Information and Coding Theory	**Autumn 2022**

Homework 2

Due: November 1, 2022

Note: *You may discuss these problems in groups. However, you must write up your own solutions and mention the names of the people in your group. Also, please do mention any books, papers or other sources you refer to. It is recommended that you typeset your solutions in LaTeX.*

1. **Biased coins strike back.** **[3 + 3 = 6 points]**

   In class we considered the problem of distinguishing coins distributed according to the following two distributions:

   $$P = \begin{cases} 1 & \text{w.p. } \frac{1}{2} - \varepsilon \\ 0 & \text{w.p. } \frac{1}{2} + \varepsilon \end{cases} \quad \text{and} \quad Q = \begin{cases} 1 & \text{w.p. } \frac{1}{2} \\ 0 & \text{w.p. } \frac{1}{2} \end{cases}$$

   We derived matching upper and lower bounds (up to constants) of the form $\Theta(1/\varepsilon^2)$ on the number of coin tosses required to distinguish the two distributions. Consider now the problem of distinguishing two extremely biased coins with slightly differing biases:

   $$P' = \begin{cases} 1 & \text{w.p. } \varepsilon \\ 0 & \text{w.p. } 1 - \varepsilon \end{cases} \quad \text{and} \quad Q' = \begin{cases} 1 & \text{w.p. } 2\varepsilon \\ 0 & \text{w.p. } 1 - 2\varepsilon \end{cases}$$

   Find tight upper and lower bounds (up to constants) on the number of independent coin tosses required to distinguish coins distributed according to P' and Q'.

2. **Jensen-Shannon divergence.** **[2 + 3 + 4 + 3 = 12 points]**

   While KL-divergence is sometimes used as a measure of the difference between two distributions, it is asymmetric and can be infinite. In some applications, one can instead consider the Jensen-Shanon divergence which addresses these issues.

   (a) For two distributions P and Q, we define the Jensen-Shannon divergence as

   $$\mathrm{JSD}(P, Q) := \frac{1}{2} \cdot D(P \| M) + \frac{1}{2} \cdot D(Q \| M) \quad \text{where} \quad M = \frac{P + Q}{2}.$$

   Show that $0 \le \mathrm{JSD}(P, Q) \le 1$.

   (b) Show that $\mathrm{JSD}(P, Q) \ge \frac{1}{8 \ln 2} \cdot \|P - Q\|_1^2$.

(c) Show that $\mathrm{JSD}(P, Q) \ \le\ \frac{1}{2} \cdot \|P - Q\|_1$.

(d) The notion of Jensen-Shannon divergence can be generalized to an arbitrary number of distributions and an arbitrary convex combination. Let $P_1, \ldots, P_k$ be distributions on the same universe and let $\lambda = (\lambda_1, \ldots, \lambda_k)$ be a tuple of non-negative weights such that $\sum_i \lambda_i = 1$. We define

$$\mathrm{JSD}_\lambda(P_1, \ldots, P_k) \ :=\ \sum_i \lambda_i \cdot D\left(P_i \| M\right) \quad \text{where} \quad M = \sum_i \lambda_i P_i\,.$$

Show that $0 \le \mathrm{JSD}_\lambda(P_1, \ldots, P_k) \le H(\lambda)$, where $H(\lambda)$ denotes the entropy of λ, when viewed as a distribution over $[k]$.

3. **Counting using method of types (Problem 11.5 from the book).** **[5 points]**

Let $\mathcal{X}$ be a finite universe with $|\mathcal{X}| = r$ and let $g : \mathcal{X} \to \mathbb{R}$ be a real valued function. Let $S \subseteq \mathcal{X}^n$ be the set of sequences $x_1, \ldots, x_n$ with each $x_i \in \mathcal{X}$ defined as

$$S \ =\ \left\{ (x_1, \ldots, x_n) \in \mathcal{X}^n \ \middle|\ \frac{1}{n} \sum_{i=1}^{n} g(x_i) \ge \alpha \right\}.$$

Let $\Pi = \{P \ |\ \sum_{a \in \mathcal{X}} P(a) g(a) \ge \alpha\}$. Show that

$$|S| \ \le\ (n+1)^r \cdot 2^{nH^*}\,,$$

where $H^* = \sup_{P \in \Pi} H(P)$.

4. **Differential entropy of a Gaussian.** **[2 + 3 = 5 points]**

We saw in class that if the differential entropy $h(X)$ exists for a continuous random variable X taking values in $\mathbb{R}^n$, and $A \in \mathbb{R}^{n \times n}$ is a non-singular matrix, then

$$h(AX) \ =\ h(X) + \log|A|\,,$$

where $|A|$ denotes |Det(A)|. We can use this to compute the entropy of a Gaussian random variable.

(a) Let $X \sim N(\mu, \Sigma)$ be an n-dimensional Gaussian with mean μ and covariance matrix Σ i.e.,

$$\mathbb{E}\left[X\right] = \mu \qquad \text{and} \qquad \mathbb{E}\left[(X-\mu)(X-\mu)^\mathsf{T}\right] = \Sigma\,.$$

Assume that the covariance matrix Σ is *positive definite* and hence there exists a non-singular matrix R such that $\Sigma = R^2$. Use this to show that

$$h(X) \ =\ \frac{n}{2} \cdot \log(2\pi e) + \frac{1}{2} \cdot \log|\Sigma|\,.$$

(b) Use the above to show that for any two positive definite matrices Σ_1 and Σ_2, and $\alpha \in [0,1]$, we have

$$|\alpha \cdot \Sigma_1 + (1-\alpha) \cdot \Sigma_2| \;\geq\; |\Sigma_1|^{\alpha} \cdot |\Sigma_2|^{1-\alpha} \,.$$

5. **Dual definition of KL-diveregnce** **[6+6 = 12 Points]**

Let P, Q be two distributions supported on a finite universe $\mathcal{X}$. In class, we defined the KL-divergence $D(P\|Q)$ between P and Q as

$$D(P\|Q) \;=\; \sum_{x \in U} P(x) \log \frac{P(x)}{Q(x)},$$

but there is an alternate definition known as the Donsker-Varadhan variational representation where

$$D(P\|Q) \;=\; \sup_{f:\mathcal{X}\to\mathbb{R}_{>0}} \mathbb{E}_{x\sim P}[\log f(x)] - \log(\mathbb{E}_{x\sim Q} f(x)).$$

(a) In the first part of this problem, we will prove one side of this equality. In particular, we would like to show that for any $f : \mathcal{X} \to \mathbb{R}_{>0}$ (i.e., taking only positive values),

$$D(P\|Q) \;\geq\; \mathbb{E}_{x\sim P}[\log f(x)] - \log(\mathbb{E}_{x\sim Q} f(x))\,.$$

Observe that, without loss of generality, it suffices to consider the case where $\mathbb{E}_{x\sim Q} f(x) = 1$, since we can always rescale $f(x)$ to $\tilde{f}(x) = \frac{f(x)}{\mathbb{E}_{x\sim Q} f(x)}$. Thus, prove the following: for all functions $f : \mathcal{X} \to \mathbb{R}_{>0}$ satisfying $\mathbb{E}_{x\sim Q}[f(x)] = 1$, we have

$$\mathbb{E}_{x\sim P}[\log f(x)] \;\leq\; D(P\|Q)\,.$$

(b) We will now see that the above property can be used to prove a "Pinsker-like" inequality for "Gaussian-like" random variables, which may not necessarily be bounded in absolute value. A random variable Z with mean μ is said to be σ-subgaussian if it satisfies $\mathbb{E}e^{\lambda(Z-\mu)} \leq e^{\lambda^2\sigma^2/2}$ $\forall \lambda \in \mathbb{R}$. This notion is useful because it captures random variables that enjoy some of the properties of Gaussian random variables (you can check that the inequality holds for Gaussians).

Let $g : \mathcal{X} \to \mathbb{R}$ be such that $g(X)$ is σ-subgaussian when X has the distribution Q. Show that

$$\left| \mathop{\mathbb{E}}_{x\sim P}[g(x)] - \mathop{\mathbb{E}}_{x\sim Q}[g(x)] \right| \;\leq\; \sqrt{2\ln 2 \cdot \sigma^2 \cdot D(P\|Q)}\,.$$

Hint: Apply the inequality from part (a) on an appropriately chosen $\tilde{g}$ function defined in terms of g. Use the subgaussianity property, and then optimize λ.

Note that this inequality is qualitatively similar to what we proved in class (Lecture 6). If g was bounded in absolute value, then the LHS could be bound in terms of the total variation distance, and then use Pinsker's inequality. The key difference here is that $g(X)$ is not necessarily bounded, but subgaussian (when X is distributed according to Q).

6. **Extra problem (no need to submit): Chernoff bound for read-k families.**

We used Sanov's theorem to derive the Chernoff bound for independent random variables $X_1, \ldots, X_n$ taking values uniformly in $\{0, 1\}$. In particular, we showed that

$$\mathbb{P}\left[X_1 + \cdots + X_n \geq \left(\frac{1}{2} + \varepsilon\right) n\right] \;\leq\; (n+1) \cdot 2^{-n \cdot D\left(\frac{1}{2}+\varepsilon \| \frac{1}{2}\right)},$$

where $D\left(\frac{1}{2} + \varepsilon \| \frac{1}{2}\right)$ denotes the KL-divergence of two distributions on $\{0, 1\}$, with probabilities $(\frac{1}{2} + \varepsilon, \frac{1}{2} - \varepsilon)$ and $(\frac{1}{2}, \frac{1}{2})$. In this problem, we will consider functions $f_1, \ldots, f_r$ depending on the variables $X_1, \ldots, X_n$ and prove a concentration bound on the expression $f_1 + \cdots + f_r$.

Let $S_1, \ldots, S_r$ be subsets of $[n]$ for each $i \in [r]$, let $f_i : \{0, 1\}^{S_i} \to \{0, 1\}$ be a function which depends only on the variables in S_i. We use the shorthand X_{S_i} to denote the variables $\{X_j\}_{j \in S_i}$. Moreover, we have the property that each variable is involved in only k functions i.e., $\forall j \in [n], \; |\{i \in [r] \mid j \in S_i\}| = k$. Such a family of functions is called a read-k family (it is not too hard to see that the lower bound extends to the case when each variable is in *at most* k functions).

(a) Recall that for two random variables Z_1 and Z_2 distributed on *same universe* $\mathcal{Z}$ with distributions P_1 and P_2, we also use $D(Z_1 \| Z_2)$ to mean $D(P_1 \| P_2)$. Let $Y_1, \ldots, Y_n$ be (not necessarily independent) random variables jointly distributed on $\{0, 1\}^n$ and let $X_1, \ldots, X_n$ be random variables as above, distributed uniformly and independently on $\{0, 1\}^n$. Let the sets $\{S_i\}_{i \in [r]}$ be as above. Use Shearer's lemma to show that

$$k \cdot D(Y_1, \ldots, Y_n \| X_1, \ldots, X_n) \;\geq\; \sum_{i \in [r]} D\left(Y_{S_i} \| X_{S_i}\right).$$

(b) Let $A = \left\{(a_1, \ldots, a_n) \in \{0, 1\}^n \;\middle|\; \sum_{i \in [r]} f_i\left(\{a_j\}_{j \in S_i}\right) \geq t\right\}$. Let $(Y_1, \ldots, Y_n)$ be uniformly distributed over the set A (note that $Y_1, \ldots, Y_n$ are not necessarily independent). Prove that

$$\mathop{\mathbb{P}}_{X_1, \ldots, X_n}\left[\sum_{i \in [r]} f_i(X_{S_i}) \geq t\right] \;=\; 2^{-D(Y_1, \ldots, Y_n \| X_1, \ldots, X_n)},$$

where the probability is over the uniform distribution for $X_1, \ldots, X_n$.

(c) For each $i \in [r]$, let $\mathbb{E}\left[f_i\left(X_{S_i}\right)\right] = \mu_i$ and $\mathbb{E}\left[f_i\left(Y_{S_i}\right)\right] = \nu_i$. Prove that

$$D\left(Y_{S_i} \| X_{S_i}\right) \;\geq\; D\left(\nu_i \| \mu_i\right) \,,$$

where $D\left(\nu_i \| \mu_i\right)$ denotes the divergence of two distributions on $\{0,1\}$ with probabilities $(\nu_i, 1-\nu_i)$ and $(\mu_i, 1-\mu_i)$.

(d) Use the above bounds and the convexity of KL-divergence in both its arguments to show that for $\mu = \frac{1}{r} \cdot (\mu_1 + \cdots + \mu_r)$,

$$\mathop{\mathbb{P}}_{X_1,\ldots,X_n}\left[f_1(X_{S_1}) + \cdots + f_r(X_{S_r}) \geq (\mu + \varepsilon) \cdot r\right] \;\leq\; 2^{-(r/k) \cdot D(\mu+\varepsilon \| \mu)} \,.$$