ECE598: Information-theoretic methods in high-dimensional statistics Spring 2016

Lecture 1: Introduction

Lecturer: Yihong Wu *Scribe: AmirEmad Ghassami, Jan 21, 2016 [Ed. Jan 31]*

Outline:

- Introduction of the framework

  Parametric model vs. non-parametric model

- Best estimator

1.1 Basis of Statistical Decision Theory

- **Statistical Experiment:** A collection of probability distributions (over a common measurable space $(\mathcal{X}, \mathcal{F})$).
$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}$$

- **Data:**
$$X \sim P_\theta \qquad \text{for some } \theta \in \Theta$$
X could be a random variable, vector, process, etc, depending on $\mathcal{X}$.

- **Objective:**
$$\begin{aligned} T: \ & \Theta \to \mathcal{Y} \\ & \theta \mapsto T(\theta) \end{aligned}$$
The value $T(\theta)$ is what we want to estimate, which can be θ itself, or a relevant aspect of θ, e.g., a function of θ such as $\|\theta\|_2$.

- **Estimator (Decision Rule):**
$$\hat{T} : \mathcal{X} \to \hat{\mathcal{Y}}$$
Note the that $\hat{\mathcal{Y}}$ need not be the same as $\mathcal{Y}$.

  Remark 1.1. $\hat{T}$ can be a deterministic or randomized estimator:

  - deterministic estimator: $\hat{T} = \hat{T}(X)$.
  - randomized estimator: $\hat{T} = \hat{T}(X,$ external randomness$)$. In this case $\hat{T}$ should be viewed as a conditional probability distribution $P_{\hat{T}|X}$ (Markov transition kernel).

  The problem in statistical experiment is as follows: By choosing the parameter θ, nature picks a distribution that generates the data X. The statistician observes the data and computes an estimation $\hat{T}$ of $T(\theta)$. The goal is for $\hat{T}$ to be close to T. To that end, we need to introduce a metric to quantify how good $\hat{T}$ is:

- **Loss Function:**

$$l : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}$$
$$T \times \hat{T} \mapsto l(T, \hat{T})$$

Since we are dealing with loss, all the negative (converse) results are lower bound and all the positive (achievable) results are upper bound.

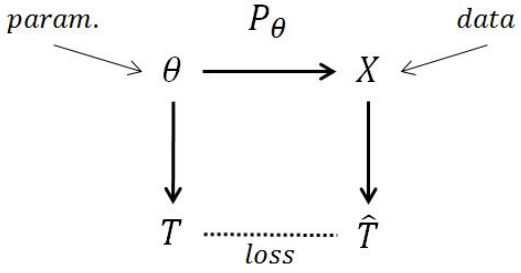
Note: Since X is a random variable, the estimator is also a random variable. Hence, $l(T, \hat{T})$ is a random variable. Therefore, to make sense of "minimizing the loss", we define the following:

- **Risk:**

$$R_\theta(\hat{T}) = \mathbb{E}_\theta[l(T, \hat{T})] = \int P_\theta(dx) P_{\hat{T}|X}(d\hat{t}|x) l(t(\theta), \hat{t}),$$

which we refer to as the risk of $\hat{T}$ at θ. Note that the expected risk depends on the strategy as well as where the truth is. The subscript indicates the distribution with respect to which the expectation is taken.

The following diagram summarizes the process:



Example 1.1.
Gaussian Location Model (GLM): or Normal Mean Model, Additive Gaussian-Noise Channel

– Model:

$$\mathcal{P} = \{\mathcal{N}(\theta, I_p) : \theta \in \Theta\}$$

where I_p is the p-dimensional identity matrix and $\Theta \subset \mathbb{R}^p$. Equivalently,

$$X = \theta + Z \qquad Z \sim \mathcal{N}(0, I_p), \theta \in \Theta \subset \mathbb{R}^p.$$

- $p = 1$: scalar case
- $p > 1$: vector case
  We also encompass matrix case: By arranging a p^2-dimensional vector into a $p \times p$ matrix. In this case $\Theta \subset \mathbb{R}^{p \times p}$.

– Objective: Examples of the objective include $T(\theta) = \theta$, $\|\theta\|_2$, $\theta_{max} = \max_{i \in [p]} \theta_i$, where $[p] = \{1, \cdots, p\}$.

– Loss function: Examples of the loss function include the following:

$$l(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_2^2, \|\theta - \hat{\theta}\|_1, \cdots$$
$$\text{In the matrix case}: l(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_F^2, \|\theta - \hat{\theta}\|_{op}, \cdots$$

– Estimator: Examples of the estimator include the following:

$$\text{Maximum Likelihood Estimator: } \hat{\theta} = X$$
$$\text{James-Stein estimator: } \hat{\theta}_{JS} = \left(1 - \frac{p-2}{\|X\|_2^2}\right) X$$

The choice of the estimator mainly depends on the objective.

– Parameter space: Examples of the parameter space include the following:

a) $\Theta = \mathbb{R}^p$: unstructured.

b) $\Theta = \{\text{all } k\text{-sparse vectors}\} = \{\theta \in \mathbb{R}^p : \|\theta\|_0 \leq k\}$, where $\|\theta\|_0 \triangleq |\{i : \theta_i \neq 0\}|$ denotes the size of the support.

$\Theta = l_q$-norm balls, $0 \leq q \leq \infty$, where $\|\theta\|_q = (\sum |\theta_i|^q)^{\frac{1}{q}}$.

c) Matrix case: low-rank matrices: $\Theta = \{\theta : rank(\theta) \leq r\}$.

Note that by definition, more structure (smaller paramater space) always leads to smaller risk; but it need not simplify the computation issue.

– Testing: We have two scenarios and based on the observed data X, we want to determine which one is the true scenario.

$*$ Simple Hypothesis:

$$H_0 : \quad \theta = \theta_0$$
$$H_1 : \quad \theta = \theta_1$$

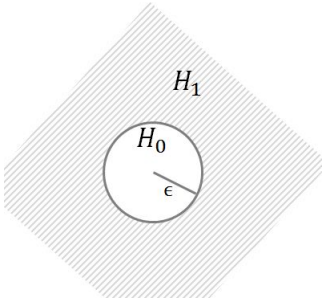
For instance θ_0 could be the all zero vector and θ_1 could be all one vector. Then this corresponds to sending a single bit repeatedly in Gaussian noise.

parameter space $= \Theta = \{\theta_0, \theta_1\} = \hat{\Theta} =$ decision space

$l(\theta, \hat{\theta}) = 1_{\{\theta \neq \hat{\theta}\}}$: This is Hamming loss (zero-one loss).

$*$ Composite Hypothesis:
Example 1: One of the hypothesis is composite.

$$H_0 : \quad \theta = 0$$
$$H_1 : \quad \|\theta\|_2 \geq \epsilon$$

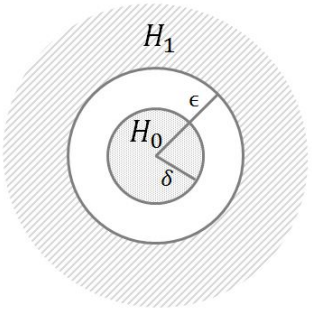


Here, H_0 and H_1 could be interpreted as pure noise case and the case where signal is present, respectively.

$\Theta = \{0\} \cup \{\theta : \|\theta\|_2 \geq \epsilon\}$

Example 2: Both hypothesis are composite.

$$H_0: \quad \|\theta\|_2 \le \delta$$
$$H_1: \quad \|\theta\|_2 \ge \epsilon$$



Here, H_0 and H_1 could be interpreted as the case with weak signal and strong signal, respectively.

Remark 1.2 (Parametric model versus non-parametric model)**.** According to statistical conventions, parametric model refers to the case that the parameter of interest is finite-dimensional while non-parametric model refers to the case that the parameter is infinite-dimensional.

In this class, we are mostly interested in high-dimensional parametric model.

Parametric Model

Examples of parametric model:

- GLM or more generally exponential family. We start with distribution P on $\mathbb{R}^p$, and for $\theta \in \mathbb{R}^p$, consider the tilted distribution

$$dP_\theta = \frac{e^{\langle \theta, X \rangle}}{\mathbb{E}_{X \sim P} e^{\langle \theta, X \rangle}} dP$$

- Covariance matrix estimation:

  $X = (X_1, \cdots, X_n) \overset{iid}{\sim} \mathcal{N}(0, \Sigma)$. In this case, Σ is our parameter and $P_\theta = \mathcal{N}(0, \Sigma)^{\otimes n}$.

  If we want to estimate Σ, we can use the loss function $l(\Sigma, \hat{\Sigma}) = \|\Sigma - \hat{\Sigma}\|$.

  If we want to estimate a function of Σ, $T : \Sigma \to v$ (principle component) we can use loss function $l(v, \hat{v}) = \|span(v) - span(\hat{v})\| = \{vv' - \hat{v}\hat{v}'\}$.

- Stochastic block model: We observe the graph G of size n and the goal is to estimate a subset of nodes C:

$$X = G$$
$$\Theta = \{C : C \subset [n], |C| = n/2\}$$

  For each two nodes i and j, we denote $i \sim j$ as the event that they belong to the same partition, that is either they both belong to C or C^c.

$$P(i \sim j) = \begin{cases} p & \text{if } i, j \in C \text{ or } C^c \\ q & o.w. \end{cases}$$

$$l(C, \hat{C}) = 1_{\{C \neq \hat{C}\}} \text{ or } |C \Delta \hat{C}|$$

- Large alphabet: Estimating a discrete distribution.

$$\mathcal{P} = \{\text{all distributions on } [k]\}$$
$$X = (X_1, \cdots, X_n) \sim P \in \mathcal{P}$$
$$l(P, \hat{P}) = \|P - \hat{P}\| \text{ or } D(P\|\hat{P})$$

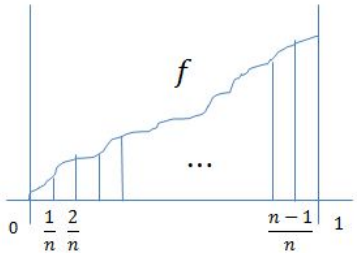
Non-parametric Model

Examples of non-parametric model:

- Density estimation: Here the parameter is a pdf, for example:

$$f \in \mathcal{F} = \{\text{smooth, log concave, monotone}\}$$
$$X = (X_1, \cdots, X_n) \overset{iid}{\sim} f \text{ on } \mathbb{R}^p$$
$$l(f, \hat{f}) = \|f - \hat{f}\|_2^2$$

- Regression: We observe noisy samples at discrete points. The parameter is the unknown function f.



$$X = (X_i) = f(i/n) + Z_i$$

- White Gaussian noise model: we observe a wave form:

$$dX_t = f(t)dt + dB_t$$
$$X_t = \int_0^t f(\tau)d\tau + B_t$$

where B_t is a Brownian motion. Equivalently, if $f \in L^2$ where $\{\phi_i\}$ is an orthonormal basis, then

$$X_i = \langle X, \phi_i \rangle = \theta_i + Z_i \qquad i = 0, 1, \cdots$$

This is called Gaussian Sequence Model (which is GLM with $p = \infty$).

Remark 1.3. Testing:

simple vs. simple

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta = \theta_1 \qquad \Theta = \{\theta_0, \theta_1\}$$

simple vs. composite

$$H_0 : \theta = \theta_0 \ vs. \ H_1 : \theta \in \Theta_1 \qquad \Theta = \{\theta_0\} \cup \Theta_1$$

composite vs. composite

$$H_0 : \theta \in \Theta_0 \ vs. \ H_1 : \theta \in \Theta_1 \qquad \Theta = \Theta_0 \cup \Theta_1$$

$\hat{T}(X) \in \{0, 1\}$ $\qquad l(\theta, \hat{T}) = 1_{\{\theta \notin \Theta_{\hat{T}}\}}$.

Remark 1.4. Confidence interval/region/bond: For example to estimate a function we output a region in which the function lies w.h.p.

$$\hat{T} = \text{some subset}$$
$$l(\theta, \hat{T}) = 1_{\{\theta \notin \hat{T}\}} + \text{size of } \hat{T}$$

Remark 1.5. We frequently deal with *independent sampling model*. In this case:

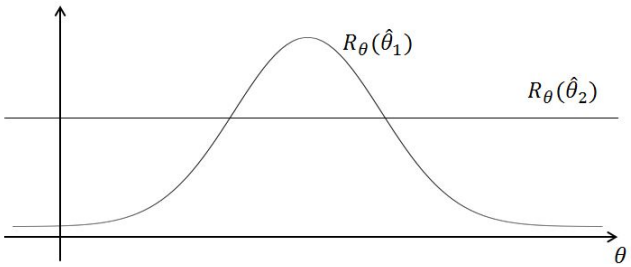
$$X = \underbrace{(X_1 \cdots, X_n)}_{\text{i.i.d. samples}}$$
$$\mathcal{P} = \{P_\theta^{\otimes n} : \theta \in \Theta\}$$

1.2 How to define the "best Estimator"

One of the main objectives of this course is to investigate the fundamental limit, that is, to find the performance of the best estimator. We use the risk of an estimator to quantify its performance. As mentioned in the framework, for an estimator $\hat{\theta}$, we define the risk as follows:

$$R_\theta(\hat{\theta}) = \mathbb{E}_\theta[l(\theta, \hat{\theta})]$$

Note that, $R_\theta(\hat{\theta})$ could be viewed as a function of θ. As an example, the following figure depicts the risk curves for two different estimators.



To find the best estimator, we first need to define the figure of merit.

Naive Method: Find the estimator which is better that all other estimators at all points, i.e., find $\hat{\theta}$, such that

$$R_\theta(\hat{\theta}) \leq R_\theta(\hat{\theta}'), \qquad \forall \theta, \forall \hat{\theta}'.$$

It is easy to see that this method is typically too greedy to be realistic and an estimator that satisfies the requirement above does not exist. For example, consider $\theta_1 \neq \theta_2$ in Θ and $l(\theta, \hat{\theta})$ is

some norm. Consider the estimator $\hat{\theta}_1 = \theta_1$ which throws away data and always spits out θ_1. Then $R_{\theta_1}(\hat{\theta}) \leq R_{\theta_1}(\hat{\theta}_1) = 0$ means $\hat{\theta} = \theta_1$, which means it cannot beat $\hat{\theta}_2 = \theta_2$ now. Therefore, we need other methods to compare estimators.

Method 1 Limit the class of competitors (of $\hat{\theta}$):

In some cases, by restricting the class of estimators, we can find a strategy which is uniformly the best. For example,

- Restricting to unbiased estimators: Frequently it is good to have be biased.
- Restricting to invariant estimators

Method 1 is difficult to generalize to high dimensional problems.

Method 2 Bayes approach: average-case analysis.

Method 3 Minimax approach: worst-case analysis

As mentioned before, finding a curve that dominates all other curves at all points is not always feasible. Hence, in Methods 2 and 3, we summarize a curve to a number so that we can compare them. In Method 2, we give weights to each point and take the average. The weights are called the prior. But the problem is which prior to choose. In Method 3, we consider the worst prior. For example, according to Method 3, in the figure above, $\hat{\theta}_2$ is a better strategy.