36-782: Homework 1
Due on 09/21/2023

1. Properties of information measures.

(a) Suppose an urn contains r red, w white, and b black balls. Let $X_1, \ldots, X_k$ denote k draws from the urn with replacement, and $Y_1, \ldots, Y_k$ denote k draws from the urn without replacement. What is the relation between $H(X_1, \ldots, X_k)$ and $H(Y_1, \ldots, Y_k)$? Give a formal argument.
Hint: what is the marginal distribution of Y_j, for $1 \leq j \leq k$? use this distribution along with "conditioning reduces entropy".

(b) For some $n \geq 2$, let $X^n = (X_1, \ldots, X_n)$ denote a random variable on $\mathcal{X}^n$ with joint distribution $Q \equiv Q_{X^n}$, and let $Y_1, \ldots, Y_n$ denote n independent $\mathcal{X}$-valued random variables with joint distribution $P \equiv \prod_{i=1}^n P_{Y_i}$. For any $i \in [n]$, we use $Q^{(i)}$ and $P^{(i)}$ to denote the distributions of $X^{(i)} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ and $Y^{(i)} = (Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n)$ respectively. Then, prove the following inequality:

$$D_{\mathrm{kl}}(Q \parallel P) \geq \frac{1}{n-1} \sum_{i=1}^n D_{\mathrm{kl}}(Q^{(i)} \parallel P^{(i)}).$$

Hint: start with Han's inequality for entropy of Q. Then, use the fact that $D(Q \parallel P) = -H(Q) + \sum_{x^n} q(x^n) \log(1/p(x^n))$.

(c) Suppose X is an $\mathcal{X}$-valued random variable, with $|\mathcal{X}| = m$. Let $\pi : \mathcal{X} \to \mathcal{X}$ denote a random bijection, drawn independently of X (i.e., drawn from the set of $m!$ possible bijections, or permutations of the elements of $\mathcal{X}$). Then, show that $H(\pi X) \geq H(X)$.

(3 + 4 + 3 points)

2. Entropy of stationary processes. Let $\{X_n : n \in \mathbb{N}\}$ denote an $\mathcal{X}$-valued stationary stochastic process, with $|\mathcal{X}| < \infty$. Recall that the distribution of a stationary process is shift-invariant: that is, for all $j, k \in \mathbb{N}$, we have

$$P(X_{i_1} = x_1, X_{i_2} = x_2, \ldots, X_{i_j} = x_j) = P(X_{i_1+k} = x_1, X_{i_2+k} = x_2, \ldots, X_{i_j+k} = x_j).$$

(a) Show that the following is true:

$$\lim_{n \to \infty} \frac{H(X_1, X_2, \ldots, X_n)}{n} = \lim_{n \to \infty} H(X_n | X_1, \ldots, X_{n-1}).$$

Hint: show that $H(X_n|X^{n-1})$ is nonincreasing, and use it to argue that $\lim_{n\to\infty} H(X_n|X^{n-1})$ exists. Next, use the fact that if a real-valued sequence $(a_n)_{n\geq 1}$ converges to a, then so does the sequence $(b_n)_{n\geq 1}$, with $b_n = (1/n)\sum_{i=1}^n a_i$, to show the equality.

(b) Which one is larger; $\lim_{n\to\infty} H(X^n)/n$ or $H(X_1)$? Why?

(c) What is the value of $\lim_{n\to\infty} \frac{1}{n} I(X^n; X_{n+1}^{2n})$?

(d) Let $\mathcal{Y}$ denote another finite alphabet, and suppose $\varphi : \mathcal{X} \to \mathcal{Y}$ is a (deterministic) mapping from $\mathcal{X}$ to $\mathcal{Y}$. For any $i \in \mathbb{N}$, let Y_i denote the random variable $\varphi(X_i)$. Then, show that

$$\lim_{n\to\infty} \frac{H(Y^n)}{n} \leq \lim_{n\to\infty} \frac{H(X^n)}{n}.$$

(3 + 1 + 2 + 4 points)

3. The method of types. Let $\mathcal{X} = \{1, 2, \ldots, m\}$, and for any sequence $\boldsymbol{x}^n = (x_1, \ldots, x_n) \in \mathcal{X}^n$, we define the "type" of $\boldsymbol{x}^n$ as $\widehat{P}_{\boldsymbol{x}^n} = (n_1/n, n_2/n, \ldots, n_m/n) \in [0,1]^m$, where $n_i \equiv n_i(\boldsymbol{x}^n) = \sum_{j=1}^n \mathbf{1}_{x_j = i}$. In other words, the type of $\boldsymbol{x}^n$ is simply the empirical probability distribution defined by the observations $\boldsymbol{x}^n = (x_1, \ldots, x_n)$ on the alphabet $\mathcal{X}$.

(a) Let $\mathcal{P}_n$ denote the set of all possible types with denominator n (that is, constructed using sequences $\boldsymbol{x}^n$ of length n). Then, what is $|\mathcal{P}_n|$? Express your answer as a binomial coefficient.

(b) Show that $|\mathcal{P}_n| \leq (n+1)^m$. That is, the number of distinct types grows polynomially with n.

(c) Let $X_1, X_2, \ldots, X_n$ be i.i.d. draws of an $\mathcal{X}$-valued random variable with a distribution Q. Then, show that for any $\boldsymbol{x}^n \in \mathcal{X}^n$, we have

$$Q^n(X^n = \boldsymbol{x}^n) = 2^{-n\left(H(\widehat{P}_{\boldsymbol{x}^n}) + D_{\mathrm{kl}}(\widehat{P}_{\boldsymbol{x}^n} \| Q)\right)}.$$

(d) For any (non-random) $\widehat{P} \in \mathcal{P}_n$ (i.e., a possible empirical distribution with denominator n), let $T(\widehat{P}) \subset \mathcal{X}^n$ denote all sequences $\boldsymbol{x}^n$ of length n with type $\widehat{P}$. For example, if $\mathcal{X} = \{1, 2\}$, and $\widehat{P} = (1/3, 2/3)$, then for $n = 3$, we have $T(\widehat{P}) = \{(1,2,2), (2,1,2), (2,2,1)\}$. Using the result of part (*c*), show that

$$|T(\widehat{P})| \leq 2^{nH(\widehat{P})}.$$

Note that it is also possible to show that $|T(\widehat{P})| \geq 2^{nH(\widehat{P})}/|\mathcal{P}_n|$, which can be further lower bounded, due to (*b*), by $2^{nH(\widehat{P})}(n+1)^{-m}$.

(e) Let Δ_m denote all pmfs on $\mathcal{X}$, and let E denote any closed subset of Δ_m. For $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} Q$, let $\widehat{P}_{X^n}$ denote the empirical distribution of X^n. Using the notation $Q^n(E)$ to denote $Q^n(P_{X^n} \in E)$, show that

$$Q^n(E) \leq (n+1)^m 2^{-nD_{\mathrm{kl}}(P^* \| Q)}, \quad \text{where} \quad P^* := \operatorname*{arg\,min}_{P \in E} D_{\mathrm{kl}}(P \parallel Q).$$

Use this to conclude that

$$\liminf_{n\to\infty} -\frac{1}{n} \log\left(Q^n(E)\right) \leq D_{\mathrm{kl}}(P^* \parallel Q).$$

(2 + 1 + 2 + 2 + 3 points)

4. Counting via entropy. Let $\mathcal{X}$ denote a finite alphabet, and for some $n \in \mathbb{N}$, let $\mathcal{F}_N$ denote a collection of subsets of $[n]$ satisfying the property that

$$|\{E \in \mathcal{F} : i \in E\}| \geq N, \quad \text{for all } i \in [n].$$

In words, each $i \in [n]$ appears in at least N distinct subsets of $[n]$ contained in $\mathcal{F}_N$.

(a) Let $X_1, X_2, \ldots, X_n$ denote $\mathcal{X}$ valued random variables. Then, show that

$$H(X_1, \ldots, X_n) \leq \frac{1}{N} \sum_{E \in \mathcal{F}_N} H(X_E), \tag{1}$$

where we use X_E to denote $\{X_i : i \in E\}$.
Hint: Consider any set $E \in \mathcal{F}_N$, and use chain rule to expand $H(X_E)$. Then, lower bound this quantity by conditioning on additional terms. Finally, sum over all $E \in \mathcal{F}_N$, and use the defining property of $\mathcal{F}_N$, to lower bound the sum with $NH(X^n)$.

(b) Show that by suitable choices of the class $\mathcal{F}_N$, we can recover (i) the result that entropy is subadditive (i.e., $H(X^n) \leq \sum_{i=1}^n H(X_i)$), and (ii) Han's inequality for entropy, from (1).

(c) Let $S_n \subset \mathbb{R}^3$ denote n distinct points in a three dimensional euclidean space. Suppose these points have n_1 distinct projections on the XY plane, n_2 distinct projections on the YZ plane, and n_3 distinct projections on the ZX plane. Then, use (1) to show that

$$n^2 \leq n_1 n_2 n_3.$$

(d) Generalize the result of part (*c*) to arbitrary dimensions $d \geq 3$. Formally, let $S_n \subset \mathbb{R}^d$ denote n points in $\mathbb{R}^d$. Suppose the projection of S_n along the i^{th} coordinate (i.e., along the hyperplane normal to the i^{th} coordinate axis) has n_i distinct points. Then, we have

$$n^{d-1} \leq \prod_{i=1}^{d} n_i.$$

(3 + 1 + 4 + 2 points)

5. Generalized Fano's inequality for statistical applications. In this problem, we will derive a general form of Fano's inequality that is useful in obtaining minimax lower bounds in various statistical problems.

Let $\mathcal{P}(\mathcal{X})$ denote a class of probability distributions on a finite alphabet $\mathcal{X}$, and let Θ denote a space of parameters, with an associated pseudo-metric $d : \Theta \times \Theta \to [0, \infty)$. Let $\theta : \mathcal{P}(\mathcal{X}) \to \Theta$ denote a mapping, that assigns a parameter in Θ to each distribution in $\mathcal{P}(\mathcal{X})$. Consider r distributions $P_1, \ldots, P_r \in \mathcal{P}(\mathcal{X})$, and introduce $\theta_i = \theta(P_i)$ for $i \in [r]$. Assume the following two statements hold (for some constants α, β):

$$d(\theta_i, \theta_j) \geq \alpha, \text{ for all } i \neq j, \quad \text{and} \quad D_{\mathrm{kl}}(P_i, P_j) \leq \beta, \text{ for all } i, j.$$

Let U be a uniformly distributed random variable over $[r]$, and let X denote the random variable with $X|_{U=i} \sim P_i$. Finally, let $Z = \arg\min_{i \in [r]} d(\theta_i, \hat{\theta}(X))$, with ties broken arbitrarily. Note that $U \to X \to Z$ form a Markov chain.

(a) Show that the worst case estimation risk can be lower bounded by the probability of error in a hypothesis test:

$$\max_{i \in [r]} \mathbb{E}_i[d(\theta_i, \hat{\theta}(X))] \geq \frac{\alpha}{2} \mathbb{P}\left(Z \neq U\right). \tag{2}$$

(b) Obtain the following bound on the mutual information between U and X:

$$I(X; U) = I(U; X) \leq \frac{1}{r^2} \sum_{i=1}^{r} \sum_{j=1}^{r} D_{\mathrm{kl}}(P_i \parallel P_j).$$

Hint: recall that relative entropy is a convex functional.

(c) Use the previous result, along with the usual Fano's inequality, to show

$$\mathbb{P}\left(Z \neq U\right) \log(r-1) \geq \log r - \frac{1}{r^2} \sum_{i,j} D_{\mathrm{kl}}(P_i \parallel P_j) - 1. \tag{3}$$

(d) Combine (2) and (3) to get

$$\max_{i \in [r]} \mathbb{E}_i[d(\theta_i, \hat{\theta}(X))] \geq \frac{\alpha}{2} \left(1 - \frac{\beta + 1}{\log r}\right).$$

Thus the minimax estimation error (LHS above) is large, if there exist many distributions (i.e., large r) whose parameters are well-separated in Θ (i.e., large α), but they are statistically almost indistinguishable (i.e., small β).

(3+3+3+1 points)