Benign overfitting

Last lecture we showed that uniform convergence allows for arguments like, wp $> 1-\delta$, $\forall f \in \mathcal{F}$, $|\hat{L}(f) - L(f)| \leq \sqrt{\frac{\text{Complexity}(\mathcal{F})}{n}}$

i.e. we can guarantee small population error if empirical error is small & # samples is sufficiently large.

However, we also saw in Zhang et al's paper that <u>interpolators</u> ($\hat{L}(f)=0$) can achieve good performance even on noisy problems ($L(f) \geq c > 0$). In particular, in many modern deep learning settings, we have

$$c \leq L(f) = |L(f) - \hat{L}(f)| \leq 2 \cdot \min_{f \in \mathcal{F}} L(f).$$

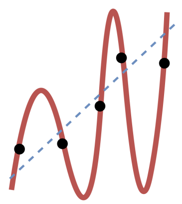
Clearly, such settings cannot have a simple uniform convergence argument. We'll now show a result which <u>provably</u> allows for

$$\min_{f \in \mathcal{F}} L(f) = c \leq L(f_n) = |L(f_n) - \hat{L}(f_n)| \leq \min_{f \in \mathcal{F}} L(f) + o_n(1).$$

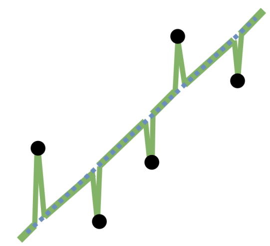
This is called "benign overfitting":

- "overfitting" since $c \leq L(f)$ & $\hat{L}(f) = 0 << c$.
- "benign" since as $n \to \infty$, $L(f_n) \to \min_{f \in \mathcal{F}} L(f)$ while $\hat{L}(f_n) = 0$ $\forall n$.

"Classical" (catastrophic) form of overfitting:
think of using high degree polynomial to fit noisy linear model:


Catastrophic overfitting

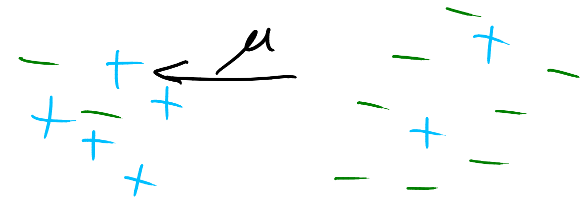
& compare w/ another "overfitting" estimator:


Benign overfitting

The setting: binary Gaussian mixture model. $\mu \in \mathbb{R}^d$. $p \in (0, \frac{1}{2})$

- $\tilde{y} \sim \mathrm{Unif}(\{\pm 1\})$, $\quad x \mid \tilde{y} \sim \tilde{y}\mu + z$, $\quad z \sim N(0, I_d)$;

$$y = \begin{cases} \tilde{y}, & \text{wp } 1-p, \\ -\tilde{y}, & \text{wp } p. \end{cases} \qquad (x, y) \sim \mathbb{P}.$$



- Suppose $(x_i, y_i) \overset{iid}{\sim} \mathbb{P}$.

Call $i \in \mathcal{N}$ if $y_i = -\tilde{y}_i$; $\quad i \in \mathcal{C}$ if $y_i = \tilde{y}_i$. (Learner does <u>not</u> know if $i \in \mathcal{C}$ or $i \in \mathcal{N}$).

Then $\mathcal{N} \cup \mathcal{C} = [n]$.

We will show that $n\mathrm{AVG} := \sum_{i=1}^{n} y_i x_i$ exhibits benign overfitting, under certain conditions.

Two things to show:

(1) $\hat{L}(n\mathrm{AVG}) = 0$: for each $k \in [n]$, $y_k \langle n\mathrm{AVG}, x_k \rangle > 0$

(2) $p \leq L(n\mathrm{AVG}) \leq p + o_n(1)$.

<u>Lemma</u> If training data is p-orthogonal in the sense that for some $p \geq 2$ and for $R^2 = \max_{i \neq j} \frac{\|x_i\|^2}{\|x_j\|^2} < \infty$, we have

$$\|x_k\|^2 \geq pR^2 n \cdot \max_{i \neq j} |\langle x_i, x_j \rangle| \quad \text{for } k = 1, \dots, n, \text{ then}$$

for all $k \in [n]$, $y_k \langle x_k, AVG \rangle > 0$.

<u>Pf.</u>
$$\langle y_k x_k, nAVG \rangle = \langle y_k x_k, \sum_{i=1}^{n} y_i x_i \rangle$$
$$= \|x_k\|^2 + \sum_{i \neq k} \langle y_k x_k, y_i x_i \rangle$$
$$\geq \|x_k\|^2 - n \cdot \max_{i \neq j} |\langle x_i, x_k \rangle|$$
$$\geq \tfrac{1}{2} \|x_k\|^2 > 0, \quad \text{since } R^2 \geq 1 \text{ \& } p \geq 2. \quad \square$$

Thus, AVG interpolates the training data ($\hat{L}(f) = 0$) <u>if</u> the training data is p-orthogonal. We now establish sufficient conditions for this.

Note: $\|x_k\|^2 = \|\tilde{y}\mu + z_k\|^2 = \|\mu\|^2 + \|z_k\|^2 + 2\langle \tilde{y}\mu, z_k \rangle$, so suffices to control both $\|z_k\|^2$, $\langle z_k, u \rangle$ for fixed vector u.

Lemma There is a $C_0 > 1$ s.t. for $\delta \in (0, \frac{1}{2})$, if $d > C_0^3 \log(12n/\delta)$, then wp $> 1-\delta$, we have:

① $\forall k, \left| \frac{\|z_k\|}{\sqrt{d}} - 1 \right| \le C_0 \sqrt{\frac{\log(12n/\delta)}{d}} \iff \left| \|z_k\| - \sqrt{d} \right| \le C_0 \sqrt{\log(\frac{12n}{\delta})}$

② $\forall i \neq j, \ |\langle z_i, z_j \rangle| \le C_0 \sqrt{d} \log(12n^2/\delta)$

Pf Part ① & ② were HW.

Lemma There is $C_1 > 1$ s.t. for $\delta \in (0, \frac{1}{2})$, if $d > C_0^3 \log(\frac{12n}{\delta})$, and if $d \ge C_0 (n \|\mu\|^2 \vee n^2 \log \frac{12n}{\delta})$, $\|\mu\| \ge 3C_0$, then for some absolute $C_1 > 1$, wp $> 1-\delta$,

$$\|x_k\|^2 \ge C_1 n \cdot \max_{i,j} \frac{\|x_i\|^2}{\|x_j\|^2} \cdot \max_{i \neq j} |\langle x_i, x_j \rangle|.$$

Pf $|\langle x_i, x_j \rangle| = |\langle \tilde{y}_i \mu + z_i, \tilde{y}_j \mu + z_j \rangle|$

$$\le \|\mu\|^2 + |\langle \mu, z_i \rangle| + |\langle \mu, z_j \rangle| + |\langle z_i, z_j \rangle|$$

$$\le \|\mu\|^2 + 2C_0 \|\mu\| \log(12n/\delta) + C_0 \sqrt{d} \log(12n^2/\delta).$$

$$\le \|\mu\|^2 + 2C_0 (\|\mu\| \vee \sqrt{d}) \log(12n^2/\delta).$$

$\|\mu\|^2 \le \frac{d}{C_0 n}$. $\|\mu\| \le \sqrt{\frac{d}{C_0 n}}$.

$$\|x_k\|^2 = \|\tilde{y}_k \mu + z_k\|^2 = \|\mu\|^2 + 2\tilde{y}_k \langle \mu, z_k \rangle + \|z_k\|^2.$$

By prev lemma, $-C_0 \sqrt{\log(12n/\delta)} \le \|z_k\| - \sqrt{d} \le C_0 \sqrt{\log \frac{12n}{\delta}}$. For $d > C_0^3 \log \frac{12n}{\delta}$,

$\sqrt{d} = \frac{\sqrt{d}}{2} + \frac{\sqrt{d}}{2} \ge \frac{C_0^{3/2}}{2} \sqrt{\log \frac{12n}{\delta}} + \frac{\sqrt{d}}{2} \implies \|z_k\| \ge \frac{\sqrt{d}}{2} + \left(\frac{C_0^{3/2}}{2} - C_0 \right) \sqrt{\log \frac{12n}{\delta}} \ge \frac{\sqrt{d}}{2}$ for C_0 large enough.

$$\Rightarrow \|x_k\|^2 \le \|\mu\|^2 + 2C_0\|\mu\| \log(12n/\delta) + d\left(1 + C_0^2 \cdot \log 12n/\delta\right)$$
$$\|x_k\|^2 \ge \|\mu\|^2 - 2C_0\|\mu\| \log(12n/\delta) + d/4.$$

For $d \ge C_0 n \|\mu\|^2$, $\|\mu\| \ge 3C_0$, & $d > C_0^3 \log(12n/\delta)$, we get:

$$\|x_k\|^2 \le d\left(\frac{1}{C_0} + \frac{d}{C_0}\frac{\log(12n/\delta)}{n} + 1 + \frac{1}{C_0}\right),$$
$$\|x_k\|^2 \ge \frac{d}{4}\left(1 - \frac{8C_0\|\mu\|\log(12n/\delta)}{d}\right) \ge \frac{d}{5} \text{ for } C_0 \text{ large.}$$

$$\Bigg\} \; R^2 \le \frac{(1+\frac{1}{C_0})}{(1-\frac{1}{C_0})} \le (1+2C_0)^2 \le 1.01 \text{ for } C_0 \text{ large.}$$

$$\Rightarrow \frac{\|x_k\|^2}{R^2 \max_{i\ne j} |\langle x_i, x_j\rangle|} \ge \frac{d/5}{1.01 \cdot \left(\|\mu\|^2 + 2C_0(\|\mu\| \vee \sqrt{d}) \log(12n^2/\delta)\right)} = \frac{d/5.05}{\|\mu\|^2 + 2C_0\sqrt{d}\log(\frac{12n^2}{\delta})}.$$

In order for near-orthogonality, this must be $\Omega(n)$:

$$\frac{d}{\|\mu\|^2 + \sqrt{d}\log(12n^2/\delta)} = \Omega(n) \text{ holds if } \frac{d}{\|\mu\|^2} = \Omega(n) \;\; \& \;\; \frac{\sqrt{d}}{\log(12n^2/\delta)} = \Omega(n)$$

These are precisely the assumptions in the lemma. $\boxtimes$

Now we move to generalization. We want to show $p \le \mathbb{P}(y \ne \mathrm{sgn}(\langle AVG, x\rangle)) \le p + o(1)$.

<u>Lemma</u> Suppose $w \in \mathbb{R}^d$. Then
$$\mathbb{P}(y \ne \mathrm{sgn}(\langle w, x\rangle)) \le p + \Phi(-\langle w/\|w\|, \mu\rangle),$$ where Φ is normal CDF.

<u>Remark</u>: If $\langle w, \mu\rangle \le 0$, vacuous; otherwise, decays exp-fast in $\langle \frac{w}{\|w\|}, \mu\rangle$.

Pf $\mathbb{P}(y \neq \mathrm{sgn}(\langle w, x\rangle)) = \mathbb{P}(y\langle w, x\rangle < 0)$

$= \mathbb{P}(y\langle w, x\rangle < 0,\ y = \tilde{y}) + \mathbb{P}(y\langle w, x\rangle < 0,\ y = -\tilde{y})$

$\leq p + \mathbb{P}(y\langle w, x\rangle < 0,\ y = \tilde{y})$.

$\mathbb{P}(y\langle w, x\rangle < 0,\ y = \tilde{y}) = \mathbb{P}(\langle w, \tilde{y}x\rangle < 0) = \mathbb{P}(\langle w, \mu + \tilde{y}z\rangle < 0)$

$= \mathbb{P}(\langle w, \tilde{y}z\rangle < -\langle w, \mu\rangle)$

$= \mathbb{P}\left(N(0,1) < -\left\langle \frac{w}{\|w\|}, \mu\right\rangle\right)$

$= \Phi(-\langle w/\|w\|, \mu\rangle)$. $\boxed{\Xi}$

So it sufflices to show $\langle n\mathrm{AVG}, \mu\rangle$ is large

Recall: $i \in \mathcal{N}$ means $y_i = -\tilde{y}_i$, so $(x_i, y_i) = (\tilde{y}_i\mu + z_i, -\tilde{y}_i)$; $i \in C$: $(x_i, y_i) = (\tilde{y}_i\mu + z_i, \tilde{y}_i)$.

$\langle n\mathrm{AVG}, \mu\rangle = \langle \sum_{i=1}^{n} y_i x_i, \mu\rangle$

$= \langle \sum_{i\in C} \tilde{y}_i(\tilde{y}_i\mu + z_i), \mu\rangle + \langle \sum_{i\in\mathcal{N}} -\tilde{y}_i(\tilde{y}_i\mu + z_i), \mu\rangle$

$= (|C| - |\mathcal{N}|)\cdot\|\mu\|^2 + \sum_{i\in C}\langle \tilde{y}_i z_i, \mu\rangle - \sum_{i\in\mathcal{N}}\langle \tilde{y}_i z_i, \mu\rangle$.

$= (n - 2|\mathcal{N}|)\cdot\|\mu\|^2 + \langle \sum_{i=1}^{n} y'_i z_i, \mu\rangle$, where

$y_i' = \begin{cases} \tilde{y}_i : i \in C, \\ 1 - \tilde{y}_i : i \in \mathcal{N}. \end{cases}$ Note that $y_i' \overset{iid}{\sim} \text{Unif}(\{\pm 1\})$, since $\{i \in C\}$ is indep. of $\tilde{y}_i$.

Thus it suffices to prove two things:

① an upper bound on $|\mathcal{N}|$;

② an upper bound on $|\langle \sum_{i=1}^n y_i' z_i, \mu \rangle|$.

In homework, you will need to derive bounds on each of these.

But intuitively, $|\mathcal{N}| \approx pn \pm O(\sqrt{n})$;

$\sum_1^n y_i' z_i \sim N(0, n I_d)$ by independence, so

$\langle \sum_1^n y_i' z_i, \mu \rangle \sim N(0, n\|\mu\|^2)$, so $|\langle \sum_1^n y_i' z_i, \mu \rangle| \lesssim \sqrt{n}\|\mu\|$;

thus

$$\langle n\text{AVG}, \mu \rangle \gtrsim (1-2p) n \|\mu\|^2 - \sqrt{n}\|\mu\|$$

$$\gtrsim n\|\mu\|^2 \text{ if } 1-2p < \text{const}, \|\mu\| \geq \text{const}.$$

Then need to bound $\|n\text{AVG}\|^2 = \|\sum_1^n y_i x_i\|^2$;

Key here is to use near-orthogonality.

If $\frac{n\|\mu\|^4}{d} = \omega_d(1)$ then should get $\mathbb{P}(y \neq \text{sgn}(\langle \text{AVG}, x \rangle)) \leq p + o_d(1)$.