COMS 4773: Minimax lower bounds

Daniel Hsu

May 1, 2024

1 Learning/estimation problems

- Z is the dataset (e.g., an iid sample of n training examples); $\mathcal{Z}$ is the space of possible datasets.
- P_θ is a probability distribution for Z, one per $\theta \in \Theta$ (e.g., P_θ is distribution of n iid examples where each example is drawn from p_θ, a distribution over $\mathcal{X} \times \{-1, 1\}$)
- $f : \mathcal{Z} \to \Theta$ is an estimator a.k.a. learning algorithm
- $\ell(\hat{\theta}, \theta) \in \mathbb{R}_+$ is a measure of how bad $\hat{\theta}$ is as an estimate of θ
- Risk of f under P_θ:
$$\mathbb{E}_{Z \sim P_\theta}[\ell(f(Z), \theta)]$$
- Minimax risk
$$\min_f \max_{\theta \in \Theta} \mathbb{E}_\theta[\ell(f(Z), \theta)]$$
- Upper bound UB on minimax risk: design learning algorithm with worst-case risk $\leq UB$
- Lower bound LB on minimax risk: prove that every learning algorithm has worst-case risk $\geq LB$

1.1 Example: learning binary classifier

- Z is n training examples from $\mathcal{X} \times \{-1, 1\}$
- P_θ is a distribution where examples $Z = (X_i, Y_i)_{i=1}^n$ are iid from a probability distribution p_θ over $\mathcal{X} \times \{-1, 1\}$ indexed by $\theta \in \Theta$
  - Let $h_\theta \in \mathcal{H}$ be hypothesis from hypothesis class $\mathcal{H}$ with smallest error rate under p_θ
- $f : \mathcal{Z} \to \Theta$ is an estimator that "guesses" $\theta \in \Theta$
  - Later, we'll show that if you have a good learning algorithm $A : \mathcal{Z} \to \mathcal{H}$, then you can get a good estimator $f : \mathcal{Z} \to \Theta$
- $\ell(\hat{\theta}, \theta) = \mathbb{1}\{\mathrm{err}_\theta(h_{\hat{\theta}}) - \mathrm{err}_\theta(h_\theta) \geq \varepsilon\}$
- "Risk" of $\hat{\theta} = f(Z)$:
$$\mathbb{E}_\theta[\ell(\hat{\theta}, \theta)] = \Pr_\theta[\mathrm{err}_\theta(h_{\hat{\theta}}) - \mathrm{err}_\theta(h_\theta) \geq \varepsilon]$$

2 Le Cam's "two-point" method

- Only two possible data distributions, P_{-1} and P_{+1}
- You get data $Z \sim P_\sigma$, and then make a guess of σ; what is the probability you guess σ incorrectly?
- Lemma:

$$\begin{aligned}\min_f \max_{\sigma \in \{-1,1\}} \Pr_{Z \sim P_\sigma}[f(Z) \neq \sigma] &\geq \frac{1}{2} \sum_{z \in \mathcal{Z}} \min\{P_{-1}(z), P_{+1}(z)\} \\ &= \frac{1}{2}(1 - \|P_{-1} - P_{+1}\|_{\mathrm{TV}})\end{aligned}$$

- Proof: Draw $\sigma \sim \mathrm{unif}\{-1,1\}$, and then draw $Z \mid \sigma \sim P_\sigma$.
  - "Bayes (optimal) classifier" that minimizes $\Pr[f(Z) \neq \sigma]$ is

$$f^\star(z) = \begin{cases} +1 & \text{if } P_{+1}(z) > P_{-1}(z) \\ -1 & \text{if } P_{+1}(z) \leq P_{-1}(z) \end{cases}$$

  - Therefore

$$\begin{aligned}\Pr[f^\star(Z) \neq \sigma] &= \frac{1}{2} P_{-1}[f^\star(Z) = +1] + \frac{1}{2} P_{+1}[f^\star(Z) = -1] \\ &= \frac{1}{2} \sum_{z : P_{+1}(z) > P_{-1}(z)} P_{-1}(z) + \frac{1}{2} \sum_{z : P_{+1}(z) \leq P_{-1}(z)} P_{+1}(z) \\ &= \frac{1}{2} \sum_{z \in \mathcal{Z}} \min\{P_{-1}(z), P_{+1}(z)\}\end{aligned}$$

- Moreover,

$$\begin{aligned}\sum_{z \in \mathcal{Z}} \min\{P_{-1}(z), P_{+1}(z)\} &= \sum_{z \in \mathcal{Z}} \left(\frac{P_{-1}(z) + P_{+1}(z)}{2} - \frac{|P_{-1}(z) - P_{+1}(z)|}{2} \right) \\ &= 1 - \frac{1}{2} \sum_{z \in \mathcal{Z}} |P_{-1}(z) - P_{+1}(z)| \\ &= 1 - \|P_{-1} - P_{+1}\|_{\mathrm{TV}}\end{aligned}$$

  Therefore

$$\min_f \max_{\sigma \in \{-1,1\}} \Pr_\sigma[f(Z) \neq \sigma] \geq \frac{1}{2}(1 - \|P_{-1} - P_{+1}\|_{\mathrm{TV}})$$

3 Using Le Cam's method

- Suppose $\mathcal{H}$ has at least two hypotheses h_{-1} and h_{+1} that disagree on a point $x_0 \in \mathcal{X}$, with $h_\sigma(x_0) = \sigma$ for each $\sigma \in \{-1,1\}$

- Consider two data distributions P_{-1} and P_{+1} for iid examples $(X_i, Y_i)_{i=1}^n$
  - Under P_σ: $X_i = x_0$ with probability 1, and
$$Y_i = \begin{cases} +\sigma & \text{with probability } \frac{1+\varepsilon}{2}, \\ -\sigma & \text{with probability } \frac{1-\varepsilon}{2}. \end{cases}$$
  - So $\mathrm{err}_\sigma(h_\sigma) = 0.5 - \varepsilon < 0.5 + \varepsilon = \mathrm{err}_\sigma(h_{-\sigma})$
  - We'll show, using Le Cam's method, that if $n \lesssim \frac{1}{\varepsilon^2} \log \frac{1}{\delta}$, then
$$\min_f \max_{\sigma \in \{-1,1\}} P_\sigma[f(Z) \neq \sigma] > \delta$$
- Application to statistical learning:
  - For arbitrary h, we have
$$\mathrm{err}_\sigma(h) = \begin{cases} \frac{1-\varepsilon}{2} & \text{if } h(x_0) = +\sigma \\ \frac{1+\varepsilon}{2} & \text{if } h(x_0) = -\sigma \end{cases}$$
  - Given learning algorithm $A \colon \mathcal{Z} \to \mathcal{H}$, define $f_A(Z) = A(Z)(x_0)$
  - If A can guarantee
$$\Pr[\mathrm{err}(A(Z)) - \min_{h \in \mathcal{H}} \mathrm{err}(h) \geq \varepsilon] \leq \delta,$$
    then
$$\max_{\sigma \in \{-1,1\}} P_\sigma[f_A(Z) \neq \sigma] \leq \delta.$$
  - Therefore, no algorithm A can guarantee
$$\Pr[\mathrm{err}(A(Z)) - \min_{h \in \mathcal{H}} \mathrm{err}(h) \geq \varepsilon] \leq \delta$$
    if
$$n \lesssim \frac{1}{\varepsilon^2} \log \frac{1}{\delta}.$$
- Now let us prove the minimax lower bound
  - Given dataset $z = (x_i, y_i)_{i=1}^n$, let $m(z) = |\{i \in [n] : y_i = +1\}|$
  - Then
$$P_{+1}(z) = \left(\frac{1+\varepsilon}{2}\right)^{m(z)} \left(\frac{1-\varepsilon}{2}\right)^{n-m(z)},$$
$$P_{-1}(z) = \left(\frac{1-\varepsilon}{2}\right)^{m(z)} \left(\frac{1+\varepsilon}{2}\right)^{n-m(z)}.$$
    So
$$\begin{aligned} \frac{P_{+1}(z)}{P_{-1}(z)} \geq 1 &\Leftrightarrow \left(\frac{1+\varepsilon}{2}\right)^{2m(z)-n} \left(\frac{1-\varepsilon}{2}\right)^{n-2m(z)} \geq 1 \\ &\Leftrightarrow \left(\frac{1+\varepsilon}{1-\varepsilon}\right)^{2m(z)-n} \geq 1 \\ &\Leftrightarrow m(z) \geq n/2 \end{aligned}$$

– So by Le Cam's lemma

$$\min_f \max_{\sigma \in \{-1,1\}} P_\sigma[f(Z) \neq \sigma] \geq \frac{1}{2} \sum_{z \in \mathcal{Z}} \min\{P_{-1}(z), P_{+1}(z)\}$$

$$= \frac{1}{2} \left(\sum_{z \in \mathcal{Z}: m(z) \geq n/2} P_{-1}(z) + \sum_{z \in \mathcal{Z}: m(z) < n/2} P_{+1}(z) \right)$$

$$= \frac{1}{2} (P_{-1}(m(Z) \geq n/2) + P_{+1}(m(Z) < n/2))$$

– By Slud's inequality (Slud, 1977),

$$P_{-1}(m(Z) \geq n/2) \geq 1 - \Phi\left(\frac{n\varepsilon/2}{\sqrt{n(1-\varepsilon^2)/4}} \right)$$

and

$$P_{+1}(m(Z) < n/2) \geq 1 - \Phi\left(\frac{n\varepsilon/2}{\sqrt{n(1-\varepsilon^2)/4}} \right)$$

where Φ is the CDF for N(0, 1)

– Conclusion: if $n \lesssim \frac{1}{\varepsilon^2} \log \frac{1}{\delta}$, then

$$\min_f \max_{\sigma \in \{-1,1\}} P_\sigma[f(Z) \neq \sigma] > \delta$$

- Difficulty distilled to testing problem between two possible distributions
- What about dependence on "complexity" of $\mathcal{H}$? Need to consider testing problem with many possible distributions (not just two)
- Two typical approaches: Assouad's method, Fano's method
- Nice reference for statistics applications: Yu (1997)

4 Assouad's "hypercube" method

- Suppose there are 2^d distributions P_σ on $\mathcal{Z}$, one per $\sigma \in \{-1,1\}^d$
- Consider Hamming loss $\ell(\hat{\sigma}, \sigma) = \sum_{j=1}^d \mathbb{1}\{\hat{\sigma}_j \neq \sigma_j\}$
- Define $\sigma^{\oplus j}$ to be the vector in $\{-1,1\}^d$ that differs from σ in only the j-th position
- Assouad's Lemma:

$$\min_f \max_{\sigma \in \{-1,1\}^d} \mathbb{E}_\sigma[\ell(f(Z), \sigma)] \geq \frac{d}{2} - \frac{1}{2} \sum_{j=1}^d \max_{\sigma \in \{-1,1\}^d} \|P_\sigma - P_{\sigma^{\oplus j}}\|_{\text{TV}}$$

- Proof: Define the mixture distributions

$$M_{+j} = \frac{1}{2^{d-1}} \sum_{\sigma \in \{-1,1\}^d : \sigma_j = +1} P_\sigma$$

$$M_{-j} = \frac{1}{2^{d-1}} \sum_{\sigma \in \{-1,1\}^d : \sigma_j = -1} P_\sigma$$

  - First we bound $\|M_{+j} - M_{-j}\|_{\mathrm{TV}}$:

$$\begin{aligned} \|M_{+j} - M_{-j}\|_{\mathrm{TV}} &= \left\| \frac{1}{2^{d-1}} \sum_{\sigma \in \{-1,1\}^d : \sigma_j = +1} P_\sigma - \frac{1}{2^{d-1}} \sum_{\sigma \in \{-1,1\}^d : \sigma_j = -1} P_\sigma \right\|_{\mathrm{TV}} \\ &= \frac{1}{2^{d-1}} \left\| \sum_{\sigma \in \{-1,1\}^d : \sigma_j = +1} (P_\sigma - P_{\sigma \oplus j}) \right\|_{\mathrm{TV}} \\ &\leq \frac{1}{2^{d-1}} \sum_{\sigma \in \{-1,1\}^d : \sigma_j = +1} \|P_\sigma - P_{\sigma \oplus j}\|_{\mathrm{TV}} \\ &\leq \max_{\sigma \in \{-1,1\}^d} \|P_\sigma - P_{\sigma \oplus j}\|_{\mathrm{TV}} \end{aligned}$$

  - Next, just like in the proof of Le Cam's lemma, we consider $\sigma \sim \mathrm{unif}\{-1,1\}^d$ and $Z \mid \sigma \sim P_\sigma$. For an estimator f, let $\hat{\sigma} = f(Z)$:

$$\mathbb{E}\, \ell(\hat{\sigma}, \sigma) = \sum_{j=1}^{d} \Pr[\hat{\sigma}_j \neq \sigma_j]$$

  - Now observe that $\Pr[\hat{\sigma}_j \neq \sigma_j]$ is the expected zero-one loss of an estimator f that just has to guess σ_j based on Z, where the two possible distributions are M_{+j} and M_{-j}
  - By Le Cam's lemma and our first step,

$$\begin{aligned} \Pr[\hat{\sigma}_j \neq \sigma_j] &\geq \frac{1}{2}(1 - \|M_{+j} - M_{-j}\|_{\mathrm{TV}}) \\ &\geq \frac{1}{2}\left(1 - \max_{\sigma \in \{-1,1\}^d} \|P_\sigma - P_{\sigma \oplus j}\|_{\mathrm{TV}}\right) \end{aligned}$$

    Plugging back into the previous displayed equation gives the claim.

5 Using Assouad's method

- Let $\mathcal{H}$ have VC dim d, and let $S = \{s_1, \ldots, s_d\}$ be shattered by $\mathcal{H}$
- Define distribution P_σ for data $Z = (X_i, Y_i)_{i=1}^d$
  - Under P_σ: $X_i = s_j$ with probability $1/d$; and given $X_i = s_j$,

$$Y_i = \begin{cases} +\sigma_j & \text{with probability } \frac{1+\varepsilon}{2}, \\ -\sigma_j & \text{with probability } \frac{1-\varepsilon}{2}. \end{cases}$$

  - Let $h_\sigma \in \mathcal{H}$ be a hypothesis satisfying $h_\sigma(s_j) = \sigma_j$ for all $j \in [d]$ (existence of h_σ is guaranteed since S is shattered by $\mathcal{H}$)
  - Note that $\min_{h\in\mathcal{H}} \operatorname{err}_\sigma(h) = \operatorname{err}_\sigma(h_\sigma) = \frac{1-\varepsilon}{2}$

- For any h,

$$\begin{aligned}\operatorname{err}_\sigma(h) &= \frac{1-\varepsilon}{2} + \frac{\varepsilon}{2}\sum_{j=1}^{d} \mathbb{1}\{h(s_j) \neq \sigma_j\} \\ &= \operatorname{err}_\sigma(h_\sigma) + \frac{\varepsilon}{2}\sum_{j=1}^{d} \mathbb{1}\{h(s_j) \neq \sigma_j\}\end{aligned}$$

- Suppose there is a learning algorithm A that can guarantee

$$\mathbb{E}[\operatorname{err}(A(Z)) - \min_{h\in\mathcal{H}} \operatorname{err}(h)] \leq \frac{\varepsilon}{4}$$

- Define $f_A \colon \mathcal{Z} \to \{-1,1\}^d$ by

$$f_A(Z) = (h(s_1), \ldots, h(s_d))$$

  where $h = A(Z)$

- This f_A satisfies

$$\mathbb{E}_\sigma[\ell(f_A(Z), \sigma)] \leq \frac{1}{2}$$

  for all $\sigma \in \{-1,1\}^d$

- By Assouad's lemma,

$$\max_{\sigma\in\{-1,1\}^d} \mathbb{E}_\sigma[\ell(f_A(Z), \sigma)] \geq \frac{d}{2} - \frac{1}{2}\sum_{j=1}^{d} \max_{\sigma\in\{-1,1\}^d} \|P_\sigma - P_{\sigma\oplus j}\|_{\mathrm{TV}}$$

- Therefore it must be that

$$\sum_{j=1}^{d} \max_{\sigma\in\{-1,1\}^d} \|P_\sigma - P_{\sigma\oplus j}\|_{\mathrm{TV}} \geq d-1 \tag{1}$$

- Pinsker's inequality:

$$\|P - Q\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2}\operatorname{RE}(P,Q)}$$

  Recall

$$\operatorname{RE}(P,Q) = \sum_{z\in\mathcal{Z}} P(z) \ln\frac{P(z)}{Q(z)}$$

- Let p_σ be the marginal distribution of (X_1, Y_1) under P_σ; since examples from P_σ are iid, we have

$$\operatorname{RE}(P_\sigma, P_{\sigma\oplus j}) = n \cdot \operatorname{RE}(p_\sigma, p_{\sigma\oplus j}) \tag{2}$$

- Note that p_σ and $p_{\sigma \oplus j}$ differ only in the probabilities assigned to (s_j, σ_j) and $(s_j, -\sigma_j)$
- Therefore

$$\begin{aligned} \mathrm{RE}(p_\sigma, p_{\sigma\oplus j}) &= p_\sigma(s_j,\sigma_j) \ln \frac{p_\sigma(s_j,\sigma_j)}{p_{\sigma\oplus j}(s_j,\sigma_j)} + p_\sigma(s_j,-\sigma_j) \ln \frac{p_\sigma(s_j,-\sigma_j)}{p_{\sigma\oplus j}(s_j,-\sigma_j)} \\ &= \frac{1}{d} \cdot \frac{1+\varepsilon}{2} \ln \frac{\frac{1}{d}\cdot\frac{1+\varepsilon}{2}}{\frac{1}{d}\cdot\frac{1-\varepsilon}{2}} + \frac{1}{d} \cdot \frac{1-\varepsilon}{2} \ln \frac{\frac{1}{d}\cdot\frac{1-\varepsilon}{2}}{\frac{1}{d}\cdot\frac{1+\varepsilon}{2}} \\ &= \frac{1}{d} \cdot \frac{1+\varepsilon}{2} \ln \frac{1+\varepsilon}{1-\varepsilon} + \frac{1}{d} \cdot \frac{1-\varepsilon}{2} \ln \frac{1-\varepsilon}{1+\varepsilon} \\ &= \frac{\varepsilon}{d} \ln \frac{1+\varepsilon}{1-\varepsilon} \end{aligned}$$

- Plugging back into (2),

$$\mathrm{RE}(P_\sigma, P_{\sigma\oplus j}) = \frac{n\varepsilon}{d} \ln \frac{1+\varepsilon}{1-\varepsilon} \approx \frac{2n\varepsilon^2}{d}$$

- Using Pinsker's, we get

$$\sum_{j=1}^{d} \max_{\sigma\in\{-1,1\}^d} \|P_\sigma - P_{\sigma\oplus j}\|_{\mathrm{TV}} \lesssim \sum_{j=1}^{d} \sqrt{\frac{n\varepsilon^2}{d}} = \sqrt{dn\varepsilon^2}.$$

- Combining with (1),

$$n \geq \frac{d-1}{\varepsilon^2} \cdot \left(1 - \frac{1}{d}\right).$$

6 Fano's mutual information method

- Assouad's method is useful if you have:
  - "separable" loss function (sum over coordinates), and
  - "hypercube" structure of difficult testing problems
- General method: Fano's (mutual information) method
- Mutual information between X and Y:

$$\begin{aligned} I(X;Y) &= \mathrm{RE}(P_{X,Y}, P_X \otimes P_Y) \quad \text{(i.e., it is symmetric w.r.t. } X \text{ and } Y) \\ &= H(Y) - H(Y \mid X) \\ &= H(X) - H(X \mid Y) \end{aligned}$$

  where

$$H(Y \mid X) = \sum_{x\in\mathcal{X}} P_X(x) H(Y \mid X = x) \quad \text{(conditional entropy)}$$

- Joint entropy:

$$\begin{aligned} H(X,Y) &= H(X) + H(Y \mid X) \\ &= H(Y) + H(X \mid Y) \end{aligned}$$

- Fano's inequality:
  - Consider a "prior" distribution π on Θ, and $\theta \sim \pi$.
  - Given θ, we draw data $Z \sim P_\theta$.
  - Guess of θ is $\hat{\theta}(Z)$.
  - If p_e is probability of guessing θ incorrectly, then

$$p_e \geq 1 - \frac{I(\theta; Z) + \ln 2}{H(\theta)}.$$

- Proof:
  - Communication scenario: Alice is given (θ, Z), but Bob is only given Z. What should Alice tell Bob so that he also knows θ?
  - Consider the following communication strategy for Alice:
    1. Compute estimator $\hat{\theta}(Z)$
    2. If $\hat{\theta}(Z) = \theta$, then send 0
    3. Else, send $(1, \theta)$
  - If this scenario is repeated independently many times, then average message length is

$$\leq 1 + p_e H(\theta)$$

  - Bob now has (θ, Z)! So he received $\geq H(\theta, Z)$ bits from Nature + Alice
  - Bob got $H(Z)$ bits from Nature, so needed $\geq H(\theta \mid Z)$ bits of information from Alice
  - Therefore

$$H(\theta \mid Z) \leq 1 + p_e H(\theta)$$

    which rearranges to

$$p_e \geq \frac{H(\theta \mid Z) - 1}{H(\theta)} = \frac{H(\theta) - I(\theta, Z) - 1}{H(\theta)} = 1 - \frac{I(\theta, Z) + 1}{H(\theta)}.$$

    (Can improve the $+1$ to $+\ln 2$ using more precise calculations...)

- Also useful: $I(\theta, Z) = \min_Q \mathrm{RE}(P_{Z|\theta}, Q \mid \theta)$ (whenever $P_{Z|\theta}/Q$ is valid)
  - Proof:

$$\begin{aligned} I(\theta, Z) &= \mathbb{E}\left[\ln \frac{P_{Z|\theta}}{P_Z}\right] \\ &= \mathbb{E}\left[\ln \frac{P_{Z|\theta}}{Q} - \ln \frac{P_Z}{Q}\right] \\ &= \mathbb{E}\left[\ln \frac{P_{Z|\theta}}{Q}\right] - \mathbb{E}\left[\ln \frac{P_Z}{Q}\right] \\ &= \mathrm{RE}(P_{Z|\theta}, Q \mid \theta) - \mathrm{RE}(P_Z, Q). \end{aligned}$$

- Pick Q in a strategic way to make it easy to upper-bound $\mathrm{RE}(P_{Z|\theta}, Q \mid \theta)$
- Suppose π is uniform on $\{\theta_1, \ldots, \theta_N\} \subset \Theta$, and

$$Q = \frac{1}{N}\sum_{j=1}^{N} P_{Z|\theta=\theta_j},$$

then by Jensen's inequality and convexity of RE in second argument,

$$\begin{aligned}
\mathrm{RE}\left(P_{Z|\theta}, \frac{1}{N}\sum_{j=1}^{N} P_{Z|\theta=\theta_j} \mid \theta\right) &\le \frac{1}{N}\sum_{j=1}^{N}\mathrm{RE}(P_{Z|\theta}, P_{Z|\theta=\theta_j} \mid \theta) \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{RE}(P_{Z|\theta=\theta_i}, P_{Z|\theta=\theta_j}) \\
&\le \max_{i,j}\mathrm{RE}(P_{Z|\theta=\theta_i}, P_{Z|\theta=\theta_j}) \\
&= \max_{i,j}\mathrm{RE}(P_{\theta_i}, P_{\theta_j})
\end{aligned}$$

- Han and Verdú's "generalized" Fano inequality: Let $\widetilde{\Theta} = \{\theta_1, \ldots, \theta_N\}$. Then

$$p_e \ge 1 - \frac{\max_{i,j}\mathrm{RE}(P_{\theta_i}, P_{\theta_j}) + \ln 2}{\ln N}$$

7 Using Fano's mutual information method

- Recipe to use the "generalized" Fano inequality:
  1. Find $\widetilde{\Theta} \subset \Theta$ such that for every distinct pair $\theta, \theta' \in \widetilde{\Theta}$, every estimator $\hat{\theta}$ has ϵ loss with respect to at least one of θ and θ'
     - If $\ell(\cdot,\cdot)$ is a distance function, then it suffices to make all pairwise distances in $\widetilde{\Theta}$ at least 2ϵ
  2. Prove upper-bound on $\mathrm{RE}(P_\theta, P_{\theta'})$ for $\theta, \theta' \in \widetilde{\Theta}$.
     This typically requires understanding details of the data distribution.
- The two steps in the recipe seem to be conflicting, but really it just means that one needs to carefully balance the two concerns when choosing $\widetilde{\Theta}$

7.1 Covering and packing

- Let (T, ℓ) be a (pseudo)metric space
  - T is a set of points
  - $\ell \colon T \times T \to \mathbb{R}_+$ is a symmetric function that satisfies $\ell(t,t) = 0$ for all $t \in T$ and $\ell(s,t) \le \ell(s,u) + \ell(u,t)$ for all $s,t,u \in T$ (triangle inequality)
- Say $C \subseteq T$ is an ϵ-cover of (T,ℓ) if for all $t \in T$, there exists $\tilde{t} \in C$ such that $\ell(t,\tilde{t}) \le \epsilon$

- Balls of radius ϵ centered around all points in C will "cover" all of T

- Say $P \subseteq T$ is an ϵ-packing of (T, ℓ) if $\ell(s,t) > \epsilon$ for all distinct $s, t \in P$
- Let $\mathcal{N}(\epsilon, T, \ell)$ denote the size of the smallest ϵ-cover of (T, ℓ)
- Let $\mathcal{M}(\epsilon, T, \ell)$ denote the size of the largest ϵ-packing of (T, ℓ)
- If an ϵ-packing P is maximal (i.e., for all $t \in T \setminus P$, the set $P \cup \{t\}$ is not an ϵ-packing), then P is also an ϵ-cover
  - This is because if P weren't an ϵ-cover, then there is a point $t \in T$ not already in P that we could add to P, and still have an ϵ-packing
- This implies that $\mathcal{N}(\epsilon, T, \ell) \leq \mathcal{M}(\epsilon, T, \ell)$
- On the other hand, we have $\mathcal{M}(2\epsilon, T, \ell) \leq \mathcal{N}(\epsilon, T, \ell)$
  - Let C be an ϵ-cover with $|C| = \mathcal{N}(\epsilon, T, \ell)$
  - If $|S| > |C|$, then by Pigeonhole principle, there are two points $s, t \in S$ that are "covered" by the same point $\tilde{t} \in C$
  - So by triangle inequality, $\ell(s,t) \leq \ell(s,\tilde{t}) + \ell(\tilde{t},t) \leq 2\epsilon$
  - This means S is not a 2ϵ-packing
- Example: $\mathcal{N}(\epsilon, [0,1]^d, \ell_\infty) \leq (1/\epsilon)^d$
  - Let $C = \{0, \epsilon, 2\epsilon, \ldots, 1-\epsilon\}^d$, so $|C| = (1/\epsilon)^d$
  - This C is an ϵ-cover
- Example: $\mathcal{M}(\epsilon, B^d, \ell_2) \leq (1 + 2/\epsilon)^d$
  - Let P be an ϵ-packing of (B^d, ℓ_2)
  - Balls of radius $\epsilon/2$ centered around points in P are disjoint
    Let V_1 be the total volume of these balls:
    $$V_1 = |P|(\epsilon/2)^d v_d$$
    where v_d is the volume of B^d itself
  - All of these balls are contained in a larger ball of radius $1 + \epsilon/2$
    Let V_2 be the volume of this larger ball:
    $$V_2 = (1 + \epsilon/2)^d v_d$$
  - $V_1 \leq V_2$, which implies
    $$|P| \leq (1 + 2/\epsilon)^d$$
- Example: $\mathcal{N}(\epsilon, B^d, \ell_2) \geq (1/\epsilon)^d$
  - If C is ϵ-cover of (B^d, ℓ_2), then B^d is contained in the union of $|C|$ balls of radius ϵ
  - By union bound, the latter has volume at most $|C|\epsilon^d \operatorname{vol}(B^d)$
  - Therefore $|C| \geq (1/\epsilon)^d$

7.2 Gaussian mean estimation in ℓ_2

- Consider Gaussian mean estimation problem in ℓ_2:
  - $P_\mu = \mathrm{N}(\mu, I_d)^{\otimes n}$ for $\mu \in \mathbb{R}^d$
  - Loss is ℓ_2 distance
$$\ell_2(\mu', \mu) = \|\mu' - \mu\|_2$$
- Let $\widetilde{\Theta}$ be a finite set from $\mathbb{R}^d$
- Given an estimator $\hat{\mu} = \hat{\mu}(Z)$ of μ, construct $f \colon \mathcal{Z} \to \widetilde{\Theta}$ by
$$f(Z) = \arg\min_{\mu \in \widetilde{\Theta}} \|\hat{\mu} - \mu\|_2$$
- Suppose $\widetilde{\Theta}$ is an ϵ-packing of $(\mathbb{R}^d, \ell_2)$:
  - Suppose true parameter is $\mu \in \widetilde{\Theta}$
  - If $f(Z) \neq \mu$, then:
$$\begin{aligned} \epsilon &< \ell_2(f(Z), \mu) \\ &\leq \ell_2(f(Z), \hat{\mu}) + \ell_2(\hat{\mu}, \mu) \\ &\leq 2\ell_2(\hat{\mu}, \mu), \end{aligned}$$
    which implies
$$\ell_2(\hat{\mu}, \mu) > \epsilon/2$$
  - Hence
$$\mathbb{E}_\mu \, \ell_2(\hat{\mu}, \mu) \geq \frac{\epsilon}{2} \Pr_\mu[f(Z) \neq \mu]$$
- $\mathrm{RE}(P_\mu, P_{\mu'}) = n \, \mathrm{RE}(\mathrm{N}(\mu, I_d), \mathrm{N}(\mu', I_d)) = \frac{n}{2}\|\mu - \mu'\|_2^2 = \frac{n}{2}\ell_2(\mu, \mu')^2$
- Balanced choice: choose $\widetilde{\Theta}$ to be ϵ-packing of $((2\epsilon)B^d, \ell_2)$ of cardinality 2^d
  - Every $\mu, \mu' \in (2\epsilon)B^d$ has $\ell_2(\mu, \mu') \leq 4\epsilon$
  - So by "Generalized" Fano inequality, we get
$$\max_{\mu \in \widetilde{\Theta}} \Pr_\mu[f(Z) \neq \mu] \geq 1 - \frac{\frac{n}{2}(4\epsilon)^2 + \ln 2}{\ln 2^d} = 1 - \frac{8n\epsilon^2 + \ln 2}{d \ln 2}$$
    which is at least $1/2$ if $n \lesssim d/\epsilon^2$ and $d \gtrsim 1$
  - In this case, we get
$$\max_{\mu \in \widetilde{\Theta}} \mathbb{E}_\mu \, \ell_2(\hat{\mu}, \mu) > \frac{\epsilon}{4}$$

7.3 Gaussian mean estimation in ℓ_∞

- Same as before, but now loss is ℓ_∞ distance $\ell_\infty(\mu', \mu) = \|\mu' - \mu\|_\infty$
- Let $\widetilde{\Theta} = \{\epsilon e_1, \ldots, \epsilon e_d\}$, so it is an ϵ-packing of $(\mathbb{R}^d, \ell_\infty)$ of cardinality d
- Every $\mu, \mu' \in \widetilde{\Theta}$ has $\ell_2(\mu, \mu') \leq \sqrt{2}\epsilon$, so by "Generalized" Fano inequality,

$$\max_{\mu \in \widetilde{\Theta}} \Pr_\mu[f(Z) \neq \mu] \geq 1 - \frac{\frac{n}{2}(\sqrt{2}\epsilon)^2 + \ln 2}{\ln d} = 1 - \frac{n\epsilon^2 + \ln 2}{\ln d}$$

  which is at least $1/2$ if $n \lesssim (\log d)/\epsilon^2$ and $d \gtrsim 1$

References

Eric V Slud. Distribution inequalities for the binomial law. *The Annals of Probability*, pages 404–412, 1977.

Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pages 423–435. Springer, 1997.