Lecture 9: kernels; architectural benefits.

Announcements
1. Hw1 due next week.
2. Zoom lectures next week.
3. OPT lec simplified.
4. Gen lec?

Recall $F_0(x;w) = F(x;w_0) + \langle \bar\partial F(x;w_0), w-w_0\rangle$.

① F_0 affine in w; in general nonlinear in x (nonlinearity can fail for specific w_0).

② Suppose have fixed $(x_1, \dots, x_n)$,

then might as well consider $w - w_0 \in \mathrm{Span}(\{\bar\partial F(x_1;w_0), \dots, \bar\partial F(x_n;w_0)\})$

└ doesn't affect $(F_0(x_1;w), \dots, F_0(x_n;w)) \in \mathbb{R}^n$.

Define
$$\bar\partial F(X;w_0) = \begin{bmatrix} -\partial f(x_1;w_0)^T- \\ | \\ -\partial f(x_n;w_0)^T- \end{bmatrix} \in \mathbb{R}^{n\times p};$$

suffices to consider w such that $w = w_0 + \bar\partial F(X;w_0)^T v$, $v \in \mathbb{R}^n$.

Then
$$F_0(x_j;w) = F(x_j;w_0) + \underbrace{\langle \partial F(x_j;w_0), w-w_0\rangle}_{\langle \partial F(x_j;w_0),\ \bar\partial F(X;w_0)^T v\rangle}$$
$$= F(x_j;w_0) + \sum_{i=1}^n v_i \underbrace{\langle \bar\partial F(x_j;w_0), \bar\partial F(x_i;w_0)\rangle}_{\text{"kernel" (fancy inner product)}\ =:\ k_m(x_j, x_i)}.$$

③ More explicitly consider least squares regression:
$$\mathcal{L}(w) = \frac12 \sum_{i=1}^n (F_0(w) - y)^2$$
using this directly since in the regime $F \approx F_0$.
$$= \frac12 \sum_{i=1}^n \left\| \bar\partial F(X;w_0)(w-w_0) - (y - F(X;w_0)) \right\|^2$$

normal equations:
$$\bar\partial F(X;w_0)^T \bar\partial F(X;w_0)(w-w_0) = \partial F(X;w_0)^T(y - F(X;w_0))$$

(minimum norm) ols solution:
$$w - w_0 = \underbrace{\left[\bar\partial F(X;w_0)^T \bar\partial F(X;w_0)\right]}_{n\times n \text{ matrix, "gram matrix"}}^{+} \bar\partial F(X;w_0)^T [y - F(X;w_0)].$$

pseudoinverse

Remarks. ① We are fitting $\boxed{y - F(x;w_0)}$ not y.

(e.g., have to work hard even if $y = 0$)

② To ensure $L(w) = 0$ for all y,

one approach is to require $\mathrm{rank}(\bar\partial F(X;w_0)) = n$,

$\Rightarrow$ gram matrix is invertible

$\Rightarrow$ $\exists$ w s.t. $F_0(X;w) = y$.

Many papers use this perspective & require or prove gram matrix is invertible.

(often implies width $\geq$ #training points.)

Recall our scaling $\frac{e}{\sqrt m} F_0(x;w_0)$ [Suppose $F(x;w) = \sum_j a_j \sigma(v_j^T x)$, $a_j \in \{\pm 1\}$]

$$\underbrace{k_\infty(x_i, x_j)}_{\text{"neural tangent kernel"}} = \lim_{m\to\infty} \left\langle \bar\partial \frac{e}{\sqrt m} F(x_i;w_0), \bar\partial \frac{e}{\sqrt m} F(x_j;w_0)\right\rangle$$

(… using shallow form …)

$$= \lim_{m\to\infty} \sum_{k=1}^m \frac{e^2}{m} \langle a_k \sigma'(v_k^T x_i) x_i, a_k \sigma'(v_k^T x_j) x_j\rangle$$
$$= \lim_{m\to\infty} \frac{e^2}{m} \sum_{k=1}^m \underbrace{a_k^2}_{1} \langle x_i, x_j\rangle \underbrace{\boxed{\sigma'(v_k^T x_i)\sigma'(v_k^T x_j)}}_{Z_k}$$
$$\xrightarrow[m\to\infty\ (a.s.)]{\text{SLLN}}\ = e^2 \langle x_i, x_j\rangle \mathop{\mathbb{E}}_{v\sim\mathcal{N}(0, I_d)} \sigma'(v_k^T x_i)\sigma'(v_k^T x_j).$$

$\frac1m \sum Z_k \to \mathbb{E} Z_1$

Proposition in the notes

(In notes: F_∞ rewritten in terms of k_∞.)

Proposition. For $\sigma(z) = \max\{0, z\}$ (ReLU), and $\|x\| = \|x'\| = 1$,
$$k_\infty(x, x') = e^2 \langle x, x'\rangle \left(\frac{\pi - \arccos(\langle x, x'\rangle)}{2\pi}\right).$$

Remark. Literature typically works with $(x^T, \sqrt{1 - \|x\|^2}) \in \mathbb{R}^{d+1}$ whereby $\|x\| \leq 1$ in $\mathbb{R}^d$ then has $\|\cdot\| = 1$.

Proof. Suffices to consider $\mathop{\mathbb{E}}_{v\sim\mathcal{N}} \sigma'(v^T x)\sigma'(v^T x')$

$\{v^T x = 0\} \cup \{v^T x' = 0\}$ are $\mathcal{N}$-null

$$= \mathop{\mathbb{E}}_{v\sim\mathcal{N}} \mathbb{1}[v^T x \geq 0] \cdot \mathbb{1}[v^T x' \geq 0]$$

algebra

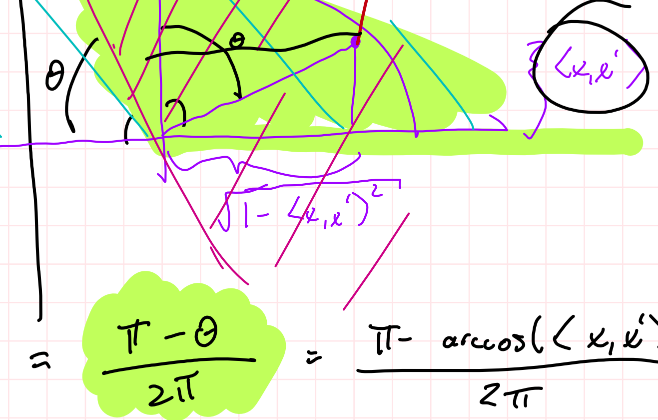
rotational invariance

$v \mapsto Mv$

$$M = \begin{bmatrix} | & | \\ x & \frac{(I - xx^T)x'}{\|(I - xx^T)x'\|} \\ | & | \end{bmatrix}$$ orthonormal

$$= \mathop{\mathbb{E}}_{v\sim\mathcal{N}} \boxed{\mathbb{1}[v_1 \geq 0]} \cdot \boxed{\mathbb{1}[v_1\langle x, x'\rangle + \sqrt{1 - \langle x, x'\rangle^2}\, v_2 \geq 0]}$$

$\|v\|$ $\quad$ $\|v\|$ $\quad$ $\|v\|$



$$= \frac{\pi - \theta}{2\pi} = \frac{\pi - \arccos(\langle x, x'\rangle)}{2\pi}$$

$$= \Pr_{(u_1, u_2)\sim \text{boundary of circle in } \mathbb{R}^2} \left[u_1 \geq 0 \text{ and } u_1\langle x, x'\rangle + u_2\sqrt{1 - \langle x, x'\rangle^2} \geq 0 \right]$$