**10-704: Information Processing and Learning** **Spring 2015**

# Lecture 19: March 31st

*Lecturer: Aarti Singh,Akshay Krishnamurthy* *Scribes: Rohan Varma*

**Note**: *LaTeX template courtesy of UC Berkeley EECS dept.*



## 19.1 Applications

### 19.1.1 Privacy

We present some channel capacity results.

$$Y = AX + Z, \text{ where } \mathbb{E}\|X\|^2 \leq P, Z \sim^{iid} (0, \sigma^2 I)$$

where A is random $m \times n$ projection. We then have

$$\sup_{p(x)} I(X,Y) \leq \frac{m}{2n}\log(1+\frac{P}{\sigma^2}) \to 0 \text{ at a rate of } \frac{m}{n}, C \leq \frac{m}{2}\log(2\pi eP) \tag{19.1}$$

Example: Compressed Linear Regression. $Y = AX\beta + \epsilon$ where $\beta$ is of dimension $p$ and $s$-sparse, $X$ of dimension $n \times p$, If $m = s^2\log(np)$, then $MSE \to 0$ and $supp(\beta) = supp(\hat{\beta})$. This latter property is known as sparsistency in the literature.

### 19.1.2 Differential Privacy

Differential privacy is a mathematical formalism for a privacy-preserving algorithm. We say an algorithm is $(\epsilon, \delta)$-differentially private if for all inputs $X, X'$ differing in at most one value, and for all possible outcomes $S$:

$$Pr[\mathcal{A}(x) \in S] \leq e^{\epsilon} Pr[\mathcal{A}(x') \in S] + \delta \tag{19.2}$$

where $\mathcal{A}$ refers to the algorithm under consideration.

One can use random projections to achieve differential privacy. If we let $Y = AX + Z$, where $X$ is the original data matrix and $Z$ has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries, then we can achieve $(\epsilon, \delta)$ differential privacy as long as:

$$\sigma^2 \geq (\max_j \|a_j\|_2)\frac{\sqrt{2(\log\frac{1}{2\delta} + \epsilon)}}{\epsilon} \tag{19.3}$$

where $a_j$ are the columns of the matrix $A$.

### 19.1.3 Rate Distortion Approach

$$\min_{\Pi(T|X)} I(X;T) \text{ s.t } \mathbb{E}[\hat{R}_X(T)] \leq \gamma \longleftrightarrow_{Blahut-Arimoto} \Pi(\theta|X) \propto \Pi(\theta)e^{-\beta\hat{R}_X(\theta)} \tag{19.4}$$

using the exponential mechanism. Here $\hat{R}_X(T)$ represents the empirical loss ($\frac{1}{n}\sum_{i=1}^{n} loss_{X_i}(T)$). In addition $\Pi(\theta|X)$ has $(2\beta\Delta_{\ell_1}(\hat{R}_X(\theta)), 0)$ differential privacy, where $\Delta_{\ell_1}(\hat{R}_X(\theta)) = \max_{X\sim X'} \|\hat{R}_X(\theta) - \hat{R}_{X'}(\theta)\|_1$.

## 19.2 Converse of Channel Coding Theorem

The converse of the channel coding theorem states that any rate $R \geq C$ is not achievable.

*Proof.* We use Fano's inequality which states that for $W \to Y$ ,

$$Pr(\hat{W}(Y) \neq W) \geq \frac{H(W|Y) - 1}{\log|W|} \tag{19.5}$$

where W is a rate R code (i.e. $W \in \{1, 2, \cdots 2^{nR}\}$ and $W$ is drawn uniformly at random.). Hence we can write for the setting where $W$ is the message sent over a discrete memoryless channel:

$$W \to X_1^n \to \text{channel} \to Y_1^n$$

and

$$P(\hat{W} \neq W) = \frac{H(W|Y) - 1}{nR} = \frac{H(W) - I(W, Y^n) - 1}{nR} = \frac{nR - I(W, Y^n) - 1}{nR} \tag{19.6}$$

We can additionally bound:

$$\begin{aligned} I(W, Y^n) &\leq I(X^n, Y^n) \\ &= H(Y^n) - H(Y^n|X^n) \\ &\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|Y_{i-1}, \cdots Y_1, X^n) \\ &\leq \sum_{i=1}^{n} H(Y_i) - H(Y_i|X_i) =\leq \sum_{i=1}^{n} I(X_i; Y_i) \leq nC \end{aligned}$$

Hence, we can conclude that

$$P(\hat{W} \neq W) \geq \frac{nR - nC - 1}{nR} \tag{19.7}$$

So that one cannot achieve rates smaller than the capacity. □

## 19.3 Minimax Theory For Testing Problems

The goal of minimax theory broadly is to understand the minimax risk

$$\inf_{T} \sup_{\theta} \mathbb{E}_\theta\left[\ell(T(x_1^n), \theta\right] \tag{19.8}$$

where T is an estimator, $\theta$ is some parameter and the inner term represents the risk.

*Example:* If the range of T is a distribution and $\ell$ is the log-loss, then this is equivalent to "minimax redundancy".

What are alternative definitions: *Pointwise* is not useful because if $\theta$ is fixed then taking infimum over all estimators can do extremely well. Without the supremum, there is a deterministic estimator that does not look at the data and simple outputs $\text{argmin}_{\hat{\theta}} \ell(\hat{\theta}, \theta)$. The *Bayesian* characterization, where we replace the supremum with an expectation, is useful and in fact we will use it and draw connections with the Redundancy-Capacity Theorem studied earlier this semester.

For testing problems, we will let $\Theta$ be finite and let $\ell$ be the indicator function. Hence we define:

$$R(\Theta) = \inf_T \sup_{\theta \in \Theta} \mathbb{E}_\theta \big[ 1[T(X^n \neq \theta] \big] = \inf_T \sup_\theta \mathbb{P}_\theta [T \neq \theta] \tag{19.9}$$

### 19.3.1 Examples

- *Normal Means Testing:* Let $\Theta = \{-\mu, \mu\}$ and consider the probability of error. The goal now is to derive a test for determining the mean of the Gaussian. This is a simple-vs-simple hypothesis test.
- *Simple vs. Composite Normal Means:* The null hypothesis $H_0 : X_1^n \sim \mathcal{N}(0, I), x_i \in \mathbb{R}^d$, and the alternative is $H_1 : \mathcal{N}(\mu v, 1), \|v\| \geq 1, v \in \mathbb{R}^d$. This is a simple vs composite normal means problem and we will see how to get bounds here as well.
- Multiple Hypothesis Test $H_v : \mathcal{N}(\mu v, 1), v \in \{-1, 1\}^d$, so that there are $2^d$ hypotheses. We will see how to derive lower bounds for this type of testing problem as well.

## 19.4 Simple vs Simple

We first study simple versus simple testing problems. Let $P_0$ and $P_1$ be the two measures corresponding to the null and alternative hypotheses. We first have :

$$\inf_T \sup_{\theta \in 0,1} \mathbb{P}_\theta [T \neq \theta] \geq \inf_T \frac{1}{2} \mathbb{P}_0 [T \neq 0] + \frac{1}{2} \mathbb{P}_1 [T \neq 1] \tag{19.10}$$

We have replaced the supremum with an expectation. This is a general technique that we shall see over and over.

**Lemma 1** (Neyman-Pearson)**.** *For any distributions $P_0$ and $P_1$ over a space $\mathcal{X}$.*

$$\inf_T \{\mathbb{P}_0(T \neq 0) + \mathbb{P}_1(T \neq 1)\} = 1 - \|P_0 - P_1\|_{TV} \tag{19.11}$$

*where the infimum is over all deterministic mappings $T$.*

**Definition 2** (Total Variation Distance)**.** *The total variation distance between two measures is defined as:*

$$\|P_0 - P_1\|_{TV} = \sup_{A \subseteq \mathcal{X}} (P_1(A) - P_2(A)) = \frac{1}{2} \int |\frac{\partial P_0(x)}{\partial \mu(x)} - \frac{\partial P_1(x)}{\partial \mu(x)}| d\mu(x) = \frac{1}{2} \int |p_1(x) - p_0(x)| dx \tag{19.12}$$

*Proof.* Any deterministic test $T : \mathcal{X} \to \{0,1\}$ has an acceptance region $A = \{x \in \mathcal{X} : T(x) = 1\}$. Then

$$\mathbb{P}_0(T \neq 0) + \mathbb{P}_1(T \neq 1) = \mathbb{P}_0(A) + \mathbb{P}_1(A^c) = 1 - \mathbb{P}_1(A) + \mathbb{P}_0(A) \tag{19.13}$$

so

$$\inf_T\{\mathbb{P}_0(T \neq 0)+\mathbb{P}_1(T \neq 1)\} = \inf_A\{1-\mathbb{P}_1(A)+\mathbb{P}_0(A)\} = 1-\sup_A(\mathbb{P}_0(A)-\mathbb{P}_1(A)) = 1-\|P_1 - P_0\|_{TV} \tag{19.14}$$

□

For us this means that

$$\inf_T \sup_{\theta\in\{0,1\}} \mathbb{P}_{X_1^n\sim\theta}[T(X^n) \neq \theta] \geq \frac{1}{2} - \frac{1}{2}\|P_0^n - P_1^n\|_{TV} \tag{19.15}$$

Before turning to the first example, we need one more result which we have actually seen before:

**Lemma 3** (Pinsker's Inequality)**.** *For any distributions P,Q:*

$$\|P - Q\|_{TV}^2 \leq \frac{1}{2}KL(P,Q) \tag{19.16}$$

*Fact:* $KL(P^n, Q^n) = nKL(P;Q)$ where $P^n$ is the n-fold product measure of $P$

**Theorem 4** (KL-form of simple vs simple testing lower bound)**.**

$$\inf_T \sup_{\theta\in\{0,1\}} \mathbb{P}_{X_1^n\sim\theta}[T(X^n) \neq \theta] \geq \frac{1}{2} - \frac{1}{2}\sqrt{\frac{n}{2}KL(P_0||P_1)} \tag{19.17}$$

**Example 1** (Normal Means Testing)**.** *$P_0 = \mathcal{N}(-\mu, 1)$, $P_1 = \mathcal{N}(\mu, 1)$ and $\theta = \{0,1\}$ with $X_1^n \sim^{iid} P_\theta$ then $KL(P_0||P_1) = 2\mu^2$. This follows from the following*

$$KL(\mathcal{N}(\mu_0,\Sigma_0),\mathcal{N}(\mu_1,\Sigma_1)) = \frac{1}{2}\big[tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1-\mu_0)^T\Sigma_1^{-1}(\mu_1-\mu_0) - k + \log\frac{det\Sigma_1}{det\Sigma_0}\big] \tag{19.18}$$

*Hence we have*

$$\inf_T \sup_\theta \mathbb{P}[T(X^n) \neq \theta] \geq \frac{1}{2} - \frac{1}{2}\sqrt{n\mu^2} \tag{19.19}$$

*Thus, the probability of error is bounded from below by a constant $\frac{1}{2} - c$ if $\frac{1}{2}\sqrt{n\mu^2} \leq c$, i.e $\mu \leq \frac{2c}{n}$*

*As a sanity check, we know that thresholding the sample mean at 0 would give the same rate:*

$$\mathbb{P}[|\bar{X} - \mu| \geq \epsilon] \leq 2e^{-\frac{n\epsilon^2}{2}} \leq \delta \tag{19.20}$$

*This implies $\epsilon = \sqrt{\frac{2}{n}\log(\frac{1}{\delta})}$ so if $\mu \geq \epsilon$ we will succeed with probability of $1-\delta$.*