed{\partial} F(x; w_0), w - w_0 \rangle.$$

Moreover we will typically only vary V (where $w = (a, V)$), so that optimization problem is "convex".

E.g., $F(x; (a, V)) = \sum a_j \sigma(v_j^T x)$; $\qquad F_0(x; (a, V)) := \sum a_{0,j} \sigma(v_{0,j}^T x)$
$\qquad + \sum_j a_{0,j} \sigma'(v_{0,j}^T x) x^T (v_j - v_{0,j})$.

Remark (Initialization)

* These lectures: $a_j \sim \mathrm{Unif}(\{-1, +1\})$, $v_j \sim \mathcal{N}(0, I_d)$
* Literature, (NTK)
  (A) $a_j \sim \mathrm{Unif}(\{\frac{-1}{\sqrt{m}}, \frac{+1}{\sqrt{m}}\})$, $v_j \sim \mathcal{N}(0, I_d)$
  (B) $a_j \sim \mathrm{Unif}(\{\frac{-\varepsilon}{\sqrt{m}}, \frac{\varepsilon}{\sqrt{m}}\})$, $v_j \sim \mathcal{N}(0, I_d)$
  (C) a_j = some deterministic choices so that $F(x; w_0) = 0$. $\quad v_j \sim \mathcal{N}(0, I_d)$

  "symmetric"
* "Pytorch": (A) $a_j \sim \mathrm{Unif}([\frac{-1}{\sqrt{m}}, \frac{+1}{\sqrt{m}}])$, $v_{j,i} \sim \mathrm{Unif}([\frac{-1}{\sqrt{d}}, \frac{+1}{\sqrt{d}}])$

  (tensorflow different).

Open: uniform vs Gaussian

Another scaling: mean-field

(A) Near initialization $\Rightarrow$ Near Taylor expansion

Warm-up: "smooth" activations.

Proposition. Suppose σ is β-smooth $\left(|\sigma'(r)-\sigma'(s)| \le \beta|r-s|\right)$.

Then, given any $V \in \mathbb{R}^{m\times d}$,

$$\left|F(x;V) - F_0(x;V)\right| \le \frac{\beta\|a\|_\infty \|x\|^2 \underbrace{\|V-V_0\|^2}}{2}.$$

Remark. (a) open to prove matching for the ReLU. (c) no m in rhs.
(b) this theorem has no randomness.

Proof. Note

$$\left|\sigma(r) - (\sigma(s)+\sigma'(s)(r-s))\right| = \left|\int_s^r \sigma'(t)\,dt - \int_s^r \sigma'(s)\,dt\right| \le \int_s^r |\sigma'(t)-\sigma'(s)|\,dt \le \frac{\beta(r-s)^2}{2}.$$

Therefore

$$\left|F(x;V) - F_0(x;V)\right| = \left|\sum_{j=1}^m a_j\left(\sigma(v_j^T x) - \left(\sigma(v_{0,j}^T x) + \sigma'(v_{0,j}^T x)\cdots \cdots\langle x, v_j - v_{0,j}\rangle\right)\right)\right|$$

$$\le \|a\|_\infty \sum_{j=1}^m \left|\sigma(v_j^T x) \cdots\right.$$

$$\le \|a\|_\infty \sum_{j=1}^m \frac{\beta}{2}\left(v_j^T x - v_{0,j}^T x\right)^2 \le \|a\|_\infty \|x\|^2 \cdot \frac{\beta}{2}\cdot \underbrace{\sum_j \|v_j - v_{0,j}\|^2}$$ //

Let's try this for ReLU.

$$\left|F(x;V) - F_0(x;V)\right| =$$

$$\left|\sum_j a_j\left(\underbrace{\sigma(v_j^T x)}_{v_j^T x\,\sigma'(v_j^T x)} - \left(\underbrace{\cancel{\sigma(v_{0,j}^T x)}}_{3} + \underbrace{\sigma'(v_{0,j}^T x)}_{2}\left(v_j^T x - \underbrace{\cancel{v_{0,j}^T x}}_{1}\right)\right)\right)\right|$$

$$= \left|\sum_j a_j\, v_j^T x \boxed{\left(\sigma'(v_j^T x) - \sigma'(v_{0,j}^T x)\right)}\right|$$

Handled via Gaussian concentration

$\le ???$

$\le ??$ $\le \|a\|_\infty \sum_j \underbrace{|v_j^T x|} \cdot \underbrace{\left|\sigma'(v_j^T x) - \sigma'(v_{0,j}^T x)\right|}_{\le 1}$

$$\le \|a\|_\infty \sum_j \|v_j\| \le \|a\|_\infty \sqrt{m}\,\|V\|_F$$

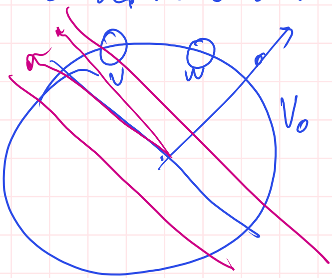
Lemma (ReLU linearization). For any $B > 0$ and any $\|a\| \le 1$, with probability $\ge 1-\delta$ (over V_0), for any $\|W - V_0\| \le B$, $\|V - V_0\| \le B$,

$$\left| F(x;V) - \left(F(x;W) + \langle \hat{\partial} F(x;W), V - W \rangle \right) \right| \le \|a\|_\infty \, m^{1/3} \left(4B^{4/3} + 2B \ln(1/\delta)^{1/4}\right).$$

Remark. (i) Maybe don't need $m^{1/3}$, perhaps if do two layers (or other change).

(ii) Vs smooth theorem: B dependence smaller (B^2), has probability.

Why proof works



Lemma. For any $\tau > 0$, $x \in \mathbb{R}^d$, $\|x\| > 0$, with pr $\ge 1-\delta$

$$\sum_j \mathbb{1}\left[|v_j^T x| \le \tau \|x\| \right] \le m\tau + \sqrt{\frac{m}{2} \ln\left(\frac{1}{\delta}\right)}.$$

Proof. rotational invariance, explicit integral of Gaussian density, & Hoeffding.