Lecture 7: Signal-to-noise in wide networks

Ann
* HW1 ?
* HW2 with easy code?

Plan for next few lectures

* $F \approx F_0$ e.g.

$F(x;V_0) + \langle \partial F(x;V_0), V - V_0 \rangle$
$= \sum_j a_j \left(\sigma(x^\top v_{0,j}) + \sigma'(x^\top v_{0,j}) \langle x, v_j - v_{0,j} \rangle \right)$

smooth act $|F(x;V) - F_0(x;V)|$
$\leq \|V - V_0\|^2 \cdot \frac{\beta}{2} \|a\|_\infty \cdot \|x\|$

ReLU $\left| F(x;W) - [F(x;V) + \langle \partial F(x;V), W - V \rangle] \right| \leq 5 m^{1/3} \left(B^{4/3} + B^{2/3} \ln(m/\delta)^{1/4} \right).$

* Signal-to-noise phenomenon

F_0 is a universal approximator.

$\Rightarrow$ implies scaling $\frac{\rho}{\sqrt{m}} F_0$ \rho (most papers use ε)

$\Rightarrow$ Then can take limits $\lim_{m \to \infty} \frac{\rho}{\sqrt{m}} F_0 \xrightarrow{a.s.} F_\infty$

* Kernels

Signal-to-noise.

[scale of noise]

* Lemma shortly: $\forall x$, $\|x\| \le 1$, w/ pr $\ge 1-\delta$, $|F(x; V_0)| \le \underbrace{16\sqrt{m}\,\ln(m/\delta)}_{\text{noise}}$.

[note whp, $\|V_0\| \approx \sqrt{m}$]

$\|V^{(m)}\| = \sqrt{m}$

* Signal property: given (a, v), $\|v\| = 1$, $a \in \{\pm 1\}$, Define $F(x; V^{(m)}) = \sum_i a\,\sigma(v^T x)$

Note $F(v; V^{(m)}) = \sum_j a\,\sigma(v^T v) = \underbrace{m \cdot a \cdot \sigma(\|v\|^2)}_{\text{signal}}$

Theorem (signal-to-noise phenomenon). [ReLU]

Let reference network $g(x) = \sum_{k=1}^{r} \alpha_k\,\sigma(\beta_k^T x)$ given with $\|\beta_k\| = 1$.

Let noise parameter $\tau > 0$ be given with $\tau < \frac{1}{2} \min_{i \ne j} \|\beta_i - \beta_j\|$.

With pr $\ge 1-\delta$ over V_0 with $m \ge 4^{d+2} \ln(1/\delta) / \tau^{2d-2}$, also $\max_j \|v_j - v_{0,j}\| \le \frac{2^{d+1}\|\alpha\|_\infty}{\tau^{d-1}\sqrt{m}}$,

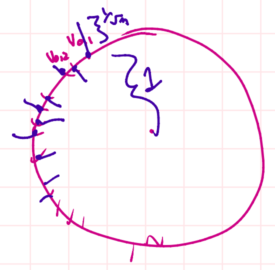
$\exists\, V \in \mathbb{R}^{m \times d}$ s.t. $\|V - V_0\|_F \le \frac{2^{d+1}\|\alpha\|_2}{\tau^{d-1}}$ (i.e, can be $\ll \sqrt{m}$, not clearing initial noise)

s.t. $\forall x$, $\|x\| \le 1$, $F(x; V) = F_0(x; V)$, and

$$\left| F(x; V) - \frac{2^{d+1}\sqrt{m}}{\tau^{d-1}} g(x) \right| \le \sqrt{m}\left(16 d \ln(m/\delta) + \|\alpha\|_1\right).$$

[aka $\left| \frac{\tau^{d-1}}{2^{d+1}\sqrt{m}} F(x; V) - g(x) \right| \le \left(\frac{\tau^{d-1}}{2^{d+1}}\right)\left(16 d \ln m/\delta + \|\alpha\|_1\right)$.]

Remark. $\|V_0\| \approx \sqrt{m}$, otoh $\|V - V_0\| = O(1)$



$\|V - V_0\| = O(1) \ll \sqrt{m} \approx \min\{\|V_0\|, \|V\|\}$

Remark: Implies universal approximation near initialization:

Given $h: \mathbb{R}^d \to \mathbb{R}$ continuous, can choose $g(x) = \sum \alpha_k\,\sigma(\beta_k^T x)$

with $\sup_{\|x\| \le 1} |h(x) - g(x)| \le \frac{\varepsilon}{2}$ [via lectures 2-4]

then use theorem to find $V \in \mathbb{R}^{m \times d}$ with $\|V - V_0\| = O(1)$

and $\left| \frac{e}{\sqrt{m}} F(x) - h(x) \right| \le \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \le \varepsilon$.

Remark. For certain functions g, it's known how to prove theorem with $\frac{2^{d+1}}{\tau^{d-1}}$ replaced by $O(d)$ or better

E.g., 2-sparse parity problem.

$\llcorner$ the example of deep networks beating kernels.

Define scaling $\frac{e_m}{\sqrt{m}} F(x; V)$

Given $(v_{0,j})_{j \ge 1}$, define $v_j := v_{0,j} + \frac{a_j}{e_m\sqrt{m}} T(v_{0,j})$,

where $T: \mathbb{R}^d \to \mathbb{R}^d$, and $(V_0^{(m)})_{m \ge 1}$ has rows $(v_{0,j}^T)_{j=1}^m$

$(V^{(m)})_{m \ge 1}$ has rows $(v_j^T)_{j=1}^m$

Theorem. Under "regularity conditions" on e_m, T, σ,

$a_j \in \{\pm 1\}$

$$\frac{e_m}{\sqrt{m}} F_0(x; V^{(m)}) \xrightarrow[\text{a.s.}]{m \to \infty} F_\infty(x; T) = \int \langle T(v), \partial_v \sigma(x; v) \rangle\, d\mathcal{N}(v)$$

$$\| \quad \frac{e}{\sqrt{m}}\left[\sum_{j=1}^{m} a_j \left(\sigma(v_{0,j}^T x) + \langle \partial_v \sigma(v_{0,j}^T x), v_j - v_{0,j} \rangle \right) \right]$$

Where "regularity conditions" mean

* $\sup_{v \in \mathbb{R}^d} |T(v)| < \infty$
* Need $\forall x$, w/ pr $\ge 1-\delta$, $|F(x; V_0)| \le c\sqrt{m} \ln(m/\delta)$.
* $\exists\, C$, s.t. $e_m \le \frac{C}{\ln(1+m)^2}$.

Proof. $\frac{e_m}{\sqrt{m}}\left[\sum_{j=1}^{m} a_j \left(\sigma(v_{0,j}^T x) + \langle \partial_v \sigma(v_{0,j}^T x), \frac{a_j}{e_m\sqrt{m}} T(v_{0,j}) \rangle \right) \right]$

$$= \boxed{\frac{e_m}{\sqrt{m}} \sum_{j=1}^{m} a_j\, \sigma(v_{0,j}^T x)} + \frac{1}{m}\sum_{j=1}^{m} \langle T(v_{0,j}), \partial_v \sigma(v_{0,j}^T x) \rangle$$

" $\to 0$ a.s." "via CLT"

$\to$ RHS a.s SLLN.

$\mathbb{E}$

Office hours

(ReLU theorem from last time.)

Want S^c s.t. $\forall j \in S$, $\mathbb{1}[v_j^T x \geq 0] = \mathbb{1}[w_j^T x \geq 0]$
$= \mathbb{1}[v_{0,j}^T x \geq 0]$

define

$$S = \left\{ j \in \{1, \dots, m\} : \begin{array}{l} |v_{0,j}^T x| \leq \tau \|x\| \\ \text{or} \\ \|v_j - v_{0,j}\| > \tau \quad \text{or} \quad \|w_j - v_{0,j}\| > \tau \end{array} \right\}$$

Consider $j \notin S \Rightarrow \left\{ \begin{array}{l} |v_{0,j}^T x| > \tau \|x\| \\ \text{and} \\ \|v_j - v_{0,j}\| \leq \tau \\ \text{and} \\ \|w_j - v_{0,j}\| \leq \tau \end{array} \right.$

Suppose $v_{0,j}^T x > \tau \|x\|$

- v: Done if can show $v_j^T x > 0$. $v_j^T x = \langle v_j - v_{0,j} + v_{0,j}, x \rangle$ ...
- w: Done if can show $w_j^T x > 0$.

otherwise $v_{0,j}^T x < -\tau \|x\|$

- v: Done if $v_j^T x < 0$.
- w: Done if $w_j^T x < 0$.