CS861: Theoretical Foundations of Machine Learning

Chapter 1: Lower Bounds

Kirthevasan Kandasamy
UW-Madison

Introduction

Why study statistical lower bounds?

- Understand the *fundamental difficulty* of a learning (estimation) problem.
- Assess whether our learning algorithm is **optimal** or how far it is from the best possible performance.

Plan of attack.

- Ch 1: Develop core techniques. Apply them to simple parameter estimation problems.
- Ch 2–6: Apply these tools to classification, regression, density estimation, bandits, online learning, online convex optimization:
  1. Define the learning problem.
  2. Design an algorithm and establish an upper bound on risk/regret.
  3. Use techniques from this chapter to derive (matching) lower bounds.

Contents

Ch 1.1: Lower bounds for point estimation

Point estimation. Estimate a single parameter $\theta(P) \in \mathbb{R}$ of distribution P using data (e.g the mean of a distribution).

A point estimation problem has the following components:

1. A known family of distributions $\mathcal{P}$.
2. A dataset S of n points drawn i.i.d from an unknown distribution $P \in \mathcal{P}$.
3. A parameter of interest $\theta = \theta(P) \in \mathbb{R}$.
4. An estimator $\widehat{\theta} = \widehat{\theta}(S) \in \mathbb{R}$. An estimator is any function of the data.
5. A loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ to evaluate how well we have estimated θ.
6. The risk $R(P, \widehat{\theta}) = \mathbb{E}_{S \sim P^n}[\ell(\theta(P), \widehat{\theta}(S))]$.

We will overload notation: a) Parameter as a scalar $\theta \in \mathbb{R}$ or a function $\theta : \mathcal{P} \to \mathbb{R}$.
b) Estimate (a random variable) $\widehat{\theta} \in \mathbb{R}$ or an estimator (function) $\widehat{\theta}$: data $\to \mathbb{R}$.

Example 1: Normal mean estimation

In the normal mean estimation example in HW0:

1. $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2); \mu \in \mathbb{R}\}$, where σ^2 is known.
2. Data $S = \{X_1, \ldots, X_n\}$ where $X_i \sim P$.
3. Parameter $\theta(P) = \mathbb{E}_{X \sim P}[X]$.
5. Loss function $\ell(\theta_1, \theta_2) = (\theta_1 - \theta_2)^2$.
6. Risk $R(\theta, \widehat{\theta}) \triangleq R(\mathcal{N}(\theta, \sigma^2), \widehat{\theta}) = \mathbb{E}_{X_1^n \sim P^n}[(\theta(P) - \widehat{\theta}(S))^2]$.
4. In HW0, you saw two possible estimators:

$$\textit{(i)}\ \text{Sample mean: } \widehat{\theta}_1(S) = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \textit{(ii)}\ \widehat{\theta}_2(S) = \frac{\alpha}{n}\sum_{i=1}^{n} X_i.$$

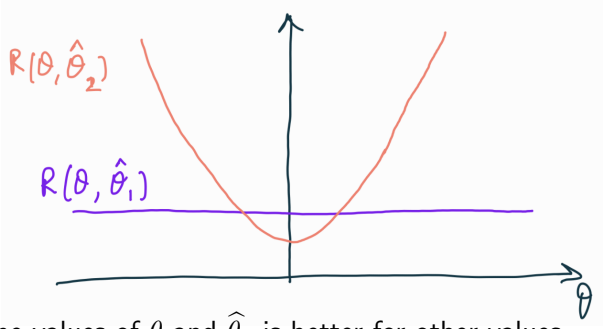
And showed

$$R(\theta, \widehat{\theta}_1) = \mathbb{E}_{X_1^n \sim P^n}[(\theta(P) - \widehat{\theta}_1(S))^2] = \frac{\sigma^2}{n}, \qquad R(\theta, \widehat{\theta}_2) = \theta^2(1-\alpha)^2 + \frac{\alpha^2\sigma^2}{n}.$$

Example 1: Normal mean estimation (cont'd)

$$\widehat{\theta}_1(S) = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad R(\theta, \widehat{\theta}_1) = \frac{\sigma^2}{n}. \qquad \widehat{\theta}_2(S) = \frac{\alpha}{n}\sum_{i=1}^{n} X_i, \quad R(\theta, \widehat{\theta}_2) = \theta^2(1-\alpha)^2 + \frac{\alpha^2\sigma^2}{n}.$$

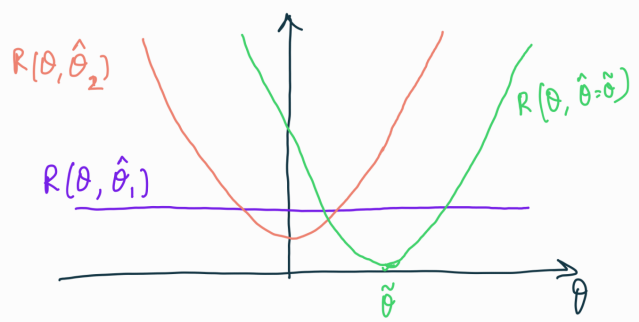
The following figure illustrates both risks as a function of θ.



$\widehat{\theta}_1$ is better for some values of θ and $\widehat{\theta}_2$ is better for other values.

Example 1: Normal mean estimation (cont'd)

Question: Can you design an estimator $\widehat{\theta}$ which minimizes $R(\theta, \widehat{\theta})$ for all θ?
That is, $R(\theta, \widehat{\theta}) \leq R(\theta, \widehat{\theta}')$ for all $\widehat{\theta}'$ and all $\theta \in \mathbb{R}$.
Ans: No. If we choose $\widehat{\theta}(S) = \widetilde{\theta}$ for some $\widetilde{\theta} \in \mathbb{R}$, it will do well when $P = \mathcal{N}(\widetilde{\theta}, \sigma^2)$.
No estimator can do better.



Optimal estimators

As 'pointwise' optimality is too strong, we usually resort to one of two versions of optimality.

1. *Minimax optimality:* $\widehat{\theta}$ minimizes the maximum risk over a class of distributions $\mathcal{P}$, i.e $\sup_{P \in \mathcal{P}} R(P, \widehat{\theta})$.
2. *Average risk optimality:* $\widehat{\theta}$ minimizes the average risk over a distribution of distributions Λ, *i.e.*, $\mathbb{E}_{P \sim \Lambda}[R(P, \widehat{\theta})]$.

Average (Bayesian) Risk Optimality

Average Risk. Introduce a probability Λ over $\mathcal{P}$ and define:

$$\overline{R}_\Lambda(\widehat{\theta}) \stackrel{\Delta}{=} \mathbb{E}_{P\sim\Lambda}[R(P,\widehat{\theta})] = \mathbb{E}_{P\sim\Lambda}\left[\mathbb{E}_{S\sim P^n}\left[\ell(\theta(P),\widehat{\theta}(S)) \,\middle|\, P\right]\right]$$

- In the Bayesian paradigm, Λ is called the **prior**.
- $\theta(P)$ is treated as a random variable, since $P \sim \Lambda$.
- An estimator $\widehat{\theta}_\Lambda$ minimizing $\overline{R}_\Lambda$ is the **Bayes' estimator**.
- The minimum $\overline{R}_\Lambda(\widehat{\theta}_\Lambda)$ is the **Bayes' risk**.

Finding the Bayes' Estimator

How do we find $\widehat{\theta}_\Lambda$? Let us write,

$$\overline{R}_\Lambda(\widehat{\theta}) = \mathbb{E}_P\left[\mathbb{E}_S\left[\ell(\theta(P), \widehat{\theta}(S)) \,|\, P\right]\right] = \mathbb{E}_S\Bigg[\underbrace{\mathbb{E}_P\left[\ell(\theta(P), \widehat{\theta}(S)) \,|\, S\right]}_{(*)}\Bigg]$$

If $\widehat{\theta}$ minimizes $(*) = \mathbb{E}_P[\ell(\theta(P), \widehat{\theta}(S))|S]$ for all S, then $\widehat{\theta}$ is the Bayes' estimator.

Bayes' Estimator under Squared Loss

Lemma. If $\ell(\theta_1, \theta_2) = (\theta_1 - \theta_2)^2$, then the Bayes' estimator is the posterior mean, *i.e.*, $\mathbb{E}_P[\theta(P)|S]$. Moreover, the Bayes' risk is $\mathbb{E}_S\left[\mathrm{Var}_P[\theta(P)|S]\right]$.

Proof. Let $\widehat{\theta}(S) = \mathbb{E}_P[\theta(P)|S]$. For any other estimator $\widehat{\theta}'$:

$$\begin{aligned}
\mathbb{E}_P[(\widehat{\theta}' - \theta)^2 \mid S] &= \mathbb{E}_P\left[(\widehat{\theta}' - \widehat{\theta} + \widehat{\theta} - \theta)^2 \mid S\right] \\
&= \mathbb{E}_P[(\widehat{\theta}' - \widehat{\theta})^2 \mid S] + \mathbb{E}_P[(\widehat{\theta} - \theta)^2 \mid S] + 2(\widehat{\theta}' - \widehat{\theta})\mathbb{E}_P[\widehat{\theta} - \theta \mid S] \\
&= \mathbb{E}_P[(\widehat{\theta}' - \widehat{\theta})^2 \mid S] + \mathbb{E}_P[(\widehat{\theta} - \theta)^2 \mid S] \\
&\geq \mathbb{E}_P[(\widehat{\theta} - \theta)^2 \mid S].
\end{aligned}$$

Thus, $\widehat{\theta}(S) = \mathbb{E}[\theta|S]$ minimizes the risk. The second statement follows from the observation that $\mathbb{E}[\theta|S] = \widehat{\theta}$ and hence $\mathbb{E}_P[(\widehat{\theta} - \theta)^2 \mid S]$ is the posterior variance $\mathrm{Var}_P[\theta|S]$.

Example: Normal–Normal Model in HW0

Setup. $S = \{X_1, \ldots, X_n\} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, σ^2 known. $\mu \sim \Lambda = \mathcal{N}(\nu, \tau^2)$.

Posterior. You will show, $\theta | S \sim \mathcal{N}(\widetilde{\nu}, \widetilde{\tau}^2)$, where

$$\widetilde{\nu} = \frac{\sigma^2/n}{\tau^2 + \sigma^2/n}\nu + \frac{\tau^2}{\tau^2 + \sigma^2/n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right), \quad \widetilde{\tau}^2 = \left(\frac{1}{\tau^2} + \frac{1}{\sigma^2/n}\right)^{-1}.$$

Bayes' Estimator and Bayes' Risk.

$$\widehat{\theta}_\Lambda(S) = \widetilde{\nu}, \qquad \overline{R}_\Lambda(\widehat{\theta}_\Lambda) = \mathbb{E}_S[\widetilde{\tau}^2] = \widetilde{\tau}^2 = \left(\frac{1}{\tau^2} + \frac{1}{\sigma^2/n}\right)^{-1}.$$

Example: Bernoulli-Beta Model in Ch0

Setup. $S = \{X_1, \ldots, X_n\} \overset{\text{i.i.d.}}{\sim} \text{Bern}(\mu)$. $\mu \sim \Lambda = \text{Beta}(a, b)$.

Posterior.

$$\theta | S \sim \text{Beta}(X + a,\ n - X + b), \qquad X = \sum_{i=1}^{n} X_i.$$

Bayes' Estimator.

$$\widehat{\theta}_\Lambda(S) = \mathbb{E}[\theta | S] = \frac{X + a}{n + a + b}.$$

Bayes' Risk.

$$\overline{R}_\Lambda(\widehat{\theta}_\Lambda) = \mathbb{E}_\mu \left[\mathbb{E}_{X \sim \text{Binomial}(n,\mu)} \left[\left(\frac{X + a}{n + a + b} - \theta \right)^2 \right] \right]$$

Minimax Optimality

Goal. We wish to find an estimator that minimizes the maximum risk over a class of distributions $\mathcal{P}$:

$$\sup_{P \in \mathcal{P}} R(P, \widehat{\theta}).$$

Minimax Risk.

$$R^\star = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} R(P, \widehat{\theta}) = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{S \sim P^n} \left[\ell(\theta(P), \widehat{\theta}(S)) \right].$$

- $R^\star$ depends on $\mathcal{P}, \ell, n, \ldots$.
- Sometimes we write $R_n^\star(\mathcal{P}, \ell)$ to make this dependence explicit.
- An estimator $\widehat{\theta}$ achieving $R^\star = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} R(P, \widehat{\theta})$ is called a **minimax estimator**.

Computing the Minimax Risk

Classical Approach: Find least favorable priors and corresponding Bayes' estimators with constant frequentist risk.

Our Recipe. We will instead use the following approach:

1. **Upper bound:** Design a "good" estimator $\widehat{\theta}$ and upper bound its risk:
$$R_n^\star \leq \sup_{P \in \mathcal{P}} R(P, \widehat{\theta}) \leq U_n.$$

2. **Lower bound:** Choose a prior Λ with $\textbf{supp}(\Lambda) \subseteq \mathcal{P}$ and lower bound the Bayes' risk by L_n. Therefore we have,
$$R_n^\star = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} R(P, \widehat{\theta}) \underbrace{\geq}_{\max \geq \text{avg}} \inf_{\widehat{\theta}} \mathbb{E}_{P \sim \Lambda}[R(P, \widehat{\theta})] \underbrace{=}_{\text{Bayes' estimator}} \mathbb{E}_{P \sim \Lambda}[R(P, \widehat{\theta}_\Lambda)] \geq L_n.$$

3. If $U_n = L_n$, we have the exact minimax risk, and $\widehat{\theta}$ is minimax optimal.

4. If $U_n \in \mathcal{O}(L_n)$, we have the minimax rate, and $\widehat{\theta}$ is rate-optimal.

Example 1: Gaussian Mean Estimation

Setup. $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}\}$, with known σ^2. Let $S = \{X_1, \ldots, X_n\}$ be i.i.d.

Claim. The sample mean $\widehat{\mu}(S) = \frac{1}{n}\sum_{i=1}^{n} X_i$ is minimax optimal for squared loss.

Upper Bound. From HW0, you know

$$\sup_{P \in \mathcal{P}} R(\mu, \widehat{\mu}) = \sup_{\mu \in \mathbb{R}} \mathbb{E}_{S \sim \mathcal{N}(\mu, \sigma^2)}[(\mu - \widehat{\mu}(S))^2] = \sup_{\mu \in \mathbb{R}} \frac{\sigma^2}{n} = \frac{\sigma^2}{n}.$$

Thus, $R^\star \leq \sigma^2/n$.

Example 1: Gaussian Mean Estimation (cont'd)

Recall, for the normal-normal model, the Bayes' risk is

$$\overline{R}_\Lambda(\widehat{\theta}_\Lambda) = \mathbb{E}_S[\widetilde{\tau}^2] = \widetilde{\tau}^2 = \left(\frac{1}{\tau^2} + \frac{1}{\sigma^2/n}\right)^{-1}.$$

Lower Bound. Take prior $\Lambda = \mathcal{N}(0, \tau^2)$. From previous example, we have,

$$R^\star \geq L_n = \frac{1}{\frac{1}{\tau^2} + \frac{1}{\sigma^2/n}}.$$

Taking $\sup_\tau$ gives $R^\star \geq \sigma^2/n$.

Therefore, the sample mean is (exactly) minimax-optimal.

Example 2: Bernoulli Mean Estimation

Setup. $\mathcal{P} = \{\mathrm{Bern}(\mu) : \mu \in [0,1]\}$. $S = \{X_1, \ldots, X_n\}$ i.i.d. from $P \in \mathcal{P}$.

Upper Bound. We will use the sample mean $\widehat{\mu}(S) = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then,

$$\begin{aligned}
\sup_{P \in \mathcal{P}} R(\mu, \widehat{\mu}) &= \sup_{\mu \in [0,1]} \mathbb{E}_{S \sim \mathrm{Bern}(\mu)}[(\mu - \widehat{\mu}(S))^2] \\
&= \sup_{\mu} \frac{\mathrm{Var}_\mu(X)}{n} = \sup_{\mu} \frac{\mu(1-\mu)}{n} \\
&= \frac{1}{4n} \triangleq U_n.
\end{aligned}$$

Thus, $R^\star \leq 1/(4n)$.

Example 2: Bernoulli Mean Estimation (cont'd)

Lower Bound. Use $\Lambda = \mathrm{Beta}(a, b)$. Then you can show (Try at home)

$$L_n = \overline{R}_\Lambda(\widehat{\theta}_\Lambda) = \mathbb{E}_\mu[\mu^2]((a+b)^2 - n) + \mathbb{E}_\mu[\mu](n - 2a(a+b)) + a^2.$$

Choosing $a = b = \sqrt{n}/2$ gives:

$$L_n = \frac{1}{4(\sqrt{n}+1)^2} = \frac{1}{4n + 8\sqrt{n} + 4}.$$

Clearly, $U_n > L_n$, but $U_n, L_n \in \Theta(1/n)$. Hence, $1/n$ is the minimax rate.

An exactly minimax-optimal estimator is:

$$\widehat{\theta}(S) = \frac{\sqrt{n}}{1+\sqrt{n}} \cdot \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{2(1+\sqrt{n})}.$$

Takeaways and Next Steps

Limitations of this approach. Need to design a prior and compute posterior. This is not easy for complex distributions.
E.g: Nonparametric regression: all Lipschitz functions in $[0,1]^d$.

Next. We will move to lower bounds beyond point estimation using *hypothesis testing*. This approach is more general.

Lessons Going Forward.

- Maximum risk $\geq$ average risk (key idea for lower bounds).
- Choose good priors, possibly depending on n.
- Still need to design good estimators.

Ch 1.2: Lower bounds for hypothesis testing

Plan for the remainder of the chapter.

- Lower bounds for hypothesis testing
  1. Le Cam's method for binary hypothesis tests
  2. Fano's method for multiple tests
- Reduction from learning (estimation) to testing

Hypothesis test

Hypothesis test. Let $\mathcal{P}$ be a class of distributions and let $\mathcal{P}_1, \ldots, \mathcal{P}_N$ be a partition of $\mathcal{P}$. Let S be a dataset drawn from some $P \in \mathcal{P}$. A (multiple) hypothesis test ψ is a function of the data which maps to $[N] = \{1, \ldots, N\}$. If $\psi(S) = j$, then the test has decided that $P \in \mathcal{P}_j$.

In this class, we will focus on cases where $\mathcal{P}_j = \{P_j\}$ is a singleton for all j.

Let $P_j(\psi \neq j) = \mathbb{P}_{S \sim P_j}(\psi(S) \neq j)$ is the probability that ψ does not correctly identify j when the data comes from P_j.
- As before, we will overload notation, and view ψ as both a function and a RV.

Lower bound. We wish to show that no test can simultaneously do well on all alternatives. Equivalently, any test will do poorly on at least one alternative:

$$\inf_{\psi} \max_{j \in [N]} P_j(\psi(S) \neq j) \geq \text{ Something large.}$$

Ch 1.2.1: Le Cam's method for binary hypothesis tests

Binary hypothesis test. A hypothesis test with just two alternatives P_0, P_1.

Neyman-Pearson test. Let data S come from either distribution P_0 or P_1, with densities p_0, p_1 respectively. The Neyman-Pearson test is a binary hypothesis test which chooses,

$$\psi_{\mathrm{NP}}(S) = \begin{cases} 0 & \text{if } p_0(S) \geq p_1(S), \\ 1 & \text{if } p_0(S) < p_1(S). \end{cases}$$

N.B. : When the dataset is an i.i.d sample, we should view p_0, p_1 as the product density.

Neyman Pearson Test (cont'd)

Theorem. The sum of errors is minimized by the Neyman-Pearson test. That is, for any other test ψ,

$$P_0(\psi \neq 0) + P_1(\psi \neq 1) \;\geq\; P_0(\psi_{\mathrm{NP}} \neq 0) + P_1(\psi_{\mathrm{NP}} \neq 1).$$

Proof. Write the LHS as,

$$\begin{aligned}
P_0(\psi = 1) + P_1(\psi = 0) &= \int_{\psi=1} p_0(x)\mathrm{d}x \;+\; \int_{\psi=0} p_1(x)\mathrm{d}x \\
&= \int_{\psi=1,\psi_{\mathrm{NP}}=1} p_0 \;+\; \int_{\psi=1,\psi_{\mathrm{NP}}=0} p_0 \;+\; \int_{\psi=0,\psi_{\mathrm{NP}}=1} p_1 \;+\; \int_{\psi=0,\psi_{\mathrm{NP}}=0} p_1 \\
&\geq \int_{\psi=1,\psi_{\mathrm{NP}}=1} p_0 \;+\; \int_{\psi=1,\psi_{\mathrm{NP}}=0} p_1 \;+\; \int_{\psi=0,\psi_{\mathrm{NP}}=1} p_0 \;+\; \int_{\psi=0,\psi_{\mathrm{NP}}=0} p_1 \\
&= \int_{\psi_{\mathrm{NP}}=1} p_0 \;+\; \int_{\psi_{\mathrm{NP}}=0} p_1 \;=\; P_0(\psi_{\mathrm{NP}} = 1) + P_1(\psi_{\mathrm{NP}} = 0)
\end{aligned}$$

Neyman Pearson Test (cont'd)

Corollary (Bretagnolle-Huber inequality). For any binary hypothesis test ψ,

$$P_0(\psi \neq 0) + P_1(\psi \neq 1) \;\geq\; \|P_0 \wedge P_1\| = 1 - \mathrm{TV}(P_0, P_1) \geq \frac{1}{2} e^{-\mathrm{KL}(P_0, P_1)}.$$

Recall from Ch0:

$$\mathrm{TV}(P, Q) = \frac{1}{2}\|P - Q\|_1 = 1 - \|P \wedge Q\|, \qquad \|P \wedge Q\| \geq \frac{1}{2} e^{-\mathrm{KL}(P, Q)}.$$

Proof. The first inequality follows from the NP test, and the observation,

$$P_0(\psi_{\mathrm{NP}} = 1) + P_1(\psi_{\mathrm{NP}} = 0) = \int_{p_0 < p_1} p_0 + \int_{p_1 < p_0} p_1 = \int \min(p_0, p_1) = \|P_0 \wedge P_1\|.$$

The remaining claims follow from the relations we proved about divergences.

Le Cam's method

LeCam's method: LeCam's method for binary hypothesis testing simply combines "max $\geq$ avg" with the BH inequality. We can summarize it as follows:

$$\begin{aligned}
\inf_{\psi} \max_{j\in\{0,1\}} P_j(\psi(S) \neq j) &\geq \frac{1}{2} \inf_{\psi} \left(P_0(\psi(S) \neq 0) + P_1(\psi(S) \neq 1)\right) && \texttt{max} \geq \texttt{avg} \\
&\geq \frac{1}{2} \left(P_0(\psi_{\mathrm{NP}}(S) \neq 0) + P_1(\psi_{\mathrm{NP}}(S) \neq 1\right) && \texttt{NP test} \\
&= \frac{1}{2} \|P_0 \wedge P_1\| \\
&\geq \frac{1}{4} e^{-\mathrm{KL}(P_0, P_1)}. && \texttt{Affinity-KL bound}
\end{aligned}$$

N.B. The KL version is the easiest to apply but you can also use TV/L1 or Chi-squared.

Example: normal vs normal

Suppose a dataset was drawn from $\mathcal{N}(\mu, \sigma^2)$ where σ^2 is known and $\mu \in \{0, \Delta\}$, where $\Delta > 0$. Consider the following hypothesis test[1],

$$\text{Choose } \psi(S) = 0 \text{ if } \quad \frac{1}{n}\sum_{i=1}^{n} X_i \leq \frac{\Delta}{2}, \qquad \text{else choose } \psi(S) = \Delta.$$

In HW0, you showed that with $\mathcal{O}\left(\sigma^2 \Delta^{-2} \log(1/\delta)\right)$ samples, this test can achieve

$$\mathbb{P}_0(\psi(S) = 0) \geq 1 - \delta, \qquad \mathbb{P}_\Delta(\psi(S) = \Delta) \geq 1 - \delta. \tag{1}$$

We will now show that $\Omega\left(\sigma^2 \Delta^{-2} \log(1/\delta)\right)$ samples are also necessary for any test that achieves (1).

[1] Try at home: Show that this is in fact the Neyman-Pearson test.

Example: normal vs normal (cont'd)

Recall:
$$\mathrm{KL}(\mathcal{N}(\mu_1,\sigma^2),\mathcal{N}(\mu_2,\sigma^2)) = \frac{(\mu_1-\mu_2)^2}{2\sigma^2}, \qquad \mathrm{KL}(P^n,Q^n) = n\mathrm{KL}(P,Q),$$

Le Cam's method: $\inf_{\psi} \max_{j\in[0,1]} P_j(\psi(S)\neq j) \geq \frac{1}{4}e^{-\mathrm{KL}(P_0,P_1)}$.

Let ψ be such that, with n samples we have $\mathbb{P}_0(\psi(S)=0)\geq 1-\delta$, and $\mathbb{P}_\Delta(\psi(S)=\Delta)\geq 1-\delta$. That is, $\max_{\mu\in\{0,\Delta\}}\mathbb{P}_\mu(\psi\neq\mu)\leq\delta$.
Hence, by Le Cam's method:

$$\begin{aligned}
\delta &\geq \max_{\mu\in\{0,\Delta\}} \mathbb{P}_\mu(\psi\neq\mu) \geq \inf_{\psi'}\max_{\mu\in\{0,\Delta\}} \mathbb{P}_\mu(\psi'\neq\mu) \\
&\geq \frac{1}{4}\exp\left(-\mathrm{KL}\left(\mathcal{N}(0,\sigma^2)^n,\mathcal{N}(\Delta,\sigma^2)^n\right)\right) && \texttt{Le Cam's} \\
&= \frac{1}{4}\exp\left(-n\cdot\frac{\Delta^2}{2\sigma^2}\right) && \texttt{KL properties}
\end{aligned}$$

Hence, $n\geq \frac{2\sigma^2}{\Delta^2}\log(1/(4\delta))$.

Ch 1.2.2: Fano's method for multiple hypothesis tests

Multiple hypothesis test. A hypothesis test with more than two alternatives P_0, P_1.

Goal. Show,

$$\inf_{\psi} \max_{j \in [N]} P_j(\psi(S) \neq j) \geq \text{ Something large.}$$

and this lower bound should grow with N.

Data Processing Inequality

Theorem. Let X, Y, Z be random variables such that $X \perp Z \mid Y$. Then,

$$I(X, Y) \geq I(X, Z) \quad \text{and hence} \quad H(X|Y) \leq H(X|Z).$$

Think of X, Y, Z as forming the Markov chain: $X \to Y \to Z$.

Intuition and Connections to Hypothesis Testing.

- We assume a prior over $\{P_1, \ldots, P_N\}$. Let $X \in [N]$ be the random variable selecting one.
- Data Y is generated from P_X.
- A test Z estimates X from Y.
- $I(X, Z) \leq I(X, Y)$ says the test contains no more information about X than Y, *i.e.*, you cannot *magically* learn more about X by processing information in Y.
- Similarly, $H(X|Y) \leq H(X|Z)$ says knowing the data Y reduces uncertainty about X at least as much as knowing only the outcome of the test Z.

Fano's Inequality

Fano's inequality. Let X be a discrete random variable with support $\mathcal{X}$. Let $X \to Y \to \widehat{X}$ form a Markov chain. Define:

$$p_e \triangleq \mathbb{P}(\widehat{X} \neq X), \qquad h(p_e) \triangleq -p_e \log(p_e) - (1 - p_e) \log(1 - p_e).$$

Then,

$$H(X|Y) \overset{(**)}{\leq} H(X|\widehat{X}) \overset{(*)}{\leq} p_e \log(|\mathcal{X}| - 1) + h(p_e).$$

Hence,

$$\mathbb{P}(\widehat{X} \neq X) \geq \frac{H(X|Y) - \log(2)}{\log(|\mathcal{X}|)}.$$

Connection to Hypothesis Testing.

- $X \in [1, \ldots, N]$ is a RV, Y is the data, $\widehat{X}$ is the test to identify X from Y.
- $p_e = \mathbb{P}(\widehat{X} \neq X)$ is the probability of error.
- Fano's inequality quantifies the relationship between p_e and $H(X|Y)$.
  e.g., If Y uniquely identifies X, then $H(X|Y) = 0$ and the lower bound is vacuous.

Proof of Fano's Inequality

Proof. Let $E = \mathbb{1}(\widehat{X} \neq X)$.

Use the chain rule for entropy in two ways:

$$H(E, X|\widehat{X}) = H(X|\widehat{X}) + H(E|X, \widehat{X}) = H(E|\widehat{X}) + H(X|E, \widehat{X}).$$

Now note:

- $H(E|X, \widehat{X}) = 0$ since $X, \widehat{X}$ determine E.
- $H(E|\widehat{X}) \leq H(E) = h(p_e)$ since conditioning reduces entropy.
- Next:

$$\begin{aligned} H(X|E, \widehat{X}) &= \mathbb{P}(E = 0)\, H(X|\widehat{X}, E = 0) + \mathbb{P}(E = 1)\, H(X|\widehat{X}, E = 1) \\ &= (1 - p_e) \cdot 0 + p_e \cdot H(X|\widehat{X}, E = 1) \\ &\leq p_e \log(|\mathcal{X}| - 1). \end{aligned}$$

Here, we used *(a)* if $E = 0$, then $X = \widehat{X}$, so $H(X|\widehat{X}, E = 0) = 0$, and *(b)* if $E = 1$, there are at most $|\mathcal{X}| - 1$ possible outcomes.

Proof of Fano's Inequality (cont'd)

Combining results:

$$H(X|\widehat{X}) \leq h(p_e) + p_e \log(|\mathcal{X}| - 1),$$

which is inequality $(*)$.

From the conditional entropy version of the data processing inequality:

$$H(X|Y) \leq H(X|\widehat{X}),$$

which is $(**)$.

Finally, since $h(p_e) = H(E) \leq \log(2)$, we get:

$$p_e \geq \frac{H(X|Y) - \log(2)}{\log(|\mathcal{X}|)}.$$

Fano's Method

Theorem (Fano's Method). Let S be drawn from some $P \in \{P_1, \ldots, P_N\} \subset \mathcal{P}$ and let ψ denote tests which map S to $[N]$. Then, the following statements hold:

1. **Global Fano method:** Denote $\overline{P} = \frac{1}{N}\sum_{i=1}^{N} P_i$. Then,

$$\inf_{\psi} \max_{j \in [N]} P_j(\psi(S) \neq j) \geq \left(1 - \frac{\frac{1}{N}\sum_{j=1}^{N} \mathrm{KL}(P_j, \overline{P}) + \log(2)}{\log(N)}\right).$$

2. **Local Fano method:**

$$\inf_{\psi} \max_{j \in [N]} P_j(\psi(S) \neq j) \geq \left(1 - \frac{\frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N} \mathrm{KL}(P_j, P_k) + \log(2)}{\log(N)}\right).$$

Global Fano is tighter, but harder to apply since computing $\mathrm{KL}(P_j, \overline{P})$ can be difficult. Local Fano is looser but easier to apply since it only requires pairwise KL divergences.

Proof of Fano's Method

Setup. Define the following data-generating process:

- Define a uniform prior over $\{P_1, \ldots, P_N\}$: $\mathbb{P}(V = j) = 1/N$.
- Given $V = j$, sample S from P_j.

The marginal distribution of S is $\overline{P}$, where for any set A,

$$\mathbb{P}(S \in A) = \sum_{j=1}^{N} \mathbb{P}(S \in A | V = j)\mathbb{P}(V = j) = \frac{1}{N} \sum_{j=1}^{N} P_j(A) = \overline{P}(A).$$

As max $\geq$ avg, and V induces a uniform prior on $[N]$, we have

$$\inf_{\psi} \max_{j \in [N]} P_j(\psi(S) \neq j) \geq \inf_{\psi} \mathbb{P}_{V,S}(\psi(S) \neq V).$$

Proof of Fano's Method (cont'd)

Fano's inequality: For $X \to Y \to \widehat{X}$, $\mathbb{P}(\widehat{X} \neq X) \geq \frac{H(X|Y)-\log(2)}{\log(|\mathcal{X}|)}$.

We want to show,

$$\max_{j\in[N]} P_j(\psi(S) \neq j) \geq \left(1 - \frac{\frac{1}{N}\sum_{j=1}^{N} \mathrm{KL}(P_j, \overline{P}) + \log(2)}{\log(N)}\right) \geq \left(1 - \frac{\frac{1}{N^2}\sum_{j,k=1}^{N} \mathrm{KL}(P_j, P_k) + \log(2)}{\log(N)}\right).$$

By Fano's inequality, for any test ψ:

$$\begin{aligned}
\mathbb{P}_{V,S}(\psi(S) \neq V) &\geq \frac{H(V|S) - \log(2)}{\log(N)} && \texttt{Fano's inequality} \\
&= \frac{H(V) - I(V,S) - \log(2)}{\log(N)} && \texttt{Using } I(X,Y) = H(X) - H(X|Y) \\
&= 1 - \frac{I(V,S) + \log(2)}{\log(N)}. && \texttt{As } H(V) = \log(N)
\end{aligned}$$

This gives:

$$\max_{j\in[N]} P_j(\psi(S) \neq j) \geq \left(1 - \frac{I(V,S) + \log(2)}{\log(N)}\right). \tag{2}$$

Proof of Fano's Method (cont.)

Bounding Mutual Information. Let p_j be the density of P_j, $\overline{p}$ be the density of $\overline{P}$, and p be the density of the joint distribution P of (V, S). We expand $I(S, V)$ as follows:

$$
\begin{aligned}
I(S, V) &= \mathbb{E}_{S,V}\left[\log\left(\frac{p(S, V)}{p(S)p(V)}\right)\right] \\
&= \sum_{j=1}^{N} \int_s \underbrace{p(S = s|V = j)}_{p_j(S)} \underbrace{\mathbb{P}(V = j)}_{1/N} \log\left(\frac{p_j(S) \cdot \frac{1}{N}}{\overline{p}(S) \cdot \frac{1}{N}}\right) \mathrm{d}s \\
&= \sum_{j=1}^{N} \int_s p_j(s) \log\left(\frac{p_j(s)}{\overline{p}(s)}\right) \mathrm{d}s \\
&= \frac{1}{N} \sum_{j=1}^{N} \mathrm{KL}(P_j, \overline{P})
\end{aligned}
$$

Proof of Fano's Method (cont.)

Jensen's inequality. For convex f, $f\big(\mathbb{E}[X]\big) \leq \mathbb{E}\big[f(X)\big]$.

By Jensen's inequality, and convexity of KL in the second argument, we have

$$\begin{aligned}\mathrm{KL}\left(P_j, \overline{P}\right) &= \mathbb{E}_{S\sim P_j}\left[\log\left(\frac{p_j(S)}{\frac{1}{N}\sum_{i=1}^N p_i(S)}\right)\right] \\ &\leq \mathbb{E}_{S\sim P_j}\left[\frac{1}{N}\sum_{i=1}^N \log\left(\frac{p_j(S)}{p_i(S)}\right)\right] = \frac{1}{N}\sum_{i=1}^N \mathrm{KL}(P_j, P_i).\end{aligned}$$

Therefore,

$$I(S,V) = \frac{1}{N}\sum_{j=1}^N \mathrm{KL}(P_j, \overline{P}) \leq \frac{1}{N^2}\sum_{j=1}^N\sum_{k=1}^N \mathrm{KL}(P_j, P_k)$$

Proof of Fano's Method (cont'd)

We have shown,

$$\max_{j\in[N]} P_j(\psi(S)\neq j) \geq \left(1-\frac{I(V,S)+\log(2)}{\log(N)}\right).$$

$$I(S,V) \overset{(a)}{=} \frac{1}{N}\sum_{j=1}^{N}\mathrm{KL}(P_j,\overline{P}) \overset{(b)}{\leq} \frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}\mathrm{KL}(P_j,P_k)$$

Putting it altogether we get,

$$\max_{j\in[N]} P_j(\psi(S)\neq j) \geq \left(1-\frac{\frac{1}{N}\sum_{j=1}^{N}\mathrm{KL}(P_j,\overline{P})+\log(2)}{\log(N)}\right) \qquad \texttt{By (a)} \rightarrow \texttt{Global Fano}$$

$$\geq \left(1-\frac{\frac{1}{N^2}\sum_{j,k=1}^{N}\mathrm{KL}(P_j,P_k)+\log(2)}{\log(N)}\right). \qquad \texttt{By (b)} \rightarrow \texttt{Local Fano}$$

Reduction from Learning to Testing

Estimation (learning) is a generalization of hypothesis testing.

A typical estimation (learning) problem. Given a class of distributions $\mathcal{P}$, data S is drawn from some $P \in \mathcal{P}$, identify[2] the distribution P.

A typical hypothesis testing problem. Data S is drawn from some $P \in \{P_1, \ldots, P_N\} \subset \mathcal{P}$. Identify the distribution P.

Hence, testing is easier than learning. From any learning algorithm, we can device a testing procedure as follows:
- Let $\widehat{P}$ be the distribution chosen by a learning algorithm.
- Choose the element in $\{P_1, \ldots, P_N\}$ that is "closest" to $\widehat{P}$.

Therefore a lower bound for testing $\implies$ a lower bound for learning.
- If we carefully design alternatives $\{P_1, \ldots, P_N\}$, we can in fact get tight lower bounds.

[2]Usually we may only be interested in learning a parameter of interest $\theta(P)$ instead of the entire distribution, but we will ignore this distinction for now.

A learning problem

- Let $\mathcal{P}$ be a *known* family of distributions.

- We observe data S drawn some *unknown* distribution $P \in \mathcal{P}$.

- An algorithm $\widehat{A}$ maps the data to an action space $\mathcal{A}$. Letting $\mathcal{D}$ denoting the data space, we can write $\widehat{A} : \mathcal{D} \to \mathcal{A}$.

- The learner incurs a loss $L(A, P)$ for choosing action A when the distribution is P, where $L : \mathcal{A} \times \mathcal{P} \to \mathbb{R}_+$.

- We will assume, for all $P \in \mathcal{P}$, we have $\inf_A L(A, P) = 0$. This is often w.l.o.g as we can always redefine, $L_{\mathcal{A}}(A, P) \triangleq L(A, P) - \inf_{A' \in \mathcal{A}} L(A', P)$.

- Define the risk of an algorithm as, $R(\widehat{A}, P) = \mathbb{E}_{S \sim P}\left[L(\widehat{A}(S), P)\right]$.

- The minimax risk: $R^\star(\mathcal{P}) = \inf_{\widehat{A}} \sup_{P \in \mathcal{P}} R(\widehat{A}, P)$.

If $\mathcal{P}$ is clear from context, we will simply write $R^\star$.
If there are n i.i.d data, we will write $R^\star_n$ to emphasize this.

Example 1: Normal mean estimation

- Let $\mathcal{P} = \{\mathcal{N}(\theta, \sigma^2); \theta \in \mathbb{R}\}$ where σ^2 is known a priori.

- We observe an i.i.d datset $S = \{X_1, \ldots, X_n\}$ from some unknown $\mathcal{N}(\theta, \sigma^2)$. Hence, dataspace $\mathcal{D} = \mathbb{R}^n$.

- We wish to estimate the mean, so the action space is $\mathbb{R}$.

- An (algorithm) estimator $\widehat{\mu} : \mathbb{R}^n \to \mathbb{R}$.

- The loss, $L(\mu', \mathcal{N}(\theta, \sigma^2)) = (\mu' - \theta)^2$. We have $\inf_{\mu' \in \mathbb{R}} L(\mu', \mathcal{N}(\theta, \sigma^2)) = 0$ for all $\theta \in \mathbb{R}$.

- Risk, $R(\widehat{\mu}, \mathcal{N}(\theta, \sigma^2)) = \mathbb{E}_{S \sim \mathcal{N}(\theta, \sigma^2)^n} \left[(\widehat{\mu}(S) - \theta)^2 \right]$.

- Minimax risk

$$R_n^\star = \inf_{\widehat{\mu}} \sup_{\theta \in \mathbb{R}} \mathbb{E}_{S \sim \mathcal{N}(\theta, \sigma^2)^n} \left[(\widehat{\mu}(S) - \theta)^2 \right]$$

Example 2: Mean estimation (more generally)

- Let $\mathcal{P}$ be a family of distributions such that $\textbf{supp}\,(P) \subset \mathbb{R}^d$ for all $P \in \mathcal{P}$.
- We observe an i.i.d datset $S = \{X_1, \ldots, X_n\}$ from some $P \in \mathcal{P}$.
- The action space is $\mathbb{R}^d$.
- An (algorithm) estimator $\widehat{\mu} : (\mathbb{R}^d)^n \to \mathbb{R}^d$.
- The loss, $L(\mu', P) = \|\mu' - \mu(P)\|_p^p$, where $p \geq 1$ and $\mu(P) = \mathbb{E}_{X \sim P}[X]$.

We have $\inf_{\mu' \in \mathbb{R}^d} L(\mu', P) = 0$.

- Risk, $R(\widehat{\mu}, P) = \mathbb{E}_{S \sim P^n}\left[\|\widehat{\mu}(S) - \mu(P)\|_p^p\right]$.
- Minimax risk

$$R_n^\star = \inf_{\widehat{\mu}} \sup_{P \in \mathcal{P}} \mathbb{E}_{S \sim P^n}\left[\|\widehat{\mu}(S) - \mu(P)\|_p^p\right]$$

Example 3: Regression in L_2 norm

- Let $\mathcal{X}$ be an input space.
- Let $\mathcal{P}$ be a family of distributions with $\textbf{supp}\,(P) \subset \mathcal{X} \times \mathbb{R}$ for all $P \in \mathcal{P}$.
- We observe an i.i.d datset $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ from some $P \in \mathcal{P}$.
- An algorithm will, based on the data, produce a function to predict Y from X. Hence, the action space is $\mathbb{R}^{\mathcal{X}} = \{g : \mathcal{X} \to \mathbb{R}\}$.
- An (algorithm) estimator $\widehat{f} : (\mathcal{X} \times \mathbb{R})^n \to \mathbb{R}^{\mathcal{X}}$.
- The loss, $L(f', P) = \|f' - f(P)\|_2^2$, where $f(P)$ is the *regression function*, *i.e.*, $f(P)(\cdot) = \mathbb{E}_P[Y|X = \cdot]$. Here, $\|f' - f(P)\|_2^2 = \int_{\mathcal{X}} (f'(x) - f(P)(x))^2 \mathrm{d}x$. We have $\inf_{f' \in \mathbb{R}^{\mathcal{X}}} L(f', P) = 0$.
- Risk, $R(\widehat{f}, P) = \mathbb{E}_{S \sim P^n}\left[\|\widehat{f}(S) - f(P)\|_2^2\right]$.
- Minimax risk

$$R_n^\star = \inf_{\widehat{\mu}} \sup_{P \in \mathcal{P}} \mathbb{E}_{S \sim P^n}\left[\|\widehat{f}(S) - f(P)\|_2^2\right]$$

Learning algorithms (cont'd)

The previous examples are instances of *parameter estimation*.
- You are estimating a parameter (property) θ of the distribution.
- E.g mean in Examples 1 and 2, regression function in Example 3.

Learning problems are not always formulated in this form.

Example 4: Regression, excess risk in a hypothesis class

- Let $\mathcal{X}$ be an input space. Let $\mathcal{P}$ be a family of distributions with $\textbf{supp}(P) \subset \mathcal{X} \times \mathbb{R}$ for all $P \in \mathcal{P}$.

- We observe an i.i.d datset $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ from some $P \in \mathcal{P}$.

- Let $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ be a hypothesis space.
- An algorithm $\widehat{h}$ will, based on the data, choose some $h \in \mathcal{H}$ as the predictor.
That is, the action space is $\mathcal{H}$ and $\widehat{h} : (\mathcal{X} \times \mathbb{R})^n \to \mathcal{H}$.

- Define the *instance loss* as $\ell(h, (X, Y)) = (h(X) - Y)^2$.

- The *population loss* is $L(h, P) = \mathbb{E}_{X,Y \sim P}[\ell(h, (X, Y))] = \mathbb{E}_{X,Y \sim P}[(h(X) - Y)^2]$.
Here, we do *not* have $\inf_{h \in \mathcal{H}} L(h, P) = 0$.
Therefore, define the *excess population loss* $L_{\mathcal{H}}(h, P) = L(h, P) - \inf_{h' \in \mathcal{H}} L(h', P)$

- Define the *excess risk*,
$R(\widehat{h}, P) = \mathbb{E}_{S \sim P^n}\left[L_{\mathcal{H}}(\widehat{h}(S), P)\right] = \mathbb{E}_{S \sim P^n}\left[L(\widehat{h}(S), P)\right] - \inf_{h' \in \mathcal{H}} L(h', P)$.

- The minimax risk $R_n^\star = \inf_{\widehat{h}} \sup_{P \in \mathcal{P}} R(\widehat{h}, P)$.

Example 5: Classification, excess risk in a hypothesis class

- Let $\mathcal{X}$ be an input space. Let $\mathcal{P}$ be a family of distributions with $\textbf{supp}(P) \subset \mathcal{X} \times \{0,1\}$ for all $P \in \mathcal{P}$.
- We observe an i.i.d datset $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ from some $P \in \mathcal{P}$.
- Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a hypothesis space.
- An algorithm $\widehat{h}$ will, based on the data, choose some $h \in \mathcal{H}$ as the predictor.
- Define the *instance loss* as $\ell(h, (X, Y)) = \mathbb{1}(h(X) \neq Y)$.
- The *population loss* is $L(h, P) = \mathbb{E}_{X,Y\sim P}[\ell(h, (X, Y))] = \mathbb{P}_{X,Y\sim P}(h(X) \neq Y)$.

Define the excess population loss $L_{\mathcal{H}}(h, P) = L(h, P) - \inf_{h' \in \mathcal{H}} L(h', P)$

- Define the *excess risk* and minimax risk similar to Example 4.

Reduction to testing

Separation between distributions. In a given learning problem with loss L and action space $\mathcal{A}$, we define the separation $\Delta(P, Q)$ between two distributions P, Q as follows:

$$\Delta(P,Q) = \sup\Big\{\delta \geq 0; \quad L(A,P) \leq \delta \implies L(A,Q) \geq \delta, \quad \forall\, A \in \mathcal{A}, \\ L(A,Q) \leq \delta \implies L(A,P) \geq \delta, \quad \forall\, A \in \mathcal{A},\Big\}$$

A set of distributions $\{P_1, \ldots, P_N\}$ are δ-separated if $\Delta(P_j, P_k) \geq \delta$ for all $j \neq k$.

Theorem (Reduction to testing). Let S be drawn from some distribution $P \in \mathcal{P}$. Let $\{P_1, \ldots, P_N\}$ be a δ-separated subset of $\mathcal{P}$. Let ψ be any test which maps the dataset to $[N]$. Then,

$$R^\star(\mathcal{P}) \geq \delta \cdot \inf_{\psi} \max_{j \in [N]} \mathbb{P}_{S \sim P_j}(\psi(S) \neq j).$$

Reduction to testing (cont'd)

This theorem gives a lower bound on the minimax risk.

- Intuition: if you cannot distinguish between N alternatives, then your estimation error also has to be large.
- We can leverage tools for proving lower bounds for hypothesis testing to now prove lower bounds for estimation.

How tight a lower bound we get depends on how well we choose our alternatives:

- If N is too large, then δ may be small and the lower bound will be small.
- If N is too small, then δ may be large, but the probability of making a mistake $\mathbb{P}_{S \sim P_j}(\psi(S) \neq j)$ will be small.

Proof of RTT Theorem

Proof of RTT theorem. For brevity, denote $\mathbb{P}_j(\cdot) = \mathbb{P}_{S\sim P_j}(\cdot)$, and $\mathbb{E}_j[\cdot] = \mathbb{E}_{S\sim P_j}[\cdot]$. As $L(A, P)$ is non-negative, we can lower bound $R^\star$ using Markov's inequality:

$$R^\star = \inf_{\widehat{A}} \sup_{P\in\mathcal{P}} \mathbb{E}_P\big[L(\widehat{A}(S), P)\big] \geq \inf_{\widehat{A}} \max_{j\in[N]} \mathbb{E}_j\big[L(\widehat{A}(S), P_j)\big]$$

$$\geq \delta \cdot \inf_{\widehat{A}} \max_{j\in[N]} \mathbb{P}_j(L(\widehat{A}(S), P_j) \geq \delta) \qquad \texttt{Markov's, } P(Z > a) \leq \frac{\mathbb{E}[Z]}{a} \texttt{ if } Z \geq 0$$

Claim: Let $\psi_{\widehat{A}}$ be the test, where $\psi_{\widehat{A}}(S) = \operatorname{argmin}_{j\in[N]} L(\widehat{A}(S), P_j)$. Suppose $S \sim P_j$. If $\psi_{\widehat{A}}(S) \neq j$, then, $L(\widehat{A}(S), P_j) \geq \delta$.

Then, by this claim we have,

$$R^\star \geq \delta \cdot \inf_{\widehat{A}} \max_{j\in[N]} \mathbb{P}_j(L(\widehat{A}(S), P_j) \geq \delta) \geq \delta \cdot \inf_{\widehat{A}} \max_{j\in[N]} \mathbb{P}_j(\psi_{\widehat{A}}(S) \neq j)$$

$$\geq \delta \cdot \inf_{\psi} \max_{j\in[N]} \mathbb{P}_j(\psi(S) \neq j)$$

Proof of RTT Theorem (cont'd)

Recall:

$$\Delta(P, Q) = \sup \Big\{ \delta \geq 0; \quad L(A, P) \leq \delta \implies L(A, Q) \geq \delta, \quad \forall\, A \in \mathcal{A}, \\ L(A, Q) \leq \delta \implies L(A, P) \geq \delta, \quad \forall\, A \in \mathcal{A}, \Big\}$$

We will now prove the claim.

Claim: Let $\psi_{\widehat{A}}$ be the test, where $\psi_{\widehat{A}}(S) = \operatorname{argmin}_{j \in [N]} L(\widehat{A}(S), P_j)$. Suppose $S \sim P_j$. If $\psi_{\widehat{A}}(S) \neq j$, then, $L(\widehat{A}(S), P_j) \geq \delta$.

Proof. Let $\psi_{\widehat{A}}(S) = k \neq j$. First suppose, $L(\widehat{A}(S), P_k) \leq \delta$. Then, $L(\widehat{A}(S), P_j) \geq \delta$ by the definition of δ-separation.

If $L(\widehat{A}(S), P_k) \geq \delta$, then as $k = \psi_{\widehat{A}}(S)$ has the smallest loss among all alternatives, we have $L(\widehat{A}(S), P_j) \geq L(\widehat{A}(S), P_k) \geq \delta$.

Le Cam's method for learning from i.i.d data

Le Cam's method. Let S be an i.i.d dataset of n points drawn from some $P \in \mathcal{P}$. Let $\{P_0, P_1\} \subset \mathcal{P}$ such that $\Delta(P_0, P_1) \geq \delta$ and $\mathrm{KL}(P_0, P_1) \leq \log(2)/n$. Then, $R_n^\star \geq \frac{\delta}{8}$.

Intuition. For tight lower bounds, we should choose P_0, P_1 to be well-separated in the loss (large $\Delta(P_0, P_1)$). But, they should be statistically indistinguishable (small KL).

Recall, Le Cam's method: $\inf_{\psi} \max_{j \in [0,1]} P_j(\psi(S) \neq j) \geq \frac{1}{4} e^{-\mathrm{KL}(P_0, P_1)}$.

Proof. As the data is i.i.d, using the properties of KL, we have

$$e^{-\mathrm{KL}(P_0^n, P_1^n)} = e^{-n\mathrm{KL}(P_0, P_1)} \geq \frac{1}{2}.$$

Therefore, by RTT

$$R_n^\star \geq \delta \cdot \inf_{\psi} \max_{j \in \{0,1\}} \mathbb{P}_j(\psi(S) \neq j) \geq \frac{\delta}{8}.$$

Fano's method for learning from i.i.d data

Local Fano method. Let S be an i.i.d dataset from some distribution $P \in \mathcal{P}$. Let $\{P_1, \ldots, P_N\} \subset \mathcal{P}$ such that $\Delta(P_j, P_k) \geq \delta$ and $\mathrm{KL}(P_j, P_k) \leq \frac{\log(N)}{4n}$ for all $j \neq k$. Suppose $N \geq 16$. Then, $R_n^\star \geq \frac{\delta}{2}$.

Recall, Local Fano method: $\inf_{\psi} \max_{j \in [N]} P_j(\psi(S) \neq j) \geq \left(1 - \frac{\frac{1}{N^2}\sum_{j,k=1}^{N} \mathrm{KL}(P_j, P_k) + \log(2)}{\log(N)}\right)$.

Proof. First note that by the KL property for i.i.d data and the given condition, $\mathrm{KL}(P_j^n, P_k^n) = n\mathrm{KL}(P_j, P_k) \leq \frac{\log(N)}{4}$. Therefore,

$$R^\star \geq \delta \cdot \inf_{\psi} \max_{j \in [N]} \mathbb{P}_{S \sim P_j}(\psi(S) \neq j) \geq \delta \cdot \left(1 - \frac{\frac{1}{N^2}\sum_{j,k} \log(N)/4 + \log(2)}{\log(N)}\right)$$

$$\overset{(a)}{\geq} \delta \cdot \left(1 - \frac{1}{4} - \frac{\log(2)}{\log(16)}\right) = \frac{\delta}{2}.$$

Here, (a) uses the fact that $N \geq 16$.

A Corollary of RTT for parameter estimation problems

Let Θ be a parameter space, and $\theta(P)$ be a parameter of the distribution, *i.e.*, $\theta : \mathcal{P} \to \Theta$. Suppose the action space is $\mathcal{A} = \Theta$, and the loss takes the form,

$$L(\theta', P) = \Phi \circ \rho(\theta', \theta(P)), \qquad \forall\, \theta' \in \Theta.$$

Here, $\rho : \Theta \times \Theta \to \mathbb{R}_+$ is a pseudo-metric and $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ is a non-decreasing function.

E.g.: *In Example 1:* $\Phi(t) = t^2$, $\rho(\theta_1, \theta_2) = |\theta_1 - \theta_2|$. *In Example 2:* $\Phi(t) = t^p$, $\rho(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|_p$. *In Example 3:* $\Phi(t) = t^2$, $\rho(f_1, f_2) = \|f_1 - f_2\|_2$.

Corollary of RTT for parameter estimation. Let $\{P_1, \ldots, P_N\} \subset \mathcal{P}$ and let $\delta = \min_{j \neq k} \rho(\theta(P_j), \theta(P_k))$. Let $\widehat{\theta}$ denote an estimator for θ. Then,

$$R^\star = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{S \sim P} \left[\Phi \circ \rho(\widehat{\theta}(S), \theta(P)) \right] \;\geq\; \Phi\left(\frac{\delta}{2}\right) \inf_{\psi} \max_{j \in [N]} \mathbb{P}_{S \sim P_j} \left(\psi(S) \neq j \right).$$

Proof of corollary

Recall: $\Delta(P, Q) = \sup \Big\{ \delta' \geq 0; \quad L(A, P) \leq \delta' \implies L(A, Q) \geq \delta', \quad \forall\, A \in \mathcal{A}, \\ L(A, Q) \leq \delta' \implies L(A, P) \geq \delta', \quad \forall\, A \in \mathcal{A}, \Big\}$

RTT: If $\Delta(P_j, P_k) \geq \delta'$ for all $P_1, \ldots, P_N$, then $R^\star \geq \delta' \inf_\psi \max_j \mathbb{P}_j \left(\psi(S) \neq j\right)$.

Proof. For simplicity, we will prove the corollary for strictly increasing Φ. Suppose $L(\theta', P_j) = \Phi \circ \rho(\theta', \theta(P_j)) \leq \Phi(\delta/2)$ for some $\theta' \in \Theta$. It is sufficient to show that $P_1, \ldots, P_N$ are $\Phi(\delta/2)$-separated in the loss L. We have,

$$\rho(\theta', \theta(P_j)) \leq \frac{\delta}{2}$$

$$\implies \quad \rho(\theta', \theta(P_k)) \geq \frac{\delta}{2} \text{ for all } k \neq j \qquad \texttt{As } \{\theta(P_i)\}_{i \in [N]} \texttt{ is a } \delta\texttt{-packing of } \Theta$$

$$\implies \quad L(\theta', P_k) = \Phi \circ \rho(\theta', \theta(P_k)) \geq \Phi\left(\frac{\delta}{2}\right).$$

Therefore, the distributions $P_1, \ldots, P_N$ are $\Phi(\delta/2)$-separated in the loss L.

Le Cam and Fano methods for parameter estimation

Le Cam's method for parameter estimation. Let S be an i.i.d dataset of n points drawn from some $P \in \mathcal{P}$. Let $P_0, P_1 \in \mathcal{P}$. Let $\rho(\theta(P_0), \theta(P_1)) \geq \delta$. If $\mathrm{KL}(P_0, P_1) \leq \frac{1}{n}\log(2)$, then

$$R_n^\star(\mathcal{P}) \geq \frac{1}{8}\Phi\left(\frac{\delta}{2}\right).$$

Local Fano method for parameter estimation. Let S be an i.i.d dataset from some distribution $P \in \mathcal{P}$. Let $\{P_1, \ldots, P_N\} \subset \mathcal{P}$ such that $N \geq 16$, $\rho(\theta(P_j), \theta(P_k)) \geq \delta$, and $\mathrm{KL}(P_j, P_k) \leq \log(N)/4n$ for all $j \neq k$. Then,

$$R_n^\star \geq \frac{1}{2}\Phi\left(\frac{\delta}{2}\right).$$

(You can verify these statements at home.)

Example 1: Normal mean estimation

Let $S = \{X_1, \ldots, X_n\}$ be drawn i.i.d from some $P \in \mathcal{P}$, where $\mathcal{P} = \{\mathcal{N}(\mu, s^2); \mu \in \mathbb{R}, s^2 \leq \sigma^2\}$ with σ^2 known. We wish to estimate the mean $\theta(P) = \mathbb{E}_{X \sim P}[X]$. Let the loss be $\Phi \circ \rho(\theta_1, \theta_2) = (\theta_1 - \theta_2)^2$.

First, we will choose $P_0 = \mathcal{N}(0, \sigma^2)$ and $P_1 = \mathcal{N}(\delta, \sigma^2)$.
We have separation $\rho(\theta(P_0), \theta(P_1)) = |\theta(P_0) - \theta(P_1)| = \delta$.

We also have, $\mathrm{KL}(P_0, P_1) = \frac{\delta^2}{2\sigma^2}$ (recall $\mathrm{KL}(\mathcal{N}(\mu_1, \sigma^2), \mathcal{N}(\mu_2, \sigma^2)) = (\mu_1 - \mu_2)^2/(2\sigma^2)$).

We need, $\mathrm{KL}(P_0, P_1) \leq \frac{1}{n}\log(2)$, so choose $\delta = \sigma\sqrt{\frac{2\log(2)}{n}}$.

Then,

$$R_n^\star \geq \frac{1}{8}\Phi\left(\frac{\delta}{2}\right) = \frac{1}{8}\frac{\delta^2}{4} = \frac{\log(2)}{16} \cdot \frac{\sigma^2}{n}.$$

The sample mean achieves risk $\mathrm{Var}_P(X)/n \leq \sigma^2/n$, and hence σ^2/n is the minimax rate.

Example 2: Mean estimation in a bounded domain

Let $\{X_1, \ldots, X_n\}$ be drawn i.i.d from some $P \in \mathcal{P}$, where $\mathcal{P} = \{P; \textbf{supp}\,(P) \subset [0,1]\}$ contains all distributions in $[0,1]$. Let the loss be $\Phi \circ \rho(\theta_1, \theta_2) = (\theta_1 - \theta_2)^2$.

Lower bound. Choose $P_0 = \mathrm{Bern}\,(1/2 + \delta)$ and $P_1 = \mathrm{Bern}\,(1/2)$. Therefore, separation is δ. Using the $\mathrm{KL} \leq \chi^2$ inequality, we have, $\mathrm{KL}(P_0, P_1) \leq \frac{(\mu_0 - \mu_1)^2}{\mu_1(1-\mu_1)} = 4\delta^2$.

We want $\mathrm{KL}(P_0, P_1) \leq \frac{\log(2)}{n}$, which is satisfied if we choose $\delta = \frac{1}{2}\sqrt{\frac{\log(2)}{n}}$.

Therefore,
$$R_n^\star \geq \frac{1}{8}\Phi\left(\frac{\delta}{2}\right) = \frac{1}{8}\frac{\delta^2}{4} = \frac{\log(2)}{128} \cdot \frac{1}{n}.$$

Upper bound. Using the sample mean, the minimax risk can be upper bounded by,
$$R_n^\star = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} R(\widehat{\theta}, P) \leq \sup_{P \in \mathcal{P}} R(\text{sample-mean}, P) = \sup_{P \in \mathcal{P}} \frac{\mathrm{Var}_P(X)}{n} = \frac{1}{4n}.$$

Hence $1/n$ is the minimax rate.

Example 3: A simplified regression problem

Let $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ where $X_i \sim \mathrm{Unif}\,(0,1)$ and Y_i is drawn from a distribution with mean $f(X_i)$, and variance bounded by σ^2. We will assume that the regression function $f(\cdot) = \mathbb{E}[Y|X = \cdot]$ is bounded in $[0,1]$ and is L-Lipschitz.

Therefore,

$$\mathcal{P} = \Big\{ P_{XY};\ P_X = \mathrm{Unif}\,(0,1),$$
$$f(\cdot) \triangleq \mathbb{E}[Y|X = \cdot] \text{ is } L\text{-Lipschitz and bounded between 0 and 1}$$
$$Y|X \text{ has variance bounded by } \sigma^2 \Big\}$$

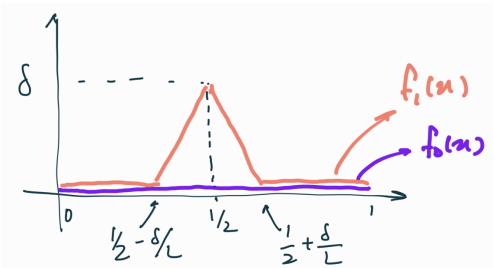
We wish to estimate $\theta = f(1/2) = \mathbb{E}[Y|X = 1/2]$ under the squared loss $\Phi \circ \rho(\theta_1, \theta_2) = (\theta_1 - \theta_2)^2$.

Example 3: A simplified regression problem (cont'd)

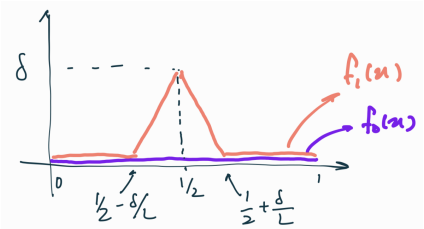
Lower bound. To apply LeCam's method, let

$$P_0 : P_0(Y|X = x) = \mathcal{N}(f_0(x), \sigma^2),$$
$$P_1 : P_1(Y|X = x) = \mathcal{N}(f_1(x), \sigma^2).$$

Therefore, $\delta = |\theta(P_0) - \theta(P_1)| = |f_0(1/2) - f_1(1/2)|$.



Example 3: A simplified regression problem (cont'd)



We will choose,

$$f_0(x) = 0, \qquad f_1(x) = \begin{cases} L(1/2 - x) + \delta & \text{if } x \in (1/2, 1/2 + \delta/L), \\ L(x - 1/2) + \delta & \text{if } x \in (1/2 - \delta/L, 1/2), \\ 0 & \text{otherwise.} \end{cases}$$

Hence, my separation is δ.

Example 3: A simplified regression problem (cont'd)

Let us now upper bound the KL divergence between the two distributions:

$$\begin{aligned}
\mathrm{KL}(P_0, P_1) &= \int_0^1 \int p_0(x,y) \log\left(\frac{p_0(x,y)}{p_1(x,y)}\right) \mathrm{d}y\mathrm{d}x \\
&= \int_0^1 \int p_0(y|x)p_0(x) \log\left(\frac{p_0(y|x)p_0(x)}{p_1(y|x)p_1(x)}\right) \mathrm{d}y\mathrm{d}x \\
&\qquad \text{As } p_0(x) = p_1(x) = 1 \\
&= \int_0^1 \underbrace{\int \phi(y; f_0(x), \sigma^2) \log\left(\frac{\phi(y; f_0(x), \sigma^2)}{\phi(y; f_0(x), \sigma^2)}\right) \mathrm{d}y}_{=\mathrm{KL}(\mathcal{N}(f_0(x),\sigma^2),\, \mathcal{N}(f_1(x),\sigma^2))} \mathrm{d}x \\
&\qquad \text{Denoting the normal pdf by } \phi. \\
&= \int_0^1 \frac{1}{2\sigma^2} \left(f_0(x) - f_1(x)\right)^2 \mathrm{d}x
\end{aligned}$$

Example 3: A simplified regression problem (cont'd)

Recall,

$$f_0(x) = 0, \qquad f_1(x) = \begin{cases} L(1/2 - x) + \delta & \text{if } x \in (1/2, 1/2 + \delta/L), \\ L(x - 1/2) + \delta & \text{if } x \in (1/2 - \delta/L, 1/2), \\ 0 & \text{otherwise.} \end{cases}$$

$$\begin{aligned} \mathrm{KL}(P_0, P_1) &= \int_0^1 \frac{1}{2\sigma^2} \left(f_0(x) - f_1(x)\right)^2 \mathrm{d}x \\ &= \frac{1}{2\sigma^2} \left(\int_{1/2-\delta/L}^{1/2} L(x - 1/2 + \delta/L)^2 \mathrm{d}x + \int_{1/2}^{1/2+\delta/L} L(1/2 + \delta/L - x)^2 \mathrm{d}x \right) \\ &= \frac{1}{2\sigma^2} \left(L^2 \left[\frac{(x - 1/2 + \delta/L)^3}{3} \right]_{1/2-\delta/L}^{1/2} + L^2 \left[\frac{(x - 1/2 - \delta/L)^3}{3} \right]_{1/2}^{1/2+\delta/L} \right) \\ &= \frac{\delta^3}{3\sigma^2 L}. \end{aligned}$$

Example 3: A simplified regression problem (cont'd)

We just showed $\mathrm{KL}(P_0, P_1) = \frac{\delta^3}{3\sigma^2 L}$.

We want $\mathrm{KL}(P_0, P_1) \leq \frac{\log(2)}{n}$, so choose $\delta = \frac{(3\sigma^2 L \log(2))^{1/3}}{n^{1/3}}$. Therefore,

$$R_n^\star \geq \frac{1}{8}\Phi\left(\frac{\delta}{2}\right) = \frac{1}{8}\frac{\delta^2}{4} = c\frac{\sigma^{4/3}L^{2/3}}{n^{2/3}}.$$

N.B. We require $\delta \leq 1$ and $\delta/L < 1/2$ (see our construction), as $f : [0,1] \to [0,1]$. So this lower bound applies only when $n \geq \max(3\sigma^2 L \log^2(2), 24\sigma^2 \log(2)/L^2)$ (larger than some constant).

Example 3: A simplified regression problem (cont'd)

Upper bound. In Appendix A, we show that the following nearest neighbor estimator $\widehat{\theta}$ for $\theta(P) = \mathbb{E}_P[Y|X = 1/2]$, achieves $R(\widehat{\theta}, \theta) \in \mathcal{O}\left(\frac{\sigma^{4/3} L^{2/3}}{n^{2/3}}\right)$, so $n^{-2/3}$ is the minimax rate for this problem.

Let h be a parameter to be chosen later (to balance bias/variance),

$$N = \sum_{i=1}^{n} \mathbb{1}\left(X_i \in (1/2 - h,\, 1/2 + h)\right),$$

$$\widetilde{\theta}(S) = \begin{cases} 1/2 & \text{if } N = 0, \\ \dfrac{1}{N} \displaystyle\sum_{i=1}^{n} Y_i \mathbb{1}\left(X_i \in (1/2 - h,\, 1/2 + h)\right). \end{cases}$$

$$\widehat{\theta}(S) = \mathrm{clip}\left(\widetilde{\theta}(S),\, 0,\, 1\right).$$

Why is Le Cam's method (binary testing) insufficient?

LeCam's method is useful only in relatively "easy" settings. For instance, they work well for point estimation problems, i.e estimating a single parameter of a distribution. In HW1, you will also apply it for estimating a categorical distribution.

However, as the following example illustrates, it does not always work well when there are a large number of estimable parameters.

Normal mean estimation in $\mathbb{R}^d$. Let $\mathcal{P} = \{\mathcal{N}(\mu, \Sigma); \mu \in \mathbb{R}^d, \Sigma_{i,i} \leq \sigma^2\}$, and σ^2 is known. We wish to estimate the mean $\theta(P) = \mathbb{E}_{X\sim P}[X]$ in the L_2 norm:

$$\Phi \circ \rho(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|_2^2, \qquad \rho(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|_2, \quad \Phi(t) = t^2.$$

Upper bound: We can consider the sample mean, $\widehat{\theta}(S) = \sum_{i=1}^n X_i$. Then,

$$R(\widehat{\theta}, P) = \mathbb{E}\left[\|\widehat{\theta}(S) - \theta(P)\|_2^2\right] = \sum_{j=1}^d \left(\frac{1}{n}\sum_{i=1}^n X_{i,j} - \theta_j\right)^2 = \frac{\sigma^2 d}{n}.$$

Why is Le Cam's method (binary testing) insufficient? (cont'd)

Lower bound: Let us try applying LeCam's method. Let,

$$P_0 = \mathcal{N}(\mathbf{0}_d, \sigma^2 I_d), \qquad P_1 = \mathcal{N}(\delta v, \sigma^2 I_d), \text{ for some } v \text{ such that } \|v\|_2 = 1.$$

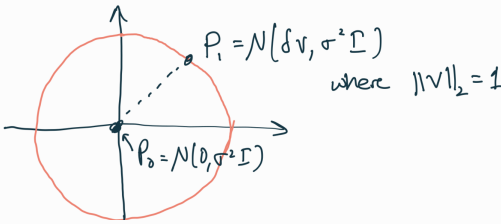
We have, $\mathrm{KL}(P_0, P_1) = \frac{\delta^2}{2\sigma^2}$. So choose $\delta = \sqrt{\frac{2\log(n)}{n}}$, so that $\mathrm{KL}(P_0, P_1) \leq \log(2)/n$.

Therefore, (via the exact same calculations in Example 1), we have, the following lower bound which is off by a factor d.

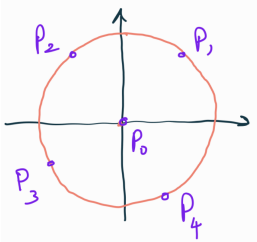
$$R_n^\star \geq \frac{1}{8}\Phi\left(\frac{\delta}{2}\right) \geq \frac{\log(2)}{16}\frac{\sigma^2}{n}$$

Why is Le Cam's method (binary testing) insufficient? (cont'd)

Intuition: The estimate can be wrong in many directions. But Le Cam's only allows you to capture one such direction.



To get the right rates, we need to reduce this estimation problem to multiple hypothesis testing.



Key challenge. We can apply Fano's method for multiple testing, but constructing alternatives for Fano's method usually requires more work than Le Cam's.

Constructing alternatives for Fano's method

We will look at two methods:

1. Tight packings
2. Gilbert-Varshamov Bound

We will start with tight packings.

Packing number. Let $(\mathcal{X}, \rho)$ be a pseudo-metric space and let $A \subset \mathcal{X}$. Let $\epsilon > 0$. A set $P \subset A$ is called an ϵ-**packing** of A if, $\rho(x, x') > \epsilon$ (note strict inequality) for all $x, x' \in P$ such that $x \neq x'$.

The ϵ-**packing number** $M(\epsilon, A, \rho)$ is the size of the largest ϵ-packing of A.

Example 4: Normal mean estimation in $\mathbb{R}^d$

Let $\mathcal{P} = \{\mathcal{N}(\mu, \Sigma); \mu \in \mathbb{R}^d, \Sigma_{i,i} \le \sigma^2\}$, and σ^2 is known. We wish to estimate the mean $\theta(P) = \mathbb{E}_{X\sim P}[X]$ in the L_2 norm:

$$\Phi \circ \rho(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|_2^2, \qquad \rho(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|_2, \quad \Phi(t) = t^2.$$

We will establish a lower bound via the following 4 steps.

Recall the following Theorem. Let $\mathcal{X} = \mathbb{R}^d$ and let $\|\cdot\|$ be any norm. Let $B = \{x \in \mathbb{R}^d; \|x\| \le 1\}$ be the unit ball. Then, $\left(\frac{1}{\epsilon}\right)^d \frac{\mathrm{vol}(A)}{\mathrm{vol}(B)} \le N(\epsilon, A, \|\cdot\|) \le M(\epsilon, A, \|\cdot\|)$.

Let U be a maximal δ packing of the L_2 ball of radius 2δ in $\mathbb{R}^d$. Let $\mathcal{P}' = \left\{\mathcal{N}(u, \sigma^2 I_d); u \in \mathcal{U}\right\}$.

By the theorem above, $|\mathcal{P}'| \ge= \left(\frac{1}{\delta}\right)^d \frac{\mathrm{vol}(2\delta B)}{\mathrm{vol}(B)} = \left(\frac{1}{\delta}\right)^d \frac{(2\delta)^d \mathrm{vol}(B)}{\mathrm{vol}(B)} = 2^d$.

Moreover, for any $u, u' \in U$, we have $\|u - u'\|_2 \ge \delta$. Hence, the separation is at least δ.

Example 4: Normal mean estimation in $\mathbb{R}^d$ (cont'd)

Recall, Local Fano for parameter estimation. Let S be an i.i.d dataset from some distribution $P \in \mathcal{P}$. Let $\{P_1, \ldots, P_N\} \subset \mathcal{P}$ such that $N \geq 16$, $\delta \geq \rho(\theta(P_j), \theta(P_k))$, and $\mathrm{KL}(P_j, P_k) \leq \log(N)/4n$ for all $j \neq k$. Then, $R_n^\star \geq \frac{1}{2}\Phi\left(\frac{\delta}{2}\right)$.

Next, let us upper bound the KL divergence. For any $P_u = \mathcal{N}(u, \sigma^2 I)$ and $P_{u'} = \mathcal{N}(u', \sigma^2 I)$, we have

$$\mathrm{KL}(P_u, P_{u'}) = \frac{\|u - u'\|_2^2}{2\sigma^2} \underbrace{\leq}_{\text{radius } 2\delta} \frac{(4\delta)^2}{2\sigma^2} = \frac{8\delta^2}{\sigma^2}.$$

We require $\mathrm{KL}(P_u, P_{u'}) \leq \frac{\log(|\mathcal{P}'|)}{4n}$ for all $u \neq u'$. As we showed that $|\mathcal{P}'| \geq 2^d$, it is sufficient if we choose, $\delta = \sigma\sqrt{\frac{d\log(2)}{32n}}$. Therefore, by the local Fano method,

$$R_n^\star \geq \frac{1}{2}\Phi\left(\frac{\delta}{2}\right) = \frac{1}{2}\frac{\delta^2}{4} = c \cdot \frac{\sigma^2 d}{n}.$$

This achieves the correct rate of d/n. As we need to satisfy the $N \geq 16$ condition, the lower bound is valid when $d \geq 4$.

Gilbert-Varshamov Bound

Often[3], it is convenient to consider alternatives in a hypercube in the following form,

$$\mathcal{P}' = \{P_\omega; \omega = (\omega_1, \ldots, \omega_m) \in \{0,1\}^m\} \subset \mathcal{P}.$$

But in this hypercube, the minimum distance between alternatives will be small relative to the largest KL.

Let us revisit our normal mean estimation example, but consider the following alternatives:

$$\mathcal{P}' = \left\{\mathcal{N}(\delta\omega, \sigma^2 I_d); \omega \in \{0,1\}^d\right\},$$

We have,

$$\min_{\omega,\omega'} \rho(\theta(P_\omega), \theta(P_{\omega'})) = \min_{\omega,\omega'} \|\delta\omega - \delta\omega'\| = \delta.$$

$$\max_{\omega,\omega'} \mathrm{KL}(P_\omega, P_{\omega'}) = \frac{\max_{\omega,\omega'} \|\delta\omega - \delta\omega'\|_2^2}{2\sigma^2} = \frac{d\delta^2}{2\sigma^2}.$$

[3]We will see several examples in the next two chapters.

Gilbert-Varshamov Bound (cont'd)

Recall, Local Fano for parameter estimation. Let S be an i.i.d dataset from some distribution $P \in \mathcal{P}$. Let $\{P_1, \ldots, P_N\} \subset \mathcal{P}$ such that $N \geq 16$, $\delta \geq \rho(\theta(P_j), \theta(P_k))$, and $\mathrm{KL}(P_j, P_k) \leq \log(N)/4n$ for all $j \neq k$. Then, $R_n^\star \geq \frac{1}{2}\Phi\left(\frac{\delta}{2}\right)$.

We just showed: $\min_{\omega,\omega'} \rho(\theta(P_\omega), \theta(P_{\omega'})) = \delta$, $\max_{\omega,\omega'} \mathrm{KL}(P_\omega, P_{\omega'}) = \frac{d\delta^2}{2\sigma^2}$.

We want the max KL to be smaller than $\frac{\log(|\mathcal{P}'|)}{4n} = \frac{d\log(2)}{4n}$.

So choose, $\delta = \sigma\sqrt{\frac{\log(2)}{2n}}$. This gives,

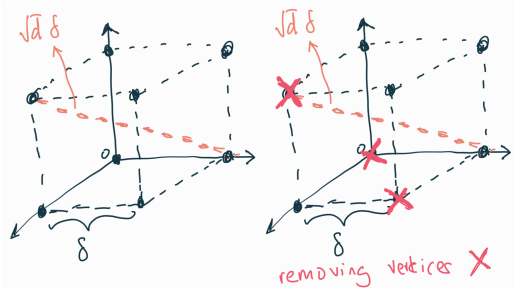
$$R_n^\star \geq \frac{1}{2}\Phi\left(\frac{\delta}{2}\right) = \frac{\delta^2}{8} \asymp \frac{\sigma^2}{n}.$$

So we don't get the d factor.

Gilbert-Varshamov Bound (cont'd)

Why does this happen? The max KL is large relative to the min distance:
$\min_{\omega,\omega'} \rho(\theta(P_\omega), \theta(P_{\omega'})) = \delta, \qquad \max_{\omega,\omega'} \mathrm{KL}(P_\omega, P_{\omega'}) = \frac{d\delta^2}{2\sigma^2}$.

We can try removing elements from this cube to make the distance large, but then the number of alternatives will become too small. We still need exponentially many alternatives to get a tight lower bound.



The GV bound says that we can find a large subset of $\{0,1\}^m$ so that the minimum distance is large.

Gilbert-Varshamov Bound (cont'd)

The Gilbert-Varshamov Bound is a classical result in coding theory. The following is one version of this result.

Theorem, Gilbert-Varshamov bound. Let $m \geq 8$. For any two $\omega, \omega' \in \{0,1\}^m$, let $H(\omega, \omega') = \sum_{i=1}^{m} \mathbb{1}(\omega_i \neq \omega'_i)$ denote the Hamming distance.
Then, there exists $\Omega_m \subset \{0,1\}^m$ such that

- $|\Omega_m| \geq 2^{m/8}$.
- for all $\omega, \omega' \in \Omega_m$, we have $H(\omega, \omega') \geq m/8$.
- $\mathbf{0}_m \in \Omega_m$.

We will refer to Ω_m as the Gilbert-Varshamov pruned hypercube.

Example 4 revisited: Normal mean estimation in $\mathbb{R}^d$

Let $\mathcal{P} = \{\mathcal{N}(\mu, \Sigma); \mu \in \mathbb{R}^d, \Sigma_{i,i} \leq \sigma^2\}$, and σ^2 is known. We wish to estimate the mean $\theta(P) = \mathbb{E}_{X\sim P}[X]$ in the L_2 norm:

$$\Phi \circ \rho(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|_2^2, \qquad \rho(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|_2, \quad \Phi(t) = t^2.$$

Lower bound. Let us consider the following alternatives,

$$\mathcal{P}' = \left\{\mathcal{N}(\delta\omega, \sigma^2 I_d); \omega \in \Omega_d\right\},$$

where Ω_d is the Gilbert-Varshamov pruned d-hypercube.

We then have,

$$\begin{aligned}
\min_{\omega,\omega' \in \Omega_d} \rho(\theta(P_\omega), \theta(P_{\omega'})) &= \min_{\omega,\omega' \in \Omega_d} \|\delta\omega - \delta\omega'\| \\
&= \delta \min_{\omega,\omega' \in \Omega_d} \sqrt{H(\omega, \omega')} \\
&\geq \delta\sqrt{d/8}. \qquad \text{By Gilbert-Varshamov bound}
\end{aligned}$$

So the separation is $\delta\sqrt{d/8}$.

Example 4 revisited: Normal mean estimation in $\mathbb{R}^d$ (cont'd)

Recall, Local Fano for parameter estimation. Let S be an i.i.d dataset from some distribution $P \in \mathcal{P}$. Let $\{P_1, \ldots, P_N\} \subset \mathcal{P}$ such that $N \geq 16$, $\rho(\theta(P_j), \theta(P_k)) \geq \delta$, and $\mathrm{KL}(P_j, P_k) \leq \log(N)/4n$ for all $j \neq k$. Then, $R_n^\star \geq \frac{1}{2}\Phi\left(\frac{\delta}{2}\right)$.

Now, let us compute the maximum KL,

$$\max_{\omega,\omega'} \mathrm{KL}(P_\omega, P_{\omega'}) = \frac{\max_{\omega,\omega'} \|\delta\omega - \delta\omega'\|_2^2}{2\sigma^2} = \frac{d\delta^2}{2\sigma^2}.$$

We want the max KL to be smaller than $\frac{\log(|\mathcal{P}'|)}{4n} = \frac{(d/8)\log(2)}{4n}$. So choose, $\delta = \sigma\sqrt{\frac{\log(2)}{16n}}$. This gives,

$$R_n^\star \geq \frac{1}{2}\Phi\left(\frac{\delta\sqrt{d/8}}{2}\right) = \frac{\delta^2 d}{64} = \frac{\log(2)}{1064}\frac{d\sigma^2}{n}$$

Therefore, $\frac{\sigma^2 d}{n}$ is the minimax rate.

N.B. We require $N \geq 16$, *i.e.*, $2^{d/8} \geq 16$. So this applies only when $d \geq 32$.

Summary

Let us quickly summarize how we prove lower bounds for learning problems.

- Reduce estimation to hypothesis testing (RTT): For alternatives that are δ-separated in the loss L,

$$R^\star = \inf_{\widehat{A}} \sup_{P \in \mathcal{P}} \mathbb{E}_{S \sim P} \left[L(\widehat{A}(S), P) \right] \geq \delta \cdot \inf_{\psi} \max_{j \in [N]} P_j(\psi \neq j).$$

  Parameter estimation: for alternatives that are δ-separated in some metric ρ,

$$R^\star = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{S \sim P} \left[\Phi \circ \rho(\theta(P), \widehat{\theta}(S)) \right] \geq \Phi\left(\frac{\delta}{2}\right) \inf_{\psi} \max_{j \in [N]} P_j(\psi \neq j).$$

- LeCam's method: reduces to a binary hypothesis testing problem:

$$\max_{j \in \{0,1\}} P_j(\psi \neq j) \underbrace{\geq}_{\max \geq \text{avg}} \frac{1}{2} \left(P_0(\psi \neq 0) + P_1(\psi \neq 1) \right) \underbrace{\geq}_{\text{NP-test}} \frac{1}{2} \| P_0 \wedge P_1 \| \geq \frac{1}{4} e^{-\mathrm{KL}(P_0, P_1)}.$$

  Useful mostly for point estimation problems.

Summary (cont'd)

- Fano's method: reduces to a multiple hypothesis testing problem:

$$\max_{j\in[N]} P_j(\psi \neq j) \underbrace{\geq}_{\max \geq \text{avg}} \underbrace{\mathbb{P}_{V,S}(\psi \neq V)}_{\frac{1}{N}\sum_{j=1}^{N} P_j(\psi \neq j)} \underbrace{\geq}_{\text{Fano's inequality}} 1 - \frac{I(V,S) + \log(2)}{\log(N)}.$$

  By bounding $I(V,S)$ we get the global and local Fano methods.

- Need to construct alternatives $\{P_1, \ldots, P_N\}$ for the local Fano method carefully, using tight packings or the Gilbert-Varshamov bound.

- Four steps to establishing a lower bound:

  1. Construct alternatives.
  2. Lower bound the separation.
  3. Upper bound the KL divergence.
  4. Compute the lower bound.

Plan going forward

We will prove lower bounds for the following problems:

- Ch 2.1: Nonparametric regression (lower and upper bounds)
- Ch 2.2: Nonparametric density estimation
- Ch 3: Lower bounds for excess risk based prediction problems.
- Ch 4 onwards: lower bounds for stochastic/adversarial bandits, mostly using the Bretagnolle-Huber inequality.

Appendix

Appendix A: Upper bound for Example 3

Upper bound. (Read at home)

We will design the following nearest neighbor estimator $\widehat{\theta}$ for $\theta(P) = \mathbb{E}_P[Y|X = 1/2]$.

$$N = \sum_{i=1}^{n} \mathbb{1}\left(X_i \in (1/2 - h,\, 1/2 + h)\right),$$

$$\widetilde{\theta}(S) = \begin{cases} 1/2 & \text{if } N = 0, \\ \dfrac{1}{N} \displaystyle\sum_{i=1}^{n} Y_i \mathbb{1}\left(X_i \in (1/2 - h,\, 1/2 + h)\right). \end{cases}$$

$$\widehat{\theta}(S) = \mathrm{clip}\left(\widetilde{\theta}(S),\, 0,\, 1\right).$$

We will specify the value of h shortly.

Appendix A: Upper bound for Example 3 (cont'd)

Note that N is a $\mathrm{Binomial}\,(n, 2h)$ random variable. Let us define $G = \{N \geq hn\}$ to be the "good event" in which a sufficient number of points fall within the $[1/2 - h, 1/2 + h]$ interval. We have, by Hoeffding's inequality,

$$\begin{aligned}\mathbb{P}(G^c) &= \mathbb{P}\left(\sum_{i=1}^{n} \mathbb{1}(X_i \in (1/2 - h,\, 1/2 + h)) \;\leq\; nh\right) \\ &= \mathbb{P}\left(\sum_{i=1}^{n} (\mathbb{1}(X_i \in (1/2 - h,\, 1/2 + h)) - 2h) \;\leq\; -nh\right) \leq e^{-2nh^2}.\end{aligned}$$

We can now write,

$$\begin{aligned}\mathbb{E}\left[(\widehat{\theta}(S) - \theta)^2\right] &\leq \mathbb{E}\left[(\widetilde{\theta}(S) - \theta)^2\right] \\ &= \mathbb{E}\left[(\widetilde{\theta}(S) - \theta)^2 \middle| G\right] \underbrace{\mathbb{P}(G)}_{\leq 1} + \underbrace{\mathbb{E}\left[(\widetilde{\theta}(S) - \theta)^2 \middle| G^c\right]}_{\leq 1/4 \text{ as } f \text{ is bounded in } [0,1]} \underbrace{\mathbb{P}(G^c)}_{\leq e^{-2nh^2}}\end{aligned}$$

Appendix A: Upper bound for Example 3 (cont'd)

To upper bound $\mathbb{E}[(\widehat{\theta}(S)-\theta)^2|G]$, let us denote $A_i = \mathbb{1}(X_i \in (1/2-h, 1/2+h))$ and expand $(\widetilde{\theta}(S)-\theta)^2$ as follows:

$$\begin{aligned}\left(\widetilde{\theta}(S)-\theta\right)^2 &= \left(\frac{1}{N}\sum_{i=1}^{n} A_i Y_i - \theta\right)^2 \\ &= \left(\underbrace{\frac{1}{N}\sum_{i=1}^{n} A_i\left(Y_i - f(X_i)\right)}_{v} + \underbrace{\frac{1}{N}\sum_{i=1}^{N} A_i\left(f(X_i)-\theta\right)}_{b}\right)^2 \\ &= v^2 + b^2 + 2bv.\end{aligned}$$

You may interpret b as the bias and v^2 as the variance. The quantity v captures the extent to which the observations Y_i deviate from their expected values $f(X_i)$, while b quantifies the deviation of $\theta = f(1/2)$ from the surrounding $f(X_i)$ values, as we consider an interval around $1/2$.

Appendix A: Upper bound for Example 3 (cont'd)

Now, let us bound the individual terms. We will start with v.

$$\begin{aligned}
\mathbb{E}[v^2] &= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{n} A_i\left(Y_i - f(X_i)\right)\right)^2 \middle| G, X_1, \ldots, X_n\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{1}{N^2}\sum_{i=1}^{n} A_i\left(Y_i - f(X_i)\right)^2 \middle| G, X_1, \ldots, X_n\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{1}{N^2}\sum_{i=1}^{n} A_i \operatorname{Var}(Y_i|X_i) \middle| G, X_1, \ldots, X_n\right]\right] \qquad \texttt{Note that } \mathbb{E}[Y_i|X_i] = f(X_i) \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{\sigma^2 N}{N^2} \middle| G\right]\right] \leq \frac{\sigma^2}{nh}. \qquad \texttt{As } N \geq nh \texttt{ under } G.
\end{aligned}$$

Appendix A: Upper bound for Example 3 (cont'd)

Next, let us consider b. For any $X_i \in (1/2 - h, 1/2 + h)$, we have

$$|f(X_i) - f(1/2)| \leq L|X_i - 1/2| \leq Lh.$$

Hence,

$$|b| = \left| \frac{1}{N} \sum_{i=1}^{N} A_i \left(f(X_i) - \theta \right) \right| \leq \frac{1}{N} \sum_{i=1}^{N} A_i \left| f(X_i) - \theta \right| \leq \frac{1}{N} \sum_{i=1}^{N} A_i Lh = Lh.$$

Therefore, $\mathbb{E}[b^2|G] \leq L^2 h^2$.

Finally, let us consider the cross-term. As $\mathbb{E}[Y_i|X_i] = f(X_i)$, we have,

$$\mathbb{E}[bv|G] = \mathbb{E}\left[b \cdot \mathbb{E}_Y \left[\frac{1}{N} \sum_{i=1}^{N} A_i \left(f(X_i) - \theta \right) \middle| G, X_1, \ldots, X_n \right] \right] = 0.$$

Appendix A: Upper bound for Example 3 (cont'd)

Therefore,

$$\mathbb{E}\left[\left(\widehat{\theta}(S)-\theta\right)^2\right] \leq e^{-2nh^2} + \frac{\sigma^2}{nh} + L^2h^2.$$

Choosing $h = \frac{\sigma^{2/3}}{L^{2/3}n^{1/3}}$, we get

$$\mathbb{E}\left[\left(\widehat{\theta}(S)-\theta\right)^2\right] \leq \exp\left(\frac{-2\sigma^{4/3}}{L^{4/3}}n^{1/3}\right) + 2\frac{\sigma^{4/3}L^{2/3}}{n^{2/3}}.$$

Therefore, the bound is tight in L, σ *if we know* L, σ.

If L, σ are unknown, we can still choose $h = n^{-1/3}$. Then, it is still tight in n,

$$\mathbb{E}\left[\left(\widehat{\theta}(S)-\theta\right)^2\right] \leq \exp\left(-2n^{1/3}\right) + \frac{1}{n^{2/3}}\left(L^2+\sigma^2\right).$$

N.B. Had we used Chernoff's instead of Hoeffding's to control $\mathbb{P}(G^c)$, we would have a slightly faster rate in the lower order $e^{-n^{1/3}}$ term.