10-704: Information Processing and Learning **Fall 2016**

Lecture 20: Nov 9

Lecturer: Aarti Singh

Note: *These notes are based on scribed notes from Spring15 offering of this course. LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

20.1 Review - Privacy

Last time, we talked about use of noisy random projections to give average privacy and differential privacy guarantees. Summarizing it:

Privacy via Noisy Random projections
One popular way to privatize the data is to release only a random projection of the original data points possibly corrupted by noise. In this case, if the original data is X, the privatized data is $Y = AX + Z$, where $A \in \mathbb{R}^{m\times n}$ is a random matrix independent of X and Z.

Using multi-antenna channel capacity, we showed that the average mutual information between the privatized data Y and the original data X, over all possible input distributions $p(X)$ $\sup_{p(X)} \frac{I(X,Y)}{n} = O(\frac{m}{n})$. In many applications the number of random projections needed $m \ll n$ and hence the mutual information decreases as n increases. This can be viewed as an **average privacy guarantee via random projections**.

Differential Privacy via Noisy Random projections
Differential privacy is a stronger mathematical formalism for a privacy-preserving algorithm. We say an algorithm is (ϵ, δ)-differentially private if for all inputs X, X' differing in at most one value, and for all possible outcomes S:

$$Pr[\mathcal{A}(x) \in S] \leq e^{\epsilon} Pr[\mathcal{A}(x') \in S] + \delta \tag{20.1}$$

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

For noisy random projections, if Z has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries, then we can achieve (ϵ, δ) differential privacy as long as:

$$\sigma \geq (\max_j \|a_j\|_2) \frac{\sqrt{2(\log \frac{1}{2\delta} + \epsilon)}}{\epsilon} \tag{20.2}$$

where a_j are the columns of the matrix A. Adding Gaussian noise to preserve privacy is known as **Gaussian mechanism** and random projection further enhances the level of privacy as $E[\|a_j\|_2^2] = O(m/n)$ if A is an $m \times n$ random Gaussian matrix with $\mathcal{N}(0, 1/n)$ entries and $m \ll n$.

20.2 Privacy via Rate Distortion

A rate-distortion approach to privacy proceeds by keeping a utility function in mind such as an empirical loss $\widehat{R}_X(T) = \frac{1}{n}\sum_{i=1}^{n} loss_{X_i}(T)$. The goal is to minimize mutual information between T and X while ensuring some amount of utility, i.e.

$$\min_{p(T|X)} I(X;T) \text{ s.t } \mathbb{E}[\widehat{R}_X(T)] \leq \gamma \tag{20.3}$$

To find the rate-distortion function, the Blhaut-Arimoto algorithm is used which starting from some initial $p(T)$ iteratively updates

$$p(T|X) \propto p(T)e^{-\beta \widehat{R}_X(T)}$$

This is precisely the **exponential mechanism** for differential privacy which outputs a random T from $p(T|X)$ proportional to $e^{-\beta \widehat{R}_X(T)}$ and is known to preserve $(2\beta\Delta_{\ell_1}(\widehat{R}_X(T)), 0)$ differential privacy, where $\Delta_{\ell_1}(\widehat{R}_X(T)) = \max_{X \sim X'} |\widehat{R}_X(T) - \widehat{R}_{X'}(T)|_1$ and $X \sim X'$ denotes two inputs that differ in a single entry bounded by 1. To see this, consider

$$\frac{P(T \text{ is generated when } X \text{ is input})}{P(T \text{ is generated when } X' \text{ is input})} = \frac{e^{-\beta \widehat{R}_X(T)} / \sum_{T'} e^{-\beta \widehat{R}_X(T')}}{e^{-\beta \widehat{R}_{X'}(T)} / \sum_{T'} e^{-\beta \widehat{R}_{X'}(T')}} \le e^{2\beta\Delta_{\ell_1}(\widehat{R}_X(T))}.$$

We now switch to a different topic - fundamental limit of communication and learning. Recall that we had shown achievability of channel capacity via construction of a random code, but not the converse i.e. that no rate above capacity can be achieved. We now investigate information theoretic tools that allow us to quantify such limits of communication. The same tools will also be extended to quantify the limits of machine learning problems. A key tool is Fano's inequality.

20.3 Fano's Inequality

Suppose that we want to predict the sent code or channel input X from the received code or channel output Y. If $H(X|Y) = 0$, then intuitively, the probability of the error p_e should be 0. Fano's inequality characterizes this relation more precisely.

Theorem 1. *Suppose X is a random variable with finite outcomes in $\mathcal{X}$. Let $\widehat{X} = g(Y)$ be the predicted value of X for some deterministic function g that also takes values in $\mathcal{X}$. Then we have:*

$$p_e \equiv p(\widehat{X} \neq X) \geq \frac{H(X|Y) - 1}{\log |\mathcal{X}|}$$

Or, stated more strongly:

$$H(Ber(p_e)) + p_e \log(|\mathcal{X}| - 1) \geq H(X|Y)$$

where $Ber(p_e)$ refers to the bernoulli error random variable E with $Pr(E = 1) = p_e$.

Proof. Define random variable $E = \begin{cases} 1 \; if \; \widehat{X} \; \neq \; X \\ 0 \; else \end{cases}$

By the Chain rule, we have two ways of decomposing $H(E, X|Y)$:

$$H(E, X|Y) = H(X|Y) + H(E|X, Y)$$

$$H(E, X|Y) = H(E|Y) + H(X|E, Y)$$

Also, $H(E|X, Y) = 0$ since E is deterministic once we know the values of X and Y (and $g(Y)$). Thus we have that

$$H(X|Y) \leq H(Ber(p_e)) + H(X|E, Y)$$

To bound $H(X|E, Y)$, we use the definition of conditional entropy:

$$H(X|E, Y) = H(X|E = 0, Y)p(E = 0) + H(X|E = 1, Y)p(E = 1)$$

We will first note that $H(X|E=0,Y)=0$ since $E=0$ implies that $X=g(Y)$ and hence, if we observe both $E=0$ and $Y, X=g(Y)$ is no longer random. Also, $P(E=1)=p_e$.

Next, we note that $H(X|E=1,Y) \leq \log(|\mathcal{X}|-1)$. This is because if we observe $E=1$ and $g(Y)$, then X cannot be equal to $g(Y)$ and thus can take on at most $|\mathcal{X}|-1$ values.

Putting everything together, we have

$$H(Ber(p_e)) + p_e \log(|\mathcal{X}|-1) \geq H(X|Y)$$

as desired. □

Next, we will use Fano's inequality to characterize when reconstruction of a code sent over a channel is not possible, i.e. the probability of error is bounded away from zero. Similarly, Fano's inequality will be used to establish the fundamental limits of inference in machine learning problems by demonstrating when the probability of error of recovering the true model from data is bounded away from zero.

20.4 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 $W \to Y$,

$$Pr(\widehat{W}(Y) \neq W) \geq \frac{H(W|Y)-1}{\log|W|} \tag{20.4}$$

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

$$P(\widehat{W} \neq W) \geq \frac{H(W|Y)-1}{nR} = \frac{H(W) - I(W, Y^n) - 1}{nR} = \frac{nR - I(W, Y^n) - 1}{nR} \tag{20.5}$$

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

$$P(\widehat{W} \neq W) \geq \frac{nR - nC - 1}{nR} \tag{20.6}$$

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

Similarly, we will use Fano's inequality to lower bound the probability of error in hypothesis testing. Later, we will show that other learning problems can be reduced to hypothesis testing, enabling us to establish fundamental limits for other learning problems.

20.5 Minimax Theory For Testing Problems

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

$$\inf_T \sup_\theta \mathbb{E}_\theta \left[\ell(T(x_1^n), \theta)\right] \tag{20.7}$$

where T is an estimator, θ is some parameter and the inner term represents the risk.

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

What are alternative definitions: *Pointwise* is not useful because if θ 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 $\arg\min_{\widehat{\theta}} \ell(\widehat{\theta}, \theta)$. 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.

For testing problems, the goal is to identify a hypothesis θ amongst candidate hypothesis Θ and we will let the number of hypothesis $|\Theta|$ be finite and let ℓ be the indicator function. Hence we define:

$$R(\Theta) = \inf_T \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[1[T(X_1^n) \neq \theta]\right] = \inf_T \sup_\theta \mathbb{P}_\theta[T \neq \theta] \tag{20.8}$$

20.5.1 Examples

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

20.6 Simple vs Simple

We first study simple versus simple testing problems. For this case, Fano's inequality is too loose and we instead use **Le Cam's method** that we discuss below. Let P_0 and P_1 be the two measures corresponding to the null and alternative hypotheses. We first have :

$$\inf_T \sup_{\theta \in 0,1} \mathbb{P}_\theta[T \neq \theta] \geq \inf_T \frac{1}{2}\mathbb{P}_0[T \neq 0] + \frac{1}{2}\mathbb{P}_1[T \neq 1] \tag{20.9}$$

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

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

$$\|P_0 - P_1\|_{TV} = \sup_{A \subseteq \mathcal{X}} (P_1(A) - P_2(A)) = \frac{1}{2}\int |\frac{\partial P_0(x)}{\partial \mu(x)} - \frac{\partial P_1(x)}{\partial \mu(x)}| d\mu(x) = \frac{1}{2}\int |p_1(x) - p_0(x)| dx \tag{20.10}$$

The following lemma relates the probability of error to the total variation distance between the probability distributions associated with the two hypothesis.

Lemma 3. *For any distributions P_0 and P_1 over a space $\mathcal{X}$.*

$$\inf_T \{\mathbb{P}_0(T \neq 0) + \mathbb{P}_1(T \neq 1)\} = 1 - \|P_0 - P_1\|_{TV} \tag{20.11}$$

where the infimum is over all deterministic mappings T.

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

$$\mathbb{P}_0(T \neq 0) + \mathbb{P}_1(T \neq 1) = \mathbb{P}_0(A) + \mathbb{P}_1(A^c) = 1 - \mathbb{P}_1(A) + \mathbb{P}_0(A) \tag{20.12}$$

so

$$\inf_T \{\mathbb{P}_0(T \neq 0) + \mathbb{P}_1(T \neq 1)\} = \inf_A \{1 - \mathbb{P}_1(A) + \mathbb{P}_0(A)\} = 1 - \sup_A (\mathbb{P}_1(A) - \mathbb{P}_0(A)) = 1 - \|P_1 - P_0\|_{TV} \tag{20.13}$$

□

For us this means that

$$\inf_T \sup_{\theta \in \{0,1\}} \mathbb{P}_{X_1^n \sim \theta}[T(X_1^n) \neq \theta] \geq \frac{1}{2} - \frac{1}{2}\|P_0^n - P_1^n\|_{TV} \tag{20.14}$$

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

Lemma 4 (Pinsker's Inequality)**.** *For any distributions P,Q:*

$$\|P - Q\|_{TV}^2 \leq \frac{1}{2} KL(P, Q) \tag{20.15}$$

Also, the following fact will be useful.

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

Theorem 5 (KL-form of simple vs simple testing lower bound)**.**

$$\inf_T \sup_{\theta \in \{0,1\}} \mathbb{P}_{X_1^n \sim \theta}[T(X^n) \neq \theta] \geq \frac{1}{2} - \frac{1}{2}\sqrt{\frac{n}{2} KL(P_0 || P_1)} \tag{20.16}$$

We also state the KL-divergence between two d-dimensional Gaussians.

$$KL(\mathcal{N}(\mu_0, \Sigma_0), \mathcal{N}(\mu_1, \Sigma_1)) = \frac{1}{2}\big[tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T \Sigma_1^{-1}(\mu_1 - \mu_0) - d + \log \frac{det\Sigma_1}{det\Sigma_0}\big] \tag{20.17}$$

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

Hence we have

$$\inf_T \sup_\theta \mathbb{P}[T(X_1^n) \neq \theta] \geq \frac{1}{2} - \frac{1}{2}\sqrt{n\|\mu\|^2} \tag{20.18}$$

Thus, the probability of error is bounded from below by a constant if $\|\mu\| = O(1/\sqrt{n})$.

As a sanity check, consider $d = 1$. We can see that probability of error tending to 0 as $n \to \infty$ is achievable if $\mu = \omega(1/\sqrt{n})$. We consider the simple test which thresholds the sample mean at 0.

$$P_e = P_0(\bar{X} > 0) + P_1(\bar{X} < 0) = P_0(\bar{X} + \mu > \mu) + P_1(\bar{X} - \mu < -\mu) \leq 2e^{-n\mu^2/2} \tag{20.19}$$

where the last step follows from Gaussian tail bound. Thus, this test can achieve probability of error going to zero as $n \to \infty$ if $\mu = \omega(1/\sqrt{n})$.

There are two more ways to use Le Cam's method, as noted below.

1. We can always throw away parameters in the supremum and lower bound the risk:

$$\inf_T \sup_\Theta \mathbb{P}_\theta\left[\cdot\right] \geq \inf_T \sup_{\Theta' \subseteq \Theta} \mathbb{P}_\theta\left[\cdot\right].$$

   Any problem with $\mathbf{1}\{\cdot\}$ loss can be lower bounded by just choosing two parameters $\theta_0, \theta_1 \in \Theta$ and computing their TV or KL. While throwing away parameters always gives you a valid lower bound, the tightest lower bound is obtained by retaining the two parameters that are hardest to distinguish. Identifying the hardest examples is a little bit of an art. For example, such a trick can be useful when separating the hypothesis $H_0 : X_1^n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 I)$ from the composite alternate $H_1 : X_1^n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2 I)$ for any $\mu \geq \mu_0 > 0$. In this case, the hardest hypothesis to distinguish from the null is $X_1^n \overset{iid}{\sim} \mathcal{N}(\mu_0, \sigma^2 I)$ and hence one can characterize the error of any test by its performance on these two hypotheses.

2. We can also separate the parameter space into two regions and mix over these sets.

$$\begin{aligned}\inf_T \sup_\Theta \mathbb{P}_\theta\left[T(x) \neq \theta\right] &\geq \inf_T \sup_{j \in \{0,1\}} \sup_{\theta \in \Theta_j} \mathbb{P}_\theta\left[T(x) \neq j\right] \\ &\geq \inf_T \{\frac{1}{2}\mathbb{E}_{\theta \sim \pi_0, x \sim P_\theta}\left[\mathbf{1}\{T(x) \neq 0\}\right] + \frac{1}{2}\mathbb{E}_{\theta \sim \pi_1, x \sim P_\theta}\left[\mathbf{1}\{T(x) \neq 1\}\right]\} \\ &\geq \frac{1}{2} - \frac{1}{2}\left\|P_{\pi_0} - P_{\pi_1}\right\|_{TV},\end{aligned}$$

   where $P_{\pi_0}(A) = \mathbb{E}_{\theta \sim \pi_0}[P_\theta(A)]$, π_0 is a distribution on Θ_0, and π_1 is a distribution on Θ_1.

   This is important for some problems. By mixing you can make the distributions much closer together to prove stronger lower bounds. Often mixing based bounds are used when the hypotheses in a subset are equally hard to distinguish. For example, when null hypothesis is $H_0 : X_1^n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 I)$ and the alternate hypothesis is composite $H_1 : X_1^n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2 I)$ where μ is s-sparse vector with entries either μ_0 or 0. But it is often challenging to compute the divergence between mixtures.

In next class, we will talk about alternate ways to prove minimax lower bounds for multiple hypothesis testing and extend it to estimation.