Lecture 1



Agenda:
* Course outline
* Testing and learning examples we would like to address
* Decision theoretic framework (minmax and Bayes formulations)
* Total variation distance and KL divergence

PART 1: Detection and Estimation

OR

Distribution learning and testing

A Motivating Examples

Ex1 (Learning bias of a coin)

→ A coin has either bias $p$ or bias $p+\epsilon$. How many times do we need to toss it to find out if it has bias $p$ or $p+\epsilon$?

→ How many times do we need to toss a coin to estimate its bias up to an accuracy of $\epsilon$?

Ex2 (Learning a Gaussian mean)

→ How many real valued observations $X_1, \ldots, X_n$ are needed to estimate the unknown mean $\mu$ of $X_i \sim \mathcal{N}(\mu, 1)$?

Ex3 (Distribution Testing)

→ Uniformity testing: How many samples $X_1, \ldots, X_n$ are needed to check if $X_i$'s are coming from unif$([k])$ or some other $P \in \mathcal{P}([k])$ s.t. $P$ is "$\epsilon$-away" from unif$([k])$.

→ Identity testing: How many samples $(X_i, Y_i)_{i=1}^n$ are needed to check if $X^n$ and $Y^n$ are coming from the same $P \in \mathcal{P}([k])$ or $P_X$ and $P_Y$ that are $\epsilon$-away?

→ Independence testing: $(X_i, Y_i)_{i=1}^n$ are from $P_{XY}$ or something that is $\epsilon$-away from an independent distribution.
$(P_{XY} \in \mathcal{P}([k_1] \times [k_2]))$.

Ex4 (Distribution learning)

→ Gaussian mixture learning: How many samples $X_1, \dots, X_n$ from $\sum_{i=1}^m \omega_i N(\mu_i, K_i)$ are needed to estimate $(\omega_i, \mu_i, K_i)_{i=1}^m$.

Support estimation; function estimation; etc.

Why? – An excuse for coming up with algorithm to be used on real data

– Generative model can actually be a good fit

How? – Classic statistics : * Schemes – ML; Bayesian (hyper-prior)

* Lower bounds : CR Bound and ??

- These bounds and schemes are typically justified by asymptotic behaviour



- In this course, we will give schemes and lower bounds that are valid for finite n.

## B Decision Theoretic Framework (A quick and dirty introduction)

By observing $X \sim P_\theta$, $\theta \in \Theta$, output an estimate $\hat\theta(X)$ of $\theta$.

**Risk or loss:** $r : \Theta \times \Theta \to \mathbb{R}_+$

in more general frameworks, $\hat\theta$ need not be in $\Theta$.

$$(\theta, \hat\theta) \mapsto r(\theta, \hat\theta)$$

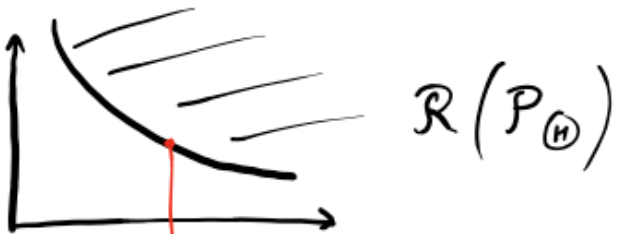
**Average risk vector:** $r_\theta(\hat\theta) = \mathbb{E}_{P_\theta}\, r(\theta, \hat\theta(X))$

$$\underline{r}(\hat\theta) = \{ r_\theta(\hat\theta),\ \theta \in \Theta \}$$

**Risk region:** $\mathcal{R}(P_\Theta) = \overline{co}\, \{ \underline{r}(\hat\theta),\ \hat\theta \in \underbrace{\mathcal{E}(\mathcal{X}; \Theta)}_{\text{set of estimators}} \}$

convex, closure

→ convex may be omitted if randomized rules are allowed




Only estimators on the boundary are of interest

→ **Admissible policies:** A policy is inadmissible if $\forall\, \theta \in \Theta$

$\exists\ \hat{\theta}'$ s.t. $R_\theta(\hat{\theta}') \leq R_\theta(\hat{\theta})$.

The policies that are not inadmissible are <u>admissible</u>.

→ Heuristically, admissible policies are those which are "best" for at least one $\theta$.

(But we will not worry about this; even inadmissible policies are okay for if they perform "close to optimal".)

Bayesian Cost: A prior $\pi$ on $\Theta$, $\pi \in \mathcal{P}(\Theta)$.

$$R_\pi(\hat{\theta}) = \mathbb{E}_{\theta\sim\pi}\left[r_\theta(\hat{\theta})\right]$$

$$R_\pi^* = \inf_{\hat{\theta}\in\mathcal{E}(\mathcal{X};\Theta)} R_\pi(\hat{\theta})$$

→ $\hat{\theta}$ is <u>Bayes</u> for $\pi$ if $R_\pi(\hat{\theta}) = R_\pi^*$.

A meta theorem of Le Cam

Under regularity conditions for $r$, $\mathcal{E}(\mathcal{X};\Theta)$, $\Theta$, a policy is admissible iff it is Bayes.

Minmax cost

$$R^* = \inf_{\hat{\theta}\in\mathcal{E}(\mathcal{X};\Theta)} \sup_{\theta\in\Theta} r_\theta(\hat{\theta}).$$

$$= \inf_{\hat{\theta}} \sup_{\pi\in\mathcal{P}(\Theta)} R_\pi(\hat{\theta})$$

$$\geq \sup_{\pi\in\mathcal{P}(\Theta)} \inf_{\hat{\theta}} R_\pi(\hat{\theta})$$

Often (for instance, when Le Cam's regularity conditions hold), the right-side above equals the left, i.e.,
$$R^* = \sup_{\pi \in \mathcal{P}(\Theta)} \inf_{\hat\theta} R_\pi(\hat\theta)$$

Suppose
$$R^* = \max_{\pi \in \mathcal{P}(\Theta)} \min_{\hat\theta} R_\pi(\hat\theta)$$
$$= \min_{\hat\theta} R_{\pi^*}(\hat\theta)$$

least-favourable prior

→ A good strategy for attaining $R^*$ is using a Bayes policy for the least-favorable prior (or something close to it).

The <u>binary hypothesis testing</u> problem corresponds to $|\Theta| = 2$. Deterministic estimators $\hat\theta$ can be described by an "<u>acceptance</u> region" $A$ where $\theta_0$ is declared when $X \in A$ is observed.

## C <u>Distances b/w distributions</u> (Total variation distance, KL divergence)

The difficulty of estimating $\theta$ is governed by how "close" $P_\theta$ are to each other. We will encounter various distances in this course to capture closeness.

## (a) Total Variation Distance

$$d(P, Q) = \sup_A P(A) - Q(A)$$

If $P, Q$ have densities $f, g$ w.r.t $\mu$, i.e.,

$$P(A) = \int_A f(x)\, \mu(dx),$$

$$d(P, Q) = \frac{1}{2} \int |f(x) - g(x)|\, \mu(dx)$$

For discrete $\mathcal{X}$: $d(P, Q) = \frac{1}{2} \sum_x |P(x) - Q(x)|$

$$P_e^* = \min_A \underbrace{\frac{1}{2} P(A^c) + \frac{1}{2} Q(A)}_{\text{Bayesian cost for unif. prior}}$$

$$= \frac{1}{2}\left(1 - d(P, Q)\right).$$ (Proof HW problem)

Lemma. $P^n = P_1 \times \dots \times P_n$ ; $Q^n = Q_1 \times \dots \times Q_n$, discrete pmf

(1) $$d(P^n, Q^n) \le \sum_{i=1}^n d(P_i, Q_i)$$

Proof.

$$d(P_1 P_2, Q_1 Q_2) = \frac{1}{2} \sum_{x,y} \left| P_1(x) P_2(y) - Q_1(x) Q_2(y) \right|$$

$$\le \frac{1}{2} \sum_{x,y} \left| P_1(x) P_2(y) - Q_1(x) P_2(y) \right|$$

$$+ \frac{1}{2} \sum_{x,y} \left| Q_1(x) P_2(y) - Q_1(x) Q_2(y) \right|$$

$$= d(P_1, Q_1) + d(P_2, Q_2)$$ ∎