Course: E2 209	Date **January 15th, 2019**

Lecture 3

Instructor: Himanshu Tyagi *Scribe: Avi Mohan*

Contents

1 Fano's inequality for M-ary hypothesis testing

We have already seen a lower bound for the probability of error for binary hypothesis testing problems (assuming a uniform prior on the two hypotheses):

$$\begin{aligned} P_e^*\left(\frac{1}{2},\frac{1}{2}\right) &\geq \frac{1}{2}\left(1-d(P,Q)\right) \\ &\geq \frac{1}{2}\left(1-\sqrt{\frac{1}{2}D(P \parallel Q)}\right). \end{aligned} \tag{1}$$

This bound allows us to quantify the difficulty of binary hypothesis tests in terms of the "distance" $D()$ between the distributions P and Q. Fano's inequality extends this to the case with $M(\geq 2)$ hypotheses.

Problem: Consider a collection of M hypotheses, $\mathscr{H}_i : X \sim P_i,\ \forall 1 \leq i \leq M$, where P_i is a measure on the space $(\mathcal{X}, \mathcal{F})$. Let $d : \mathcal{X} \to [M]$ be a (potentially randomized) map aka the hypothesis test, and assume as before, a uniform prior on $[M]$, i.e., each of the M hypotheses is chosen w.p. $\frac{1}{M}$. Let $\mathscr{P}(M)$ be the set of all probability measures on $[M]$, i.e., the $M-1$-dimensional simplex. Define the probability of error by

$$P_e^*(unif) := \inf_{d \in (\mathscr{P}(M))^{\mathcal{X}}} \frac{1}{M}\sum_{m=1}^{M} P_m\left(d(X) \neq m\right) \tag{2}$$

Theorem 1.1 (Fano)**.** With the probability of error defined as in (2), we have

$$P_e^*(unif) \geq 1 - \left[\frac{\frac{1}{M}\sum_{m=1}^{M} D\left(P_m \parallel \frac{1}{M}\sum_{i=1}^{M} P_i\right)+1}{\log M}\right] \tag{3}$$

$$\overset{(\dagger)}{\geq} 1 - \frac{\max_{m\neq m'} D(P_m \parallel P_{m'})+1}{\log M}. \tag{4}$$

Remark 1.1. 1. In case the (moderately awake) reader is wondering why the RHS of (3) is a valid probability, that is, why is $\frac{\frac{1}{M}\sum_{m=1}^{M} D\left(P_m \parallel \frac{1}{M}\sum_{i=1}^{M} P_i\right)+1}{\log M} \in [0,1]$: hold that thought - we'll explain this after the proof of the theorem.

2. The quantity $\frac{1}{M}\sum_{m=1}^{M} P_m$ on the RHS of (3) behaves like a "centroid" for the given set of probability distributions, and the numerator therefore, is a measure of the average distance of the set from its centroid.

3. Invoking the convexity of KL-divergence and Jensen's inequality, inequality (†), the RHS of (4) is derived from the RHS of (3) as follows:

$$\begin{aligned} D\left(P_m \,\|\, \frac{1}{M}\sum_{i=1}^{M} P_i\right) &\leq \frac{1}{M}\sum_{i=1}^{M} D\left(P_m \,\|\, P_i\right) \\ &\leq \max_{j\in[M]} D\left(P_m \,\|\, P_i\right). \end{aligned} \tag{5}$$

   Finally, one uses the fact that

$$\sum_{m=1}^{M} \max_{j\in[M]} D\left(P_m \,\|\, P_i\right) \leq \max_{m\neq m'} D(P_m \,\|\, P_{m'}) \tag{6}$$

   to get (4).

Before we proceed to the proof, we will require a new way to interpret the numerator on the RHS of (3).

1.1 A mutual information viewpoint

The mutual information between two random variables U and X with joint distribution P_{UX} and marginals P_U and P_X is defined as

$$I(U;X) := D(P_{UX} \,\|\, P_U P_X). \tag{7}$$

It is easy to show that $I(U;X) = H(U) - H(U|X)$, and quantifies the amount of information observing X gives about U. Obviously, $I(U;X) = 0$ if U and X are independent, which is in line with our intuition that now, X cannot tell us anything about U. Suppose $U \sim Unif([M])$, i.e., $P(U = i) = \frac{1}{M},\ i \in [M]$, and this random variable is transmitted through a channel W such that the output X has a distribution P_U. Then clearly, $P_X \equiv \frac{1}{M}\sum_{i=1}^{M} P_i$ and

$$\begin{aligned} I(U;X) &= D(P_{UX} \,\|\, P_U P_X) \\ &= \sum_{i=1}^{M}\sum_{y\in\mathcal{Y}} \frac{1}{M} P_i(y) \log \frac{\frac{1}{M}P_i(y)}{\frac{1}{M}P(y)} \\ &= \sum_{i=1}^{M}\sum_{y\in\mathcal{Y}} \frac{1}{M} P_i(y) \log \frac{\frac{1}{M}P_i(y)}{\frac{1}{M}\sum_{j=1}^{M}\frac{1}{M}P_j} \\ &= \sum_{i=1}^{M} D\left(P_i(y) \,\|\, \frac{1}{M}\sum_{j=1}^{M} P_j\right). \end{aligned} \tag{8}$$

Substituting this in (3), we get

$$\begin{aligned} P_e^*(unif) &\geq 1 - \left[\frac{I(U;X)+1}{\log M}\right] && (9) \\ &\geq 1 - \left[\frac{C(W)+1}{\log M}\right], && (10) \end{aligned}$$

where $C(W)$ is called the *Capacity* of the channel W and is defined as

$$C(W) := \max_{U\sim P\in\mathscr{P}([M])} I(U;X). \tag{11}$$

Recall that the channel was defined using a conditional distribution P_U, defined on the space $\mathcal{X}$ in which the output X of the channel takes values. In (11), the channel, i.e., this conditional is *fixed* and only the distribution of the input to the channel is varied. Since U takes values in $[M]$, its distributions come from the $(M-1)$-dimensional simplex of pmf's denoted by $\mathscr{P}([M])$.

Proof. Denote by $d : \mathcal{X} \to [M]$ the (possibly randomized) decision rule that outputs our guess of U upon observing X. Let $Q_{UX} \equiv P_U P_X$ be the product measure on $2^M \times \mathcal{F}$. Note that under this measure, U and X are independent. Since this rule need not be optimal, its error $P_e \geq P_e^*$. Now, consider the set $B \subset [M] \times \mathcal{X}$ over which d takes correct decisions, i.e., $B = \{(u,x) : d(x) = u\}$. Clearly,

$$\begin{aligned} P_{UX}(B) &= 1 - P_e \leq 1 - P_e^*, \text{ and} \\ Q_{UX}(B) &= Q_{UX}(d(X) = U) \\ &\overset{(*1)}{\leq} \frac{1}{M}. \end{aligned} \tag{12}$$

In inequality $(*1)$, we have used the fact that since U is uniformly distributed on $[M]$, and independent of X. Now, let $p = 1 - P_e$ and $q = Q_{UX}(d(X) = U)$ and a channel $W_B : [M] \times \mathcal{X} \to \{0,1\}$ such that $W_B(u,x) = \mathbb{I}_{\{d(x)=u\}} = \mathbb{I}_{\{(u,x)\in B\}}$. Observe that

$$I(U;X) = D\left(P_{UX} \parallel Q_{UX}\right) \underbrace{\geq}_{\text{data processing}} D\left(P_{UX}^{W_B} \parallel Q_{UX}^{W_B}\right) \tag{13}$$

However,

$$\begin{aligned} D\left(P_{UX}^{W_B} \parallel Q_{UX}^{W_B}\right) &= p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \\ &\overset{(*2)}{\geq} -h(p) + (1 - P_e) \log M \qquad (14) \\ &\geq -1 + (1 - P_e) \log M \qquad (15) \\ \Rightarrow I(U;X) &\geq D\left(P_{UX}^{W_B} \parallel Q_{UX}^{W_B}\right) \geq -1 + (1 - P_e) \log M, \\ \Rightarrow P_e^*(unif) &\geq 1 - \left[\frac{I(U;X) + 1}{\log M}\right]. \end{aligned}$$

where in $(*2)$, $h : [0,1] \to [0,1],\ h(p) := -p \log p - (1-p) \log 1 - p$, is the binary entropy function. We have used (12). □

Remark 1.2. 1. Strictly speaking, Fano's inequality should be

$$P_e^*(unif) \geq 1 - \left[\frac{I(U;X) + H(D)}{\log M}\right], \tag{16}$$

where $D := \mathbb{I}_{\{d(X)=U\}}$ indicates when the decision rule is correct. Now recall observation 1 in Rem. 1.1.

$$\begin{aligned} I(U;X) + H(D) &\leq \log M = H(U) \\ \iff H(U) - H(U|X) + H(D) &\leq H(U) \\ \iff H(\mathbb{I}_{\{d(X)=U\}}) &\leq H(U|X). \end{aligned}$$

But since this needs to be true regardless of the classifier d, it is easy to violate. For example, consider a binary hypothesis testing problem ($M = 2 \Rightarrow \log M = 1$) wherein $X \in \mathcal{X} = \{0,1\}$ and U is distributed uniformly over $\{0,1\}$. Also suppose that $P_0 \equiv Ber(0)$, $P_1 \equiv Ber(1)$, and a dumb detector with $d(X) = 0,\ w.p.\ 1/2$. Then $H(D) = 1$ while $H(U|X) = 0$, whereby,

$$\frac{H(U) - H(U|X) + H(D)}{\log M} = 2.$$

This means that Fano's inequality can, in fact, be vacuously true (sorry for the anti-climax).

2. Going back to 11, we see that

$$\begin{aligned} C(W) &= \max_{U\sim P\in\mathscr{P}([M])} I(U;X) \\ &= \max_{U\sim P\in\mathscr{P}([M])} \min_{Q_X\in\mathscr{P}(\mathcal{X})} D(P_{X|U} \parallel Q_X|P_U) \\ &\overset{(*3)}{=} \min_{Q_X} \max_{U\sim P\in\mathscr{P}([M])} D(P_{X|U} \parallel Q_X|P_U) \\ &\overset{(*4)}{=} \min_{Q_X} \max_{u\in[M]} D(P_{X|u} \parallel Q_X) \end{aligned} \tag{17}$$

where, $(*3)$ is true under certain regularity conditions that are satisfied here, and $(*4)$ follows from the fact that the maximum of a convex combination is attained by the distribution that puts all mass on the largest value. In the literature, the quantity $\min_{Q_X} \max_{u\in[M]} D(P_{X|u} \parallel Q_X)$ is sometimes called *information radius* [1].

3. We will frequently see that the difficulty of hypothesis testing and estimation problems can be stated in terms of the information radius and the number of hypotheses to be tested.

2 Example: Learning k-ary distributions

Let $P \in \mathscr{P}([k])$, the $(k-1)$-dimensional simplex and $X_1, \cdots, X_n$ be IID samples distributed P. We seek to estimate P from these samples, assuming k is known. Let $X^n := [X_1, , \cdots, X_n]$. A natural choice for an estimator for P is the empirical distribution $\hat{P}$ defined for every $x \in \mathcal{X}$ as

$$\hat{P}_x := \frac{1}{n}\sum_{i=1}^{N} \mathbb{I}_{\{X_i=x\}}. \tag{18}$$

Clearly, $\mathbb{E}\hat{P}_x = P_x, \ \forall x \in [k]$ and so, we have an unbiased estimator which, by the SLLN, is also strongly consistent. How well does it behave in the non-asymptotic regime?

$$\begin{aligned} \mathbb{E}_P d(P, \hat{P}_{X^n}) &= \mathbb{E}\left[\frac{1}{2}\sum_{x\in[k]} |P_x - \hat{P}_x|\right] \\ &\overset{Jensen}{\leq} \frac{1}{2}\sum_{x\in[k]} \sqrt{\left[\mathbb{E}\left(P_x - \hat{P}_x\right)^2\right]} \\ &= \frac{1}{2}\sum_{x\in[k]} \sqrt{\frac{1}{n^2}\mathbb{E}\left[\left(nP_x - \sum_{i=1}^{N}\mathbb{I}_{\{X_i=x\}}\right)^2\right]} \\ &\overset{(*5)}{=} \frac{1}{2\sqrt{n}}\sum_{x\in[k]} \sqrt{P_x(1-P_x)} \overset{C-S}{\leq} \frac{1}{2\sqrt{n}}\sqrt{k\sum_{x\in[k]} P_x} \\ &= \frac{1}{2}\sqrt{\frac{k}{n}}, \\ &\leq \epsilon, \ \forall n \geq \frac{k}{4\epsilon^2}. \end{aligned} \tag{19}$$

In $(*5)$, we have used the fact that the variance of a $\text{Bin}(n,p)$ random variable is $np(1-p)$. This, once again, shows that sample complexity is proportional to ϵ^{-2}. Note, however, that this derivation heavily depends on the IID nature of the samples. In the next lecture, we will see a more powerful method that, to a certain extent, does not require IID sampling.

3 Minimax and Probably Approximately Correct (PAC) formulations

We will focus on two different (but related) formulations to establish the efficacy of estimators/classifiers.

1. **Minimax formulation:** Given IID samples from some distribution $P \in \mathscr{P}(\mathcal{X})$ on $(\mathcal{X}, \mathcal{F})$, and an estimator $\hat{P} : \mathcal{X}^n \to \mathscr{P}(\mathcal{X}^n)$, i.e., $x^n \mapsto \hat{P}(x^n)$, the minimax risk is defined as

$$\begin{aligned} R(n,k) &:= \min_{\hat{P}} \max_{P \in \mathscr{P}(\mathcal{X})} \mathbb{E}_P d(P, \hat{P}_{X^n}) \\ &= \min_{\hat{P}} \max_{\pi \in \mathscr{P}(\mathscr{P}(\mathcal{X}))} \mathbb{E}_{P \sim \pi} \left[\mathbb{E}_P d(P, \hat{P}_{X^n}) \right] \\ &\overset{(*6)}{=} \max_{\pi \in \mathscr{P}(\mathscr{P}(\mathcal{X}))} \min_{\hat{P}} \mathbb{E}_{P \sim \pi} \left[\mathbb{E}_P d(P, \hat{P}_{X^n}) \right], \end{aligned}$$

   where equality $(*6)$ is true under certain regularity conditions and helps with analysis.

2. (ϵ, δ)**-PAC formulation:** Given the space of k-ary distributions,

$$n(\epsilon, \delta, k) := \min \left\{ n \geq 1 : \exists\ \hat{P}\ s.t.\ \max_{P \in \mathscr{P}(\mathcal{X})} P\left(d(P, \hat{P}_{X^n}) > \epsilon \right) \leq \delta \right\}. \tag{20}$$

   We will freeze $\delta = \frac{1}{3}$. One can use either Markov's inequality or a Chernoff-Hoeffding bound to transition between the two formulations.

Preview of Lecture 4:

- ⋆ Having studied Fano's inequality and the two risk formulations, we will first look at Fano's bound for minimax risk.
- ⋆ This will give a clearer picture of how information radius affects the performance of classifiers and estimators.

References

[1] I. Csiszár, P. C. Shields *et al.*, "Information theory and statistics: A tutorial," *Foundations and Trends® in Communications and Information Theory*, vol. 1, no. 4, pp. 417–528, 2004.