Advanced Topics in Artificial Intelligence: EE6180
Indian Institute of Technology Madras — August 6, 2019
Instructor: Abhishek Sinha

Problem Set 1

- This problem set is due on **August 20, 2019** in the class.
- Each problem carries 10 points.
- You may work on the problems in groups of size at most **two**. However, **each student must write their own solution**. If you work together on the problems, clearly mention the name of your collaborator.

---

1. **(Relationship between the f-divergences)** Bounding the Total-Variation distance from above is the first step in proving lower bounds for many testing and estimation problems. In this exercise, we will derive two inequalities which will help us in this task.

   (a) Recall that $d(p||q) \equiv D(\text{Bern}(p)||\text{Bern}(q))$ denotes the binary divergence function:

   $$d(p||q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}.$$

   i. Prove for all $p, q \in [0, 1]$

   $$d(p||q) \geq 2(p-q)^2 \log e.$$

   Hint: Use Taylor's series expansion.

   ii. Apply data processing inequality to prove *Pinsker's inequality:*

   $$\text{TV}(P, Q) \leq \sqrt{\frac{1}{2 \log e} D(P||Q)},$$

   where $\text{TV}(P, Q)$ is the *total variation* distance between probability distribution P and Q:

   $$\text{TV}(P, Q) \stackrel{\Delta}{=} \sup_E (P(E) - Q(E)),$$

   with the supremum taken over all events E.

   (b) Recall that the χ^2 divergence between two (discrete) probability distributions P and Q is defined as $D_{\chi^2}(P||Q) = \frac{1}{2}\mathbb{E}_Q\big(\frac{P}{Q} - 1\big)^2$. Show that

   $$D(P||Q) \leq \log\big(1 + D_{\chi^2}(P||Q)\big).$$

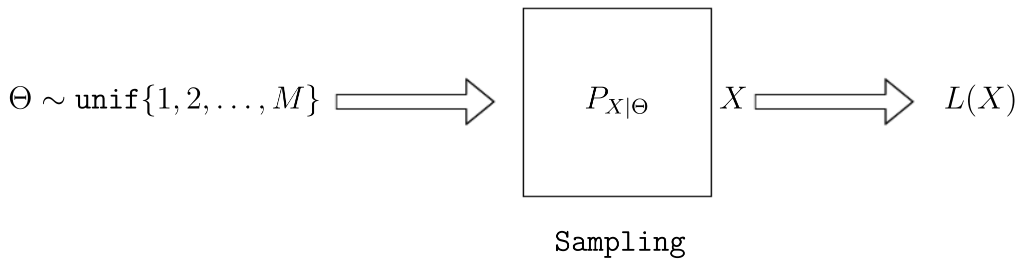

Figure 1: List Decoding

2. **(Fano's inequality for List Decoding)** List decoding is a relaxation to the usual unique decoding where we only require the true distribution to be in a predicted *output list* L of size $\mathcal{L} \geq 1$, in order to declare the decoding to be successful (see Figure 1). In this problem, we will derive a Fano-like lower-bound for the probability of unsuccessful decoding.

   Formally, let Θ be a uniformly chosen random index which may take M different values corresponding to the distributions $\{P_\theta\}_{\theta=1}^M$. The random variable X is sampled according to the conditional probability distribution $P_{X|\Theta} \equiv P_\Theta$. Upon observing X, the decoder outputs a list $L(X)$ comprising of $\mathcal{L}$ predicted indices (*i.e.*, $L(X) \subset \{1, 2, \ldots, M\}, |L| = \mathcal{L}$). We consider the decoding to be unsuccessful if $\Theta \notin L(X)$.

   Let $p_e \equiv \mathbb{P}(\Theta \notin L(X))$ be the probability that the decoding is unsuccessful. Prove that

$$p_e \geq \frac{H(\Theta|X) - 1 - \log(\mathcal{L})}{\log(\frac{M}{\mathcal{L}} - 1)}.$$

   Note that, the above lower-bound reduces to the usual Fano's lower bound when we put $\mathcal{L} = 1$.

3. **(Tensorization of Total Variation)** Using the coupling characterization of total variation discussed in the class, show that

$$d_{\texttt{TV}}(\prod_{i=1}^k P_i, \prod_{i=1}^k Q_i) \leq \sum_{i=1}^k d_{\texttt{TV}}(P_i, Q_i).$$

4. **(Gaussian Maxima)** Let $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ be the density function of a standard normal $Z \sim \mathcal{N}(0, 1)$ variate.

   (a) Show that $\phi'(z) + z\phi(z) = 0$.

(b) Use part (a) to show that

$$\phi(z)\left(\frac{1}{z}-\frac{1}{z^3}\right) \le \mathbb{P}[Z \ge z] \le \min\{\frac{\phi(z)}{z}, \phi(z)\left(\frac{1}{z}-\frac{1}{z^3}+\frac{3}{z^5}\right)\}.$$

HINT: Integration by parts!

(c) Let $\{X_i\}_{i=1}^n$ be an i.i.d. sequence of Gaussian $\mathcal{N}(0,\sigma^2)$ variables. Define $Z_n \equiv \max_{i=1}^n |X_i|$. Prove that

$$\mathbb{E}[Z_n] \le \sqrt{2\sigma^2 \log 2n}, \quad \forall n \ge 2.$$

HINT: Derive an upper bound for the MGF of Z_n and use Jensen's inequality.

(d) Using (b) and (c) or otherwise, show that

$$\lim_{n\to\infty} \frac{\mathbb{E}[Z_n]}{\sqrt{2\sigma^2 \log n}} = 1.$$

HINT: Simplify lower bound on $\mathbb{P}(Z \ge z)$ from part (a) for an appropriate z and use Markov's inequality for lower bounding the expectation.

5. **(Information Inequality)** Let $\{X_i, i = 1, 2, \ldots, d\}$ be independent random variables. Prove that $I(X_i; X_i + X_j) \ge I(X_i; X_i + X_j + X_k)$.