PS4 due Tuesday
(No Peer Eval for PS3, but will do one more for PS4.)

- Topics after break: some subset of
  { neural net learning,
    single-index models,
    leaderboards,
    chain-of-thought learning,
    active learning }

Last time:

- Linear learning paradigm
- Convex surrogate losses
- SGD

Setup for SGD from last time:

$\phi : \mathcal{X} \to \mathbb{R}^d$

$\ell : \mathbb{R} \to \mathbb{R}$ loss function.

today: $\ell = \ell_{hinge}$

$\ell_{0/1}(s) = \mathbb{1}\{s \le 0\}$

$\ell_{hinge}(s) = \max\{0, 1-s\}$




$J(w) = E[\ell(Y\langle\phi(X), w\rangle)]$ for $(X,Y) \sim P$

$= R_\ell(w)$

$$\nabla L(w;(x,y)) = \begin{cases} 0 & \text{if } y\langle\phi(x),w\rangle \ge 1 \\ -y\phi(x) & \text{o.w.} \end{cases}$$

SGD for (approx) minimizing $J$

$((X_1,Y_1), \dots, (X_T,Y_T)) \sim P^T$

$w_1 = 0$

For $t = 1, 2, \dots, T$:

$w_{t+1} = w_t - \eta \nabla_w L(w_t; (X_t, Y_t))$ $\qquad L(w;(x,y)) = \ell(y\langle\phi(x),w\rangle)$

$$= \begin{cases} w_t & \text{if } Y_t\langle\phi(X_t), w_t\rangle \ge 1 \\ w_t + \eta Y_t \phi(X_t) & \text{o.w.} \end{cases}$$

Return $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$

**Thm** Assume $\|\phi(x)\| \le G \ \forall x \in \mathcal{X}$

and we use $\eta = W/G\sqrt{T}$

then

$$E[R_\ell(\bar{w})] \le \min_{\substack{w \in \mathbb{R}^d: \\ \|w\| \le W}} R_\ell(w) + \frac{WG}{\sqrt{T}}.$$

$\frac{WG}{\sqrt{T}} = \epsilon \iff T = \left(\frac{WG}{\epsilon}\right)^2$ ← sample complexity to get $\ell$-risk at most $\min_{\substack{w \in \mathbb{R}^d: \\ \|w\| \le W}} R_\ell(w) + \epsilon$.

Time complexity: $\left(\frac{WG}{\epsilon}\right)^2 \times [\text{time to compute } \phi + \text{time to update weight}]$

$O(d)$ typically

For any $g : \mathcal{X} \to \mathbb{R}$, and any $w \in \mathbb{R}^d$

$$R_\ell(w) = R_\ell(\langle\phi, w\rangle) \le R_\ell(g) + E[|\langle\phi(x), w\rangle - g(x)|]$$

Strategy

① Identify nice $g : \mathcal{X} \to \mathbb{R}$ that has small $R_\ell(g)$

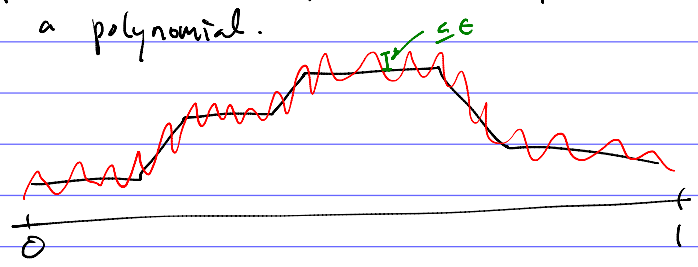
② Design $\phi : \mathcal{X} \to \mathbb{R}^d$ s.t. $g$ is well-approximated by $\langle\phi(x), w\rangle$ for some ($g$-specific) choice of $w$.

$X = \mathbb{R}^n$, $g$ is $\lambda$-Lipschitz function

$$|g(x) - g(x')| \le \lambda \cdot \|x - x'\| \quad \text{for all } x, x' \in \mathbb{R}^n.$$

$n=1$ Weierstrass Approximation Theorem:

"Any continuous function can be arbitrarily well approximated uniformly over a compact domain (like $[0,1]$) by a polynomial."




$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m$$

$$\phi(x) = (1, x, x^2, \ldots, x^m)$$
$$w = (a_0, a_1, a_2, \ldots, a_m)$$
$$\langle \phi(x), w \rangle = p(x)$$

---

Bernstein polynomials $\{b_{m,k}\}_{k=0}^{m}$ of degree $m$

$$b_{m,k}(x) = \binom{m}{k} x^k (1-x)^{m-k}$$

**Thm** For any $\lambda$-Lipschitz fn $g$

define

$$B_m g(x) = \sum_{k=0}^{m} g\left(\frac{k}{m}\right) \cdot b_{m,k}(x)$$

Then $\forall x \in [0,1]$, $\quad |B_m g(x) - g(x)| \le \frac{\lambda}{2\sqrt{m}}$

**Proof:** Fix any $x \in [0,1]$, $S \sim \text{Binomial}(m, x)$

$$E\left[g\left(\tfrac{S}{m}\right)\right] = \sum_{k=0}^{m} \Pr(S=k)\, g\left(\tfrac{k}{m}\right)$$
$$= \sum_{k=0}^{m} \binom{m}{k} x^k (1-x)^{m-k} g\left(\tfrac{k}{m}\right) = B_m g(x)$$

$$|B_m g(x) - g(x)| = \left|E\left[g\left(\tfrac{S}{m}\right) - g(x)\right]\right|$$
$$\le E\left|g\left(\tfrac{S}{m}\right) - g(x)\right| \quad \text{by Jensen's inequality}$$
$$\le \lambda \cdot E\left|\tfrac{S}{m} - x\right|$$
$$\le \lambda \sqrt{\text{Var}(S/m)}$$
$$= \lambda \sqrt{\frac{x(1-x)}{m}}$$
$$\le \frac{\lambda}{2\sqrt{m}} \quad \square$$

---

For $n \ge 1$

to get $\epsilon$-approximation, need degree $m = \frac{\lambda^2}{4\epsilon^2} \cdot n$ per variable

Dim. of $\phi$ will be like $n^{\frac{\lambda^2 n}{4\epsilon^2}} = n^{m}$

— exponential in original data dimension $n$.

"Curse of dimension"

---

Our hope of using linear learning to learn Lipschitz functions in general is dashed!

We'll make more assumptions about the learning problem to endow $g$ w/ more structure

↳ low-dim structure

**Assumption** $X = B^n := \{x \in \mathbb{R}^n \mid \|x\| \le 1\}$ (unit ball)

Assume $g(x) = \sigma(\langle x, v \rangle)$ for some $v \in B^n$

and $\sigma : \mathbb{R} \to \mathbb{R}$, $\lambda$-Lipschitz

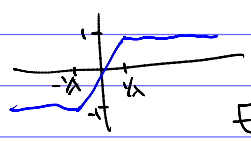
has

$$E[\ell_{\text{hinge}}(y\, g(x))] \le \epsilon.$$

---

Almost like saying $\text{LTF}_{\mathbb{R}^n}$-realizable

- except:
  - $\epsilon > 0$
  - $\sigma$ is not $\text{sign}(\cdot)$

If $\sigma =$  then

$$E[\ell_{\text{hinge}}(y\, g(x))] \le 2 \cdot \text{err}_{\gamma}(\text{sign} \circ g)$$

**Thm** $\exists\ \phi : B^n \to \mathbb{R}^d$, $d = n^{O((\lambda/\epsilon)^2)}$, $\|\phi(x)\| = O(\frac{\lambda}{\epsilon})$

s.t. for all $g(x) = \sigma(\langle x, v\rangle)$ for some $\lambda$-Lipschitz $\sigma$ and $v \in B^n$

there is $w \in \mathbb{R}^d$, $\|w\| = 2^{O((\lambda/\epsilon)^2)}$

s.t. $|g(x) - \langle \phi(x), w\rangle| \le \epsilon \quad \forall\ x \in B^n$.

$\phi(x)$ has components being monomials in $x$ of deg. $m = O((\frac{\lambda}{\epsilon})^2)$.

---

**Main idea:**

① $\sigma$ is $\lambda$-Lipschitz, so $\exists$ polynomial $p$ of deg $m = \lceil \frac{\lambda}{\epsilon} \rceil^2$ s.t. $|p(z) - \sigma(z)| \le \epsilon \quad \forall\ z \in [-1, 1]$.

$$p(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_m z^m$$
$$\text{s.t.} \sum_{i=0}^{m} |a_i| \le 2^m$$

② Then we'll use $\phi(x) = (1, x, x^{\otimes 2}, \ldots, x^{\otimes m}) \in \mathbb{R}^d$

$$d = \sum_{k=0}^{m} n^k$$

and show $\exists\ w = w^{\sigma, v}$ (specific to $\sigma$ & $v$)

s.t.

$$\langle \phi(x), w\rangle = p(\langle x, v\rangle) \approx \sigma(\langle x, v\rangle)$$

---

For $x \in \mathbb{R}^n$

$x^{\otimes k} \in \mathbb{R}^{n^k}$ whose entries are $\prod_{t=1}^{k} x_{i_t}$ for $(i_1, \ldots, i_k) \in [n]^k$

$$x^{\otimes 2} = xx^T = \begin{bmatrix} x_1^2 & x_1x_2 & \cdots & x_1x_n \\ x_2x_1 & x_2^2 & & \\ \vdots & & \ddots & \\ x_nx_1 & & & x_n^2 \end{bmatrix}$$

$\langle x^{\otimes k}, y^{\otimes k}\rangle = \langle x, y\rangle^k$

$\phi(x) = (1, x, x^{\otimes 2}, x^{\otimes 3}, \ldots, x^{\otimes m})$

$p(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_m z^m$

$w^{p,v} = (a_0, a_1 v, a_2 v^{\otimes 2}, \ldots, a_m v^{\otimes m})$

$$\langle \phi(x), w^{p,v}\rangle = \sum_{k=0}^{m} \langle x^{\otimes k}, a_k v^{\otimes k}\rangle = \sum_{k=0}^{m} a_k \langle x^{\otimes k}, v^{\otimes k}\rangle$$
$$= \sum_{k=0}^{m} a_k \langle x, v\rangle^k$$
$$= p(\langle x, v\rangle).$$