Lec 7: Minimax lower bounds: Le Cam, Assouad, and Fano

Yanjun Han

Recall: P_θ : statistical model with unknown parameter $\theta \in \Theta$

$X \sim P_\theta$: observation

$\hat{\theta} = \hat{\theta}(X)$: estimator

$L(\theta, a) \geq 0$: loss function

Target of upcoming lectures: characterize the minimax risk

$$r^* = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta [L(\theta, \hat{\theta}(X))]$$

Upper bound: construct an estimator $\hat{\theta}$ and analyze its worst-case risk

Lower bound (★): use information-theoretic arguments to show a fundamental limit for any estimator

Last lecture (LAM): asymptotic, exact constant

This lecture: non-asymptotic, focus on the optimal rate

High-level idea: $r^* \geq r^*_\pi$ for any prior π

- constructing the least favorable prior and analyzing the Bayes risk can both be hard
- try simpler priors:
  1) binary hypothesis testing : $\pi = \text{Unif}(\{\theta_0, \theta_1\})$ (Le Cam's two point method)
  2) testing multiple hypotheses: $\pi = \text{Unif}(\{\theta_1, \cdots, \theta_m\})$ (Assouad, Fano)

Le Cam's two point method

Thm. Suppose $\theta_0, \theta_1 \in \Theta$ satisfy the separation condition:

$$\inf_a (L(\theta_0, a) + L(\theta_1, a)) \geq \Delta .$$

Then

$$r^* \geq \inf_{\hat{\theta}} \frac{1}{2} (\mathbb{E}_{\theta_0} L(\theta_0, \hat{\theta}) + \mathbb{E}_{\theta_1} L(\theta_1, \hat{\theta})) \geq \frac{\Delta}{2}(1 - TV(P_{\theta_0}, P_{\theta_1})).$$

General paradigm of applying two point method: find two points $\theta_0, \theta_1 \in \Theta$ satisfying

- separation condition: $\inf_a (L(\theta_0, a) + L(\theta_1, a)) \geq \Delta$

  (no single action performs uniformly well under θ_0, θ_1)

- indistinguishability condition: $TV(P_{\theta_0}, P_{\theta_1}) \leq 1 - \Omega(1)$

  (no test can reliably distinguish between P_{θ_0} and P_{θ_1})

By the joint range of TV vs. H^2, KL, χ^2, the second condition is implied by:

① $H^2(P_{\theta_0}, P_{\theta_1}) \leq 2 - \Omega(1)$

② $D_{KL}(P_{\theta_0} \| P_{\theta_1}) = O(1)$ or $D_{KL}(P_{\theta_1} \| P_{\theta_0}) = O(1)$

③ $\chi^2(P_{\theta_0} \| P_{\theta_1}) = O(1)$ or $\chi^2(P_{\theta_1} \| P_{\theta_0}) = O(1)$

④ $I(\theta; X) \leq \log 2 - \Omega(1)$, with $\theta \sim \mathrm{Unif}(\{\theta_0, \theta_1\})$ (exercise!)

Pf. For any estimator $\hat{\theta} = \hat{\theta}(X)$,

$$\begin{aligned}
\mathbb{E}_{\theta_0} L(\theta_0, \hat{\theta}) + \mathbb{E}_{\theta_1} L(\theta_1, \hat{\theta}) &= \int L(\theta_0, \hat{\theta}(x)) P_{\theta_0}(x) dx + \int L(\theta_1, \hat{\theta}(x)) P_{\theta_1}(x) dx \\
&\geq \Delta \int \min\{P_{\theta_0}(x), P_{\theta_1}(x)\} dx \\
&= \Delta \int \tfrac{1}{2}(P_{\theta_0}(x) + P_{\theta_1}(x) - |P_{\theta_0}(x) - P_{\theta_1}(x)|) dx \\
&= \Delta(1 - TV(P_{\theta_0}, P_{\theta_1}))
\end{aligned}$$

Despite the simplicity of the two-point method, it has numerous applications.

Example 1.1 (normal mean estimation) $X \sim N(\theta, \sigma^2)$, unknown $\theta \in \mathbb{R}$, known σ^2.

Target: $r^* = \inf_{\hat{\theta}} \sup_{\theta} \mathbb{E}_\theta [(\hat{\theta} - \theta)^2]$.

Clearly, by choosing $\hat{\theta}(X) = X$, we achieve $r^* \leq \sigma^2$.

To lower bound r^*, apply two point method with $\theta_0 = 0$, $\theta_1 = \delta$:

- separation condition: $\Delta = \frac{\delta^2}{2}$
- indistinguishability condition: $TV(N(0, \sigma^2), N(\delta, \sigma^2)) = 2(1 - \Phi(\frac{|\delta|}{2\sigma}))$

Therefore, $r^* \geq \sup_{\delta \in \mathbb{R}} \frac{\delta^2}{2}(1 - \Phi(\frac{|\delta|}{2\sigma})) \approx 0.332\sigma^2$.

(Compare with $r^* = \sigma^2$ by Anderson's lemma in Lec 6)

Example 1.2 (Binomial model) Let $X \sim B(n, \theta)$ with unknown $\theta \in [0, 1]$.

Target: $r^* = \inf_{\hat\theta} \sup_{\theta} \mathbb{E}_\theta[(\hat\theta - \theta)^2]$

Upper bound: choose $\hat\theta(X) = \frac{X}{n}$, then

$$\mathbb{E}_\theta\left[\left(\frac{X}{n} - \theta\right)^2\right] = \frac{\theta(1-\theta)}{n} \le \frac{1}{4n} = O\left(\frac{1}{n}\right).$$

Lower bound: apply two point method to $\theta_0 = \frac{1}{2}$, $\theta_1 = \frac{1}{2} + \frac{1}{2\sqrt{n}}$

- separation: $\Delta = \frac{1}{2}\left(\frac{1}{2\sqrt{n}}\right)^2 = \Omega\left(\frac{1}{n}\right)$
- distinguishability: $D_{KL}(B(n,\theta_1) \| B(n, \theta_0)) = n \cdot D_{KL}(\mathrm{Bern}(\theta_1) \| \mathrm{Bern}(\theta_0))$

$$= \frac{n}{2}\left(\left(1+\frac{1}{\sqrt{n}}\right)\log\left(1+\frac{1}{\sqrt{n}}\right) + \left(1-\frac{1}{\sqrt{n}}\right)\log\left(1-\frac{1}{\sqrt{n}}\right)\right)$$

$$= \frac{n}{2} \cdot O\left(\left(\frac{1}{\sqrt{n}}\right)^2\right) = O(1).$$

This implies $r^* = \Omega\left(\frac{1}{n}\right)$ as well.

Example 1.3 (functional estimation) $X = (X_1, \cdots, X_n)$ are i.i.d. draws from an unknown pmf $P = (p_1, \cdots, p_k)$. Loss: $L(P, a) = |a - H(P)|$ (entropy estimation)

Result (Jiao et al. '15; Wu and Yang '16):

$$r^* = \inf_{\hat\theta} \sup_{P} \mathbb{E}_P |\hat\theta - H(P)| \asymp \frac{k}{n \log n} + \frac{\log k}{\sqrt{n}} \quad \text{if } k \lesssim n \log n$$

Pf of the simpler $\Omega\left(\frac{\log k}{\sqrt{n}}\right)$ lower bound: since $D_{KL}(P_{X|P_0} \| P_{X|P_1}) = n D_{KL}(P_0 \| P_1)$ we solve the optimization program motivated by the two point method:

$$\max \quad |H(P_0) - H(P_1)|$$

$$\text{s.t.} \quad D_{KL}(P_0 \| P_1) \le \frac{c}{n}$$

Try $P_0 = \left(\frac{1}{2}, \frac{1}{2(k-1)}, \cdots, \frac{1}{2(k-1)}\right)$, $P_1 = \left(\frac{1-\varepsilon}{2}, \frac{1+\varepsilon}{2(k-1)}, \cdots, \frac{1+\varepsilon}{2(k-1)}\right)$

Then $D_{KL}(P_0 \| P_1) = \frac{1}{2}\log\frac{1}{1-\varepsilon^2} = O\left(\frac{1}{n}\right)$ if $\varepsilon = O\left(\frac{1}{\sqrt{n}}\right)$,

$$|H(P_0) - H(P_1)| = \left|\frac{1}{2}\log 2 + \frac{1}{2}\log 2(k-1) - \frac{1-\varepsilon}{2}\log\frac{2}{1-\varepsilon} - \frac{1+\varepsilon}{2}\log\frac{2(k-1)}{1+\varepsilon}\right|$$

$$\asymp \varepsilon \log k.$$

Therefore, choosing $\varepsilon \asymp \frac{1}{\sqrt{n}}$ proves that $r^* = \Omega\left(\frac{\log k}{\sqrt{n}}\right)$. $\square$

(The other lower bound $\Omega\left(\frac{k}{n \log n}\right)$ is proven using a more involved two point method, which will be the topic of Lec 8)

Example 1.4 (two-armed bandit) $\theta = (\mu_1, \mu_2) \in [0,1]^2$

Observation: for $t \in [T]$, learner pulls an arm $\pi_t \in \{1,2\}$ based on (π^{t-1}, r^{t-1}),
and observes reward $r_t \sim N(\mu_{\pi_t}, 1)$.

Learner aims to minimize the regret $R_T(\pi) = T \max\{\mu_1, \mu_2\} - \sum_{t=1}^{T} \mu_{\pi_t}$.

We will show that $r^* = \inf_{\pi} \sup_{\substack{\mu_1, \mu_2, \\ |\mu_1-\mu_2| \geq \Delta}} \mathbb{E}_{\mu_1,\mu_2}[R_T(\pi)] = \Omega\left(\frac{1 \vee \log(T\Delta^2)}{\Delta} \wedge T\Delta\right)$

(In particular, choosing $\Delta \asymp \frac{1}{\sqrt{T}}$ gives the usual lower bound $\Omega(\sqrt{T})$ for two-armed bandit)

Pf. First, by the chain rule of KL, we have (exercise)

$$D_{KL}(P_{\mu_1,\mu_2} \| P_{\mu_1',\mu_2'}) = \sum_{t=1}^{T} \mathbb{E}_{P_{\mu_1,\mu_2}}\left[\frac{(\mu_1-\mu_1')^2}{2} \mathbb{1}(\pi_t = 1) + \frac{(\mu_2-\mu_2')^2}{2}\mathbb{1}(\pi_t = 2)\right]$$
$$= \frac{(\mu_1-\mu_1')^2}{2} \mathbb{E}_{P_{\mu_1,\mu_2}}[T_1] + \frac{(\mu_2-\mu_2')^2}{2}\mathbb{E}_{P_{\mu_1,\mu_2}}[T_2],$$

where $T_i = \sum_{t=1}^{T} \mathbb{1}(\pi_t = i)$ is the total number of times we pull arm $i \in \{1,2\}$.

Motivated by this, choose two points $(\mu_1, \mu_2) = (\Delta, 0)$, $(\mu_1', \mu_2') = (\Delta, 2\Delta)$.

The separation parameter is $T\Delta$ (why?), and

$$D_{KL}(P_{\mu_1,\mu_2} \| P_{\mu_1',\mu_2'}) = 2\Delta^2 \mathbb{E}_1[T_2] \quad (\mathbb{E}_1 := \mathbb{E}_{P_{\mu_1,\mu_2}}).$$

Two point method then gives $\quad r^* = \Omega(T\Delta \cdot \exp(-2\Delta^2 \mathbb{E}_1[T_2]))$

$(1 - TV(P,Q) \geq \frac{1}{2}\exp(-D_{KL}(P\|Q)))$

Note that $\mathbb{E}_1[T_2]$ depends on the policy π, and this lower bound is useful only if $\mathbb{E}_1[T_2]$ is small. To address it, note that we have a different lower bound:

$$r^* \geq \mathbb{E}_1[R_T(\pi)] = \Delta \cdot \mathbb{E}_1[T_2].$$

thus
$$r^* = \Omega(\max\{\Delta \cdot \mathbb{E}_1[T_2], T\Delta \cdot \exp(-2\Delta^2 \mathbb{E}_1[T_2])\})$$
$$= \Omega\left(\min_{T_0 \in [0,T]} \max\{\Delta T_0, T\Delta \cdot \exp(-2\Delta^2 T_0)\}\right)$$
$$= \Omega\left(\frac{1 \vee \log(T\Delta^2)}{\Delta} \wedge T\Delta\right) \qquad \blacksquare$$

Example 1.5 (multi-armed bandit) Same observation model, but now with K arms:

$$\theta = (\mu_1, \cdots, \mu_K) \in [0,1]^K, \quad R_T(\pi) = T \max_{i \in [K]} \mu_i - \sum_{t=1}^{T} \mu_{\pi_t}.$$

We will show that $r^* = \inf_{\pi} \sup_{\theta} \mathbb{E}_\theta [R_T(\pi)] = \Omega(\sqrt{KT})$.

(Interestingly, two points suffice for this example!)

Pf. Choose $\theta_1 = (\Delta, 0, 0, \cdots, 0)$

$\theta_{2,i} = (\Delta, 0, \cdots, 0, 2\Delta, 0, \cdots, 0)$, $i = 2, \cdots, K$.

For each pair $(\theta_1, \theta_{2,i})$, the separation parameter is always $T\Delta$.

In addition, for any policy π, $D_{KL}(P_{\theta_1} \| P_{\theta_{2,i}}) = 2\Delta^2 \mathbb{E}_1[T_i]$.

(recall $T_i = \sum_{t=1}^{T} 1(\pi_t = i)$)

Key observation: since $\sum_{i=2}^{K} \mathbb{E}_1[T_i] \le T$, there must $\exists\, i_0$ s.t. $\mathbb{E}_1[T_{i_0}] \le \frac{T}{K-1}$.

Now applying two point argument to $(\theta_1, \theta_{2,i_0})$ and $\Delta \asymp \sqrt{\frac{K}{T}}$ gives

$D_{KL}(P_{\theta_1} \| P_{\theta_{2,i_0}}) = \Theta(1)$, and therefore

$$r^* = \Omega(T\Delta) = \Omega(\sqrt{KT}). \quad \boxed{\checkmark}$$

Testing multiple hypotheses

Why two points may fail? Look at normal mean estimation in high dimensions:

$X \sim N(\theta, \sigma^2 I_n)$, with loss $L(\theta, \hat\theta) = \|\hat\theta - \theta\|_2^2$.

Two point method gives at best

$$r^* \ge \sup_{\theta_0, \theta_1} \frac{\|\theta_0 - \theta_1\|_2^2}{2} \left(1 - \Phi\left(\frac{\|\theta_0 - \theta_1\|_2}{2\sigma}\right)\right) \asymp \sigma^2.$$

This doesn't capture the dependence on the dimension n! (Recall that $r^* = n\sigma^2$ by Anderson's lemma)

At a high level, this is because testing between two hypotheses does NOT capture the true difficulty of a high-dimensional problem!

Challenge in high dimensions:

- separation condition: there might be different separation conditions
- indistinguishability condition: instead of the binary testing case where $1 - TV(P, Q)$ is a tight measure of the test measure, this tight measure in the multiple hypotheses case is often not tractable (and needs further lower bound)

Pairwise separation: Fano's inequality

Thm. Let $\theta_1, \cdots, \theta_m \in \Theta$ satisfy

$$\min_{i \neq j} \inf_a \left(L(\theta_i, a) + L(\theta_j, a) \right) \geq \Delta.$$

Then for $\pi = \mathrm{Unif}(\{\theta_1, \cdots, \theta_m\})$,

$$r_\pi^* \geq \frac{\Delta}{2} \left(1 - \frac{I(\theta; X) + \log 2}{\log m} \right), \qquad \theta \sim \pi,\ X \mid \theta \sim P_\theta.$$

Before we prove it, we establish a useful "golden formula" for mutual info:

Lemma. $I(X; Y) = \min_{Q_Y} D_{KL}(P_{XY} \| P_X Q_Y) = \min_{Q_Y} \mathbb{E}_{P_X}[D_{KL}(P_{Y|X} \| Q_Y)]$

Pf. Simply note that $I(X; Y) = D_{KL}(P_{XY} \| P_X Q_Y) - D_{KL}(P_Y \| Q_Y)$, $\forall Q_Y$. $\square$

Pf of Fano.

$$P_{\theta X} \quad \text{and} \quad P_\theta P_X \xrightarrow{(\theta, X) \mapsto \mathbb{1}(L(\theta, \hat{\theta}(X)) < \frac{\Delta}{2})} \mathrm{Bern}(\mathbb{P}(L(\theta, \hat{\theta}) < \tfrac{\Delta}{2})) \quad \text{and} \quad \mathrm{Bern}(q)$$

Note that $q \leq \frac{1}{m}$ by separation condition. Then DPI gives

$$D_{KL}(\mathrm{Bern}(\mathbb{P}(L(\theta, \hat{\theta}) < \tfrac{\Delta}{2})) \| \mathrm{Bern}(q)) \leq I(\theta; X)$$

$$\Rightarrow \mathbb{P}(L(\theta, \hat{\theta}) \geq \tfrac{\Delta}{2}) \geq 1 - \frac{I(\theta; X) + \log 2}{\log m}.$$ The rest follows from Markov's ineq. $\square$

In fact, the previous argument establishes a general result:

Thm (generalized Fano). For any prior π and $\Delta > 0$,

$$r_\pi^* \geq \Delta\left(1 - \frac{I(\theta;X) + \log 2}{\log\left(\frac{1}{p_\Delta}\right)}\right), \quad \theta \sim \pi, \; X|\theta \sim P_\theta$$

where $p_\Delta = \sup_a \pi(L(\theta, a) < \Delta)$ is the "small-ball probability".

The classical Fano's inequality is a special case with $\pi \sim \mathrm{Unif}(\{\theta_1, \cdots, \theta_m\})$ and a pairwise separation condition.

Additive separation: Assouad's lemma

Thm. For a hypercube parametrization $u \in \{\pm 1\}^d$, associate $\theta_u \in \Theta$.

Suppose

$$\inf_a \left(L(\theta_u, a) + L(\theta_{u'}, a)\right) \geq \Delta \cdot \sum_{i=1}^d 1(u_i \neq u_i')$$

then for the prior $\pi = \mathrm{Unif}(\{\theta_u\}_{u \in \{\pm 1\}^d})$,

$$r_\pi^* \geq \frac{\Delta}{4} \sum_{i=1}^d \left(1 - TV(P_{i,+}, P_{i,-})\right).$$

where

$$P_{i,+} = \frac{1}{2^{d-1}} \sum_{u: u_i = 1} P_{\theta_u}, \quad P_{i,-} = \frac{1}{2^{d-1}} \sum_{u: u_i = -1} P_{\theta_u}.$$

The following corollaries are often used:

Corollary 1: $$r_\pi^* \geq \frac{d\Delta}{4}\left(1 - \max_{u, u' \text{ neighbors}} TV(P_{\theta_u}, P_{\theta_{u'}})\right).$$ (the classical Assouad)

Corollary 2: $$r_\pi^* \geq \frac{d\Delta}{4}\left(1 - \mathbb{E}_{u \sim \mathrm{Unif}(\{\pm 1\}^d)} \mathbb{E}_{i \sim \mathrm{Unif}([d])} TV(P_{\theta_u}, P_{\theta_{\underbrace{u \oplus i}_{u \text{ with } i\text{-th bit flipped}}}})\right)$$

Pf of Assouad. For any estimator $\hat\theta$, construct $\hat u = (\hat u_1, \cdots, \hat u_d) \in \{\pm 1\}^d$ as follows:

$$\hat u = \arg\min_u L(\theta_u, \hat\theta).$$

Then $\forall u \in \{\pm 1\}^d$,

$$L(\theta_u, \hat\theta) \geq \frac{L(\theta_u, \hat\theta) + L(\theta_{\hat u}, \hat\theta)}{2} \geq \frac{\Delta}{2} \sum_{i=1}^d 1(u_i \neq \hat u_i).$$

$$\Rightarrow \frac{1}{2^d} \sum_u \mathbb{E}_{\theta_u}[L(\theta_u, \hat\theta)] \geq \frac{1}{2^d} \sum_u \frac{\Delta}{2} \sum_{i=1}^d P_{\theta_u}(u_i \neq \hat u_i)$$

$$= \frac{\Delta}{4} \sum_{i=1}^d \left(P_{i,+}(\hat u_i \neq 1) + P_{i,-}(\hat u_i \neq -1) \right)$$

$$\geq \frac{\Delta}{4} \sum_{i=1}^d (1 - TV(P_{i,+}, P_{i,-})). \qquad \boxtimes$$

(Exercise: show that $\frac{1}{d} \sum_{i=1}^d TV(P_{i,+}, P_{i,-}) = 1 - \Omega(1) \Longleftrightarrow \frac{1}{d} \sum_{i=1}^d I(U_i ; X) = \log 2 - \Omega(1)$ for $U \sim \mathrm{Unif}(\{\pm 1\}^d)$.

In addition, since $I(U; X) \geq \sum_{i=1}^d I(U_i ; X)$, under this hypercube condition, Assouad is no weaker than Fano)

Fano and Assouad are good at proving lower bounds for <u>high-dimensional</u> targets.

Example 2.1 (normal mean model) $X \sim N(\theta, \sigma^2 I_n)$, $L(\theta, \hat\theta) = \|\theta - \hat\theta\|_2^2$.

We will show that $r^* = \Omega(n\sigma^2)$.

Pf 1 (Fano). Construct a subset Θ_0 of $\{\pm\delta\}^n$ (with δ TBD) s.t.

- $m := |\Theta_0|$ is large enough ;
- $\min_{\theta \neq \theta' \in \Theta_0} \|\theta - \theta'\|_2^2 \geq \frac{\delta^2 n}{5}$.

By Gilbert–Varshamov bound below, we can make $m = e^{\Omega(n)}$. Then $\Delta = \frac{\delta^2 n}{10}$, and

$$I(\theta; X) \underset{\substack{\uparrow \\ \text{golden formula}}}{\leq} \max_{\theta \in \Theta_0} D_{KL}(N(\theta, \sigma^2 I_n) \| N(0, \sigma^2 I_n)) = \frac{n\delta^2}{2\sigma^2}.$$

Fano $\Rightarrow$ $r^* = \Omega\left(\delta^2 n \left(1 - \frac{n\delta^2/2\sigma^2 + \log 2}{\Omega(n)}\right)\right) \xrightarrow{\delta \asymp \sigma} r^* = \Omega(n\sigma^2)$.

Lemma (Gilbert-Varshamov) $\exists A \subseteq \{\pm 1\}^n$ s.t. $\min_{\substack{u,u'\in A\\ u\neq u'}} \sum_{i=1}^n 1(u_i \neq u_i') \geq d$, and

$$m \geq \frac{2^n}{\sum_{j=0}^{d-1}\binom{n}{j}} = 2^{n(1-h_2(\frac{d}{n}))+o(n)} \quad \left(h_2(x) = x\log_2\frac{1}{x} + (1-x)\log_2\frac{1}{1-x}\right)$$

Pf. Volume argument: $\forall u \in \{\pm1\}^n$, $\left|\{u' \in \{\pm 1\}^n : \sum_{i=1}^n 1(u_i \neq u_i') \leq d-1\}\right| = \sum_{j=0}^{d-1}\binom{n}{j}$.
So if $m < \frac{2^n}{\sum_{j=0}^{d-1}\binom{n}{j}}$, there must exist some $u \in \{\pm1\}^n$ with distance $\geq d$ to all existing points. $\boxtimes$

Pf 2 (Generalized Fano) Let $\theta \sim \mathrm{Unif}(\{\pm\delta\}^n)$, then $I(\theta;X) \leq \frac{n\delta^2}{2\sigma^2}$

Pick $\Delta = \frac{n\delta^2}{12}$, then the small-ball probability is

$$p_\Delta = \sup_a \pi(\|\theta - a\|_2^2 < \Delta) \leq \sup_{\hat\theta \in \{\pm\delta\}^n} \pi(\|\theta-\hat\theta\|_2^2 < 4\Delta) = \frac{1}{2^n}\sum_{j=0}^{\lceil n/3\rceil - 1}\binom{n}{j} = 2^{-\Omega(n)}.$$

$\hat\theta = \arg\min_\theta \|a-\theta\|_2^2$; Stirling

Therefore, we again get $r^* = \Omega\left(n\delta^2\left(1 - \frac{\frac{n\delta^2}{2\sigma^2} + \log 2}{\Omega(n)}\right)\right) \overset{\delta \asymp \sigma}{=} \Omega(n\sigma^2)$.

Pf 3 (Assouad). For $\delta > 0$, let $\theta_u = \delta u$, $u \in \{\pm1\}^n$. Then $\Delta = 2\delta^2$, and

$$1 - \max_{u,u' \text{ neighbors}} TV(N(\theta_u, \sigma^2 I_n), N(\theta_{u'}, \sigma^2 I_n)) = 2\left(1 - \Phi\left(\frac{\delta}{\sigma}\right)\right).$$

Choosing $\delta = \sigma$, Assouad's lemma gives $r^* = \Omega(n\Delta) = \Omega(n\sigma^2)$.

Example 2.2 (learning theory). $(X_1,Y_1),\cdots,(X_n,Y_n) \sim P_{XY}$ (unknown) with $Y_i \in \{0,1\}$.
Let $\mathcal{F}$ be a given class of functions $\mathcal{X} \to \{0,1\}$ with VC dimension d.
Excess risk for a trained classifier $\hat f : \mathcal{X} \to \{0,1\}$ based on $(X_1,Y_1),\cdots,(X_n,Y_n)$:

$$ER(\hat f) := ER(\hat f; P_{XY}) := P_{XY}(Y \neq \hat f(X)) - \min_{f\in\mathcal{F}} P_{XY}(Y \neq f(X))$$

We will show that for $n \geq d$,

$$\inf_{\hat f}\sup_{P_{XY}} \mathbb{E}_{P_{XY}}[ER(\hat f)] = \Omega\left(\sqrt{\tfrac{d}{n}}\right) \quad \text{(agnostic setting)}$$

$$\inf_{\hat f}\sup_{\substack{P_{XY}:\exists f\in\mathcal{F}\\ Y=f(X)\ P_{XY}\text{-a.s.}}} \mathbb{E}_{P_{XY}}[ER(\hat f)] = \Omega\left(\tfrac{d}{n}\right) \quad \text{(realizable setting)}$$

Recall: VC-dim$(\mathcal{F}) = d \Rightarrow \exists x_1, \cdots, x_d \in \mathcal{X}, \forall u \in \{\pm 1\}^d, \exists f_u \in \mathcal{F}$ s.t. $f_u(x_i) = u_i$
$\forall i \in [d]$.

Pf (agnostic case). Use $x_1, \cdots, x_d \in \mathcal{X}$ and $\{f_u\}_{u \in \{\pm1\}^d} \subseteq \mathcal{F}$ in the above definition.

Under $u \in \{\pm 1\}^d$, construct $P_u := P_{XY,u}$ as:

$$X \sim \mathrm{Unif}(\{x_1, \cdots, x_d\}), \qquad Y \mid X = x_i = \begin{cases} u_i & \text{w.p. } \frac{1}{2} + \delta \\ -u_i & \text{w.p. } \frac{1}{2} - \delta. \end{cases}$$

- Separation condition: $\min_{f \in \mathcal{F}} P_u(f(X) \neq Y) = \frac{1}{2} - \delta, \quad \forall u.$

$\forall f_0$:

$$\begin{aligned} & ER(f_0, P_u) + ER(f_0, P_{u'}) \\ &= P_u(Y \neq f_0(X)) + P_{u'}(Y \neq f_0(X)) - 2(\tfrac{1}{2} - \delta) \\ &\geq \sum_{i=1}^d \frac{1}{d}\left(1(u_i \neq u_i') \cdot 1 + 1(u_i = u_i') \cdot (1 - 2\delta)\right) - (1 - 2\delta) \\ &= \frac{2\delta}{d} \sum_{i=1}^d 1(u_i \neq u_i') \quad \Longrightarrow \quad \Delta = \frac{2\delta}{d}. \end{aligned}$$

- Indistinguishability condition: for neighboring u, u':

$$\begin{aligned} D_{KL}(P_u^{\otimes n} \| P_{u'}^{\otimes n}) &= n D_{KL}(P_u \| P_{u'}) = n \cdot \frac{1}{d} D_{KL}(\mathrm{Bern}(\tfrac{1}{2} + \delta) \| \mathrm{Bern}(\tfrac{1}{2} - \delta)) \\ &= O(\tfrac{n\delta^2}{d}) \quad \text{for } \delta = \tfrac{1}{2} - \Omega(1). \end{aligned}$$

Finally, choosing $\delta \asymp \sqrt{\frac{d}{n}}$, Assouad's lemma gives $r^* = \Omega(d\Delta) = \Omega(\sqrt{\frac{d}{n}})$.

(Realizable case) For $u \in \{(1, \pm1, \cdots, \pm1)\}$, now define $P_u := P_{XY,u}$ as:

$$X = \begin{cases} x_1 & \text{w.p. } 1 - \frac{d-1}{n} \\ x_i & \text{w.p. } \frac{1}{n} \quad \forall 2 \le i \le d \end{cases}, \qquad Y \mid X = x_i = u_i \quad \text{w.p. } 1.$$

Clearly $\min_{f \in \mathcal{F}} P_u(f(X) \neq Y) = 0, \quad \forall u.$

- Separation condition: $\Delta = \frac{1}{n}$ by a similar analysis
- Indistinguishability condition: for $u' = u^{\oplus i}$,

$$TV(P_u^{\otimes n}, P_{u'}^{\otimes n}) \le P(x_i \text{ appears in } X_1, \cdots, X_n) = 1 - (1 - \tfrac{1}{n})^n = 1 - \Omega(1).$$

Therefore, Assouad's lemma gives $r^* = \Omega((d-1)\Delta) = \Omega(\frac{d}{n})$.

(See HW for a further generalization.)

Assouad's lemma is also surprisingly flexible in sequential settings.

<u>Example 2.3 (distribution estimation under sequential communication protocols)</u>

Let $P = (p_1, \cdots, p_k)$ be an unknown pmf.
The observations are $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} P$, but $\hat{P}$ must be formed via the distributed diagram on the right, together with communication constraints $\ell \le k$.

communication protocol Π to be designed

$$
\left.\begin{array}{ccc}
X_1 & \longrightarrow & Y_1 \in [\ell] \\
(X_2, Y_1) & \longrightarrow & Y_2 \in [\ell] \\
\vdots & & \\
(X_n, Y^{n-1}) & \longrightarrow & Y_n \in [\ell]
\end{array}\right\} \longrightarrow \hat{P}
$$

distributed nodes | central server | final estimator

We will show that,
$$r^* = \inf_{(\hat{P}, \Pi)} \sup_{P} \mathbb{E}_P[\mathrm{TV}(P, \hat{P})] = \Omega\left(\frac{k}{\sqrt{n\ell}}\right) \quad \text{if} \quad \begin{array}{l} k \le \ell \\ n \ge k^2/\ell . \end{array}$$

<u>Pf</u>. WLOG assume k is even. For $u \in \{\pm 1\}^{k/2}$, construct
$$P_u = \left(\frac{1+\delta u_1}{k}, \frac{1-\delta u_1}{k}, \cdots, \frac{1+\delta u_{k/2}}{k}, \frac{1-\delta u_{k/2}}{k}\right), \quad \delta \in (0, \tfrac{1}{2}) \text{ TBD.}$$

Easy to check that $\Delta = \Omega(\frac{\delta}{k})$. Next we use Corollary 2 of Assouad's lemma and try to upper bound
$$
\begin{aligned}
\mathbb{E}_u \mathbb{E}_i [\mathrm{TV}(P_{Y^n|u}, P_{Y^n|u^{\oplus i}})] &\le \sqrt{\mathbb{E}_u \mathbb{E}_i [\mathrm{TV}(P_{Y^n|u}, P_{Y^n|u^{\oplus i}})^2]} \quad \text{(Jensen)} \\
&\le \sqrt{\mathbb{E}_u \mathbb{E}_i [H^2(P_{Y^n|u}, P_{Y^n|u^{\oplus i}})]} \quad (\mathrm{TV} \le H) \\
&\le C \sqrt{\sum_{t=1}^{n} \mathbb{E}_u \mathbb{E}_i \mathbb{E}_{P_{Y^{t-1}|u}} [H^2(P_{Y_t|Y^{t-1},u}, P_{Y_t|Y^{t-1},u^{\oplus i}})]}
\end{aligned}
$$
(Jayram's subadditivity of H^2 in Lec 3)

Note that $\quad P_{Y_t|Y^{t-1},u} = \sum_{x \in [k]} P_{Y_t|Y^{t-1}, X_t = x} \cdot P_u(x) \ge \frac{1}{2k} \sum_{x \in [k]} P_{Y_t|Y^{t-1}, X_1 = x}$,

so
$$
\begin{aligned}
\chi^2(P_{Y_t|Y^{t-1},u^{\oplus i}} \| P_{Y_t|Y^{t-1},u}) &\le \sum_{y \in [\ell]} \frac{(P_{Y_t=\ell|Y^{t-1},u^{\oplus i}} - P_{Y_t=\ell|Y^{t-1},u})^2}{\frac{1}{2k}\sum_{x \in [k]} P_{Y_t=y|Y^{t-1},X_t=x}} \\
&= 2k\left(\frac{2\delta}{k}\right)^2 \sum_{y \in [\ell]} \frac{(P_{Y_t=\ell|Y^{t-1},X_t=2i-1} - P_{Y_t=\ell|Y^{t-1},X_t=2i})^2}{\sum_{x \in [k]} P_{Y_t=y|Y^{t-1},X_t=x}} \\
&\le \frac{8\delta^2}{k} \sum_{y \in [\ell]} \frac{P_{Y_t=\ell|Y^{t-1},X_t=2i-1} + P_{Y_t=\ell|Y^{t-1},X_t=2i}}{\sum_{x \in [k]} P_{Y_t=y|Y^{t-1},X_t=x}} .
\end{aligned}
$$

Therefore,

$$\mathbb{E}_i\left[H^2\left(P_{Y_t|Y^{t-1},u}, P_{Y_t|Y^{t-1},u^{\oplus i}}\right)\right] \le \mathbb{E}_i\left[\chi^2\left(P_{Y_t|Y^{t-1},u^{\oplus i}} \,\|\, P_{Y_t|Y^{t-1},u}\right)\right]$$

$$= O\left(\frac{\delta^2 \ell}{k^2}\right).$$

Putting everything together, Assouad's lemma yields

$$r^* = \Omega\left(\delta\left(1 - O\left(\sqrt{\frac{n\delta^2\ell}{k^2}}\right)\right)\right) \overset{\delta \asymp \frac{k}{\sqrt{n\ell}},\ \delta < \frac{1}{2}}{=} \Omega\left(\frac{k}{\sqrt{n\ell}}\right).$$

Special topic: interactive Le Cam

Model for interactive decision making:

- unknown true model M^* in a given model class $\mathcal{M}$ (e.g. reward distribution of all arms)
- at each round $t = 1, \cdots, T$:
  1. learner chooses an action $a_t \in \mathcal{A}$;
  2. nature reveals reward $r_t \in [0,1]$ and possible additional observation o_t, with $\mathbb{E}[r_t \mid a_t = a] = r^{M^*}(a)$, and $(r_t, o_t) \sim M^*(a_t)$
- learner aims to minimize the regret

$$R_T = \sum_{t=1}^{T} \left(r_*^{M^*} - r^{M^*}(a_t) \right), \quad \text{where} \quad r_*^{M} = \max_{a \in \mathcal{A}} r^M(a).$$

Example (multi-armed bandit)

- $\mathcal{A} = [K]$
- $M^\mu = \{\mathrm{Bern}(\mu_i)\}_{i=1}^{K}$, $\mathcal{M} = \{ M^\mu : \mu \in [0,1]^K \}$.
- $M^\mu(a) = \mathrm{Bern}(\mu_a)$. $r^{M^\mu}(a) = \mu_a$.

Question: What is a general two point lower bound for R_T?

Idea: let $g^M(a) := r_*^M - r^M(a)$ denote the "gap" of $a \in \mathcal{A}$ under M, then point lower bound suggests that

$$\inf_{\{a_t\}} \sup_{M^*} \mathbb{E}[R_T] \ge T \cdot \begin{cases} \sup_{M_0, M_1 \in \mathcal{M}} \left(\inf_{a \in \mathcal{A}} g^{M_0}(a) + g^{M_1}(a) \right) \\ \text{s.t.} \quad H^2(M_0, M_1) \le \frac{c}{T} \text{ for small } c > 0. \end{cases}$$

Challenges: ① The metric $\inf_a (g^{M_0}(a) + g^{M_1}(a))$ could be too pessimistic. For example, if a policy uses an action distribution p on a under M_0, then $\mathbb{E}_{a\sim p}[g^{M_1}(a)]$ might be a better separation metric.

② $H^2(M_0, M_1)$ is NOT well-defined, as the distributions of (r_t, o_t) depend on a_t, i.e. $(r_t, o_t) \sim M^*(a_t)$. So $H^2(M_0, M_1)$ should be replaced by $\mathbb{E}_{a\sim p}[H^2(M_0(a), M_1(a))]$ for some p, and take $\inf_p$ somewhere.

③ Where to take $\inf_p$ (learner as the min player)?

- $\sup_{M_0, M_1} \inf_p$: too small, by the same reason in ①
- $\inf_p \sup_{M_0, M_1}$: too large, as the learner can adjust the action distribution in the sequential setting.

Defn. The constrained decision-to-estimation coefficient (DEC) is defined as

$$\mathrm{dec}_\varepsilon(\mathcal{M}) = \sup_{\bar{M}} \inf_{p \in \Delta(\mathcal{A})} \sup_{M \in \mathcal{M} \cup \{\bar{M}\}} \left\{ \mathbb{E}_{a\sim p}[g^M(a)] : \mathbb{E}_{a\sim p}[H^2(M(a), \bar{M}(a))] \le \varepsilon^2 \right\}.$$

- This is a sup-inf-sup structure: first choose a reference model $\bar{M}$, learner chooses an action distribution p based on $\bar{M}$, finally nature chooses an alternative model M
- The separation condition is w.r.t. $a \sim p$
- The reference model $\bar{M}$ doesn't need to belong to $\mathcal{M}$

Example 3.1 (two-armed bandit). For two-armed bandit $(\mathrm{Bern}(\mu_1), \mathrm{Bern}(\mu_2))$ with $|\mu_1 - \mu_2| \ge \Delta$, choose $\bar{M} = (\mathrm{Bern}(\frac{1}{2}+\Delta), \mathrm{Bern}(\frac{1}{2}))$ and $M \in \{ (\mathrm{Bern}(\frac{1}{2}+\Delta), \mathrm{Bern}(\frac{1}{2})),$
$(\mathrm{Bern}(\frac{1}{2}+\Delta), \mathrm{Bern}(\frac{1}{2}+\Delta+\delta)) \}$

$$\mathrm{dec}_\varepsilon(\mathcal{M}) \ge \inf_{p_2 \in [0,1]} \max\left\{ p_2 \Delta, (1-p_2)\left(\frac{c\varepsilon}{\sqrt{p_2}} - \Delta\right)_+ \right\} = \Omega\left(\Delta \wedge \frac{\varepsilon^2}{\Delta}\right).$$

Example 3.2 (multi-armed bandit) For $\mathcal{M} = \{ \prod_{i=1}^K \mathrm{Bern}(\mu_i) : \mu_1, \dots, \mu_K \in [0,1] \}$, can choose $\bar{M} = \prod_{i=1}^K \mathrm{Bern}(\frac{1}{2})$, pick $i_0 = \arg\min_i p_i$ and set $M(i_0) = \mathrm{Bern}(\frac{1}{2} + \varepsilon\sqrt{K})$ to get $\mathrm{dec}_\varepsilon(\mathcal{M}) = \Omega(\varepsilon\sqrt{K})$ if $\varepsilon\sqrt{K} = O(1)$.

$$\boxed{\begin{array}{l}\underline{\text{Thm (DEC lower bound)}}.\ \text{There exist absolute constants } c, C>0 \text{ s.t.} \\ \qquad \inf_{\{a_t\}} \sup_{M^* \in \mathcal{M}} \mathbb{E}_{M^*}[R_T] = \Omega\left(T \cdot (\mathrm{dec}_\varepsilon(\mathcal{M}) - C\varepsilon)_+\right), \quad \text{for } \varepsilon = \sqrt{\frac{c}{T}}.\end{array}}$$

Specializing to previous examples, this gives a lower bound $\Omega(\frac{1}{\Delta})$ for Example 3.1 when $\Delta \geq \frac{C}{\sqrt{T}}$, and $\Omega(\sqrt{KT})$ for $T \geq K$.

Pf. (Simpler case: $\overline{M} \in \mathcal{M}$, from (Foster, Golowich, Han '23)) Let $\Delta := \mathrm{dec}_\varepsilon(\mathcal{M})$.

$p_{\overline{M}} := \mathbb{E}_{\overline{M}}[\frac{1}{T}\sum_{t=1}^{T} p_t(\cdot \mid H^{t-1})]$: learner's average play under $\overline{M}$

M : inner maximizer under $p = p_{\overline{M}}$

$p_M := \mathbb{E}_M[\frac{1}{T}\sum_{t=1}^{T} p_t(\cdot \mid H^{t-1})]$: learner's average play under M.

By definition, $\quad \mathbb{E}_{a\sim p_{\overline{M}}}[g^M(a)] \geq \Delta \quad$ ①

$\qquad \mathbb{E}_{a\sim p_{\overline{M}}}[H^2(M(a), \overline{M}(a))] \leq \varepsilon^2 \quad$ ②

By the sake of contradiction, we can assume that

$\qquad \mathbb{E}_{a\sim p_M}[g^M(a)] \leq \frac{\Delta}{100} \quad$ ③

$\qquad \mathbb{E}_{a\sim p_{\overline{M}}}[g^{\overline{M}}(a)] \leq \frac{\Delta}{100} \quad$ ④

We introduce some useful lemmas.

Lemma 1. For $c>0$ small enough, $TV(p_M, p_{\overline{M}}) \leq 0.1$.

Pf.
$$\begin{aligned} TV(p_M, p_{\overline{M}})^2 &\leq TV(P^M_{r^T, o^T}, P^{\overline{M}}_{r^T, o^T})^2 \quad (\text{DPI}) \\ &\leq H^2(P^M_{r^T, o^T}, P^{\overline{M}}_{r^T, o^T}) \\ &\leq C\sum_{t=1}^{T} \mathbb{E}_{P^{\overline{M}}}[H^2(P^M_{r_t, o_t \mid H^{t-1}}, P^{\overline{M}}_{r_t, o_t \mid H^{t-1}})] \\ &\qquad (\text{subadditivity of } H^2) \\ &= C\sum_{t=1}^{T} \mathbb{E}_{P^{\overline{M}}}[H^2(M(a_t), \overline{M}(a_t))] \\ &= CT \cdot \mathbb{E}_{a\sim p_{\overline{M}}}[H^2(M(a), \overline{M}(a))] \\ &\leq 0.1 \qquad (\text{by } ②) \end{aligned}$$

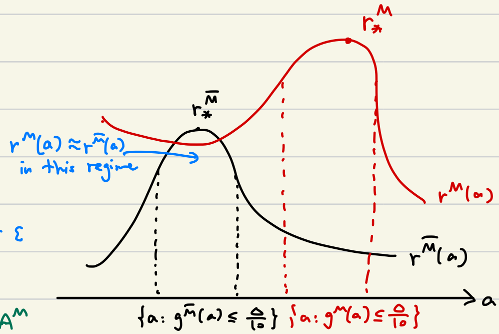
$\boxed{3}$

Lemma 2. $\mathbb{E}_{a\sim p_{\bar M}} | r^M(a) - r^{\bar M}(a) | \le \varepsilon$. (This step critically uses that the rewards are observed)

Pf. As $r_t \in [0,1]$,

$$\text{LHS} \le \mathbb{E}_{a\sim p_{\bar M}}[TV(M(a), \bar M(a))] \le \mathbb{E}_{a\sim p_{\bar M}}[H(M(a), \bar M(a))] \le \varepsilon . \quad \square$$

Next we present the proof.



1) By Lemma 2 and ①:

$$\begin{aligned} \Delta &\le r_*^M - \mathbb{E}_{a\sim p_{\bar M}}[r^M(a)] \\ &\le r_*^M - \mathbb{E}_{a\sim p_{\bar M}}[r^{\bar M}(a)] + \varepsilon \\ &= r_*^M - r_*^{\bar M} + \mathbb{E}_{a\sim p_{\bar M}}[g^{\bar M}(a)] + \varepsilon \\ &\overset{④}{\le} r_*^M - r_*^{\bar M} + \frac{\Delta}{100} + \varepsilon \end{aligned}$$

$$\Longrightarrow r_*^M - r_*^{\bar M} \ge \frac{99\Delta}{100} - \varepsilon .$$

2) By ③ and Markov: $P_M(\underbrace{\{a: g^M(a) \le \frac{\Delta}{10}\}}_{=:A^M}) \ge \frac{9}{10}$

3) By 2) and Lemma 1: $P_{\bar M}(A^M) \ge \frac{4}{5}$

4) By 1) and 3): $\mathbb{E}_{a\sim p_{\bar M}}[(r^M(a) - r^{\bar M}(a)) \mathbb{1}_{a\in A^M}]$

$$\ge (r_*^M - \frac{\Delta}{10} - r_*^{\bar M}) \cdot P_{\bar M}(A^M) \ge (\frac{89\Delta}{100} - \varepsilon) \cdot \frac{4}{5} .$$

However, Lemma 2 states that LHS $\le \varepsilon$. This is a contradiction when $\Delta > C\varepsilon$!

(General $\bar M$, from (Glasgow and Rakhlin '23))

For $\bar M \notin \mathcal{M}$, ④ is no longer a result of small regret.

Solution: a stopping time argument. Let ALG be the original learner's algorithm, define ALG' as follows: $ALG_t = ALG'_t$ as long as $\sum_{s<t} g^{\bar M}(a_s) < \frac{\Delta T}{100}$.

and ALG_t always pulls $a^* = \arg\max_a r^{\bar M}(a)$ o.w.

Now redefine $p_{\bar M}, M, p_M$ using ALG', then ①②④ + Lemma 1 & 2 still hold.

Let $\tau > 0$ be the stopping time of $\sum_{t=1}^{\tau} g^{\bar M}(a_t) \ge \frac{\Delta T}{100}$.

By Lemma 2 and Markov's inequality, w.p. ≥ 0.9 under $\mathbb{P}_{\bar{M}}^{ALG'}$,

$$\frac{1}{T}\sum_{t=1}^{T\wedge\tau} g^{M}(a_t) = \frac{1}{T}\sum_{t=1}^{T\wedge\tau}\left(r_*^{M} - r^{M}(a_t)\right)$$

$$\geq \frac{1}{T}\sum_{t=1}^{T\wedge\tau}\left(r_*^{M} - r^{\bar{M}}(a_t)\right) - \frac{1}{T}\sum_{t=1}^{T}\left|r^{M}(a_t) - r^{\bar{M}}(a_t)\right|$$

$$\overset{\text{Lemma 2}}{\geq} \frac{1}{T}\sum_{t=1}^{T\wedge\tau} g^{\bar{M}}(a_t) + \frac{T\wedge\tau}{T}\left(r_*^{M} - r_*^{\bar{M}}\right) - 10\varepsilon$$

$$\overset{1)}{\geq} \frac{1}{T}\sum_{t=1}^{T\wedge\tau} g^{\bar{M}}(a_t) + \frac{T\wedge\tau}{T}\left(0.99\Delta - \varepsilon\right) - 10\varepsilon.$$

- If $\tau \geq T$, the RHS is $\geq 0.99\Delta - 11\varepsilon = \Omega(\Delta)$ for $\Delta > C\varepsilon$;
- If $\tau < T$, the RHS $\geq \frac{1}{T}\frac{\Delta T}{100} - 10\varepsilon = \Omega(\Delta)$ for $\Delta > C\varepsilon$.

$$\Longrightarrow \quad \mathbb{P}_{\bar{M}}^{ALG'}\left(\frac{1}{T}\sum_{t=1}^{T\wedge\tau} g^{M}(a_t) = \Omega(\Delta)\right) \geq 0.9 \quad \text{in both cases.}$$

Since $TV(\mathbb{P}_{\bar{M}}^{ALG'}, \mathbb{P}_{M}^{ALG'}) \leq 0.1$ by Lemma 1, we get

$$\mathbb{P}_{M}^{ALG'}\left(\frac{1}{T}\sum_{t=1}^{T\wedge\tau} g^{M}(a_t) = \Omega(\Delta)\right) \geq 0.8.$$

Finally, since ALG′ and ALG coincide up to time $\tau\wedge T$, this gives the claimed result.

(In (Glasgow and Rakhlin '23), this stopping time argument establishes a stronger claim:

$$\inf_{\{a_t\}}\sup_{M^*} \mathbb{P}_{M^*}\left(\frac{R(T)}{T} > \left((1-c_0)\,\mathrm{dec}_\varepsilon(\gamma) - C\varepsilon\right)_+\right) = \Omega(1), \quad \varepsilon = \sqrt{\frac{c}{T}}$$

for any fixed $c_0 > 0$, where c, C depend on c_0.)