

EE378C: Homework #1

Due on Wednesday, April 21, 2021

Please hand in your homework via Gradescope before 11:59 PM. Typing your solution using \LaTeX is highly recommended.

1. Consider the following variant of the hat guessing game we covered in class. There are n players sitting together, each of whom wears a red hat or a blue hat independently with probability $1/2$. Each player could see the color of all others' hats, but not his/her own. The players cannot communicate with each other, except for a previously agreed strategy. Here is the rule: everyone must guess the color of his/her own hat, and the players win if and only if *every* player makes a correct guess. So the difference here is that the players no longer have the option to pass.
 - (a) Prove that the winning probability of the players is at most $1/2$.
 - (b) Propose a strategy to achieve the $1/2$ winning probability.
2. This problem investigates some properties of several widely-used f -divergences, many of which will be helpful in our later lectures.

- (a) Let (P_n) and (Q_n) be two sequences of probability distributions defined on (\mathcal{X}_n) . Show that as $n \rightarrow \infty$, $\|P_n^{\otimes n} - Q_n^{\otimes n}\|_{\text{TV}} \rightarrow 0$ if and only if $n \cdot H^2(P_n, Q_n) \rightarrow 0$, and $\|P_n^{\otimes n} - Q_n^{\otimes n}\|_{\text{TV}} \rightarrow 1$ if and only if $n \cdot H^2(P_n, Q_n) \rightarrow \infty$. Here $P^{\otimes n}$ denotes the n -fold product distribution of P .

This problem implies that the Hellinger distance is a good proxy of whether the distributions would be asymptotically indistinguishable or singular.

- (b) For two conditional distributions $P_{Y|X}$, $Q_{Y|X}$ and a marginal distribution P_X , we define the *conditional KL divergence* $D_{\text{KL}}(P_{Y|X} \| Q_{Y|X} | P_X)$ as

$$D_{\text{KL}}(P_{Y|X} \| Q_{Y|X} | P_X) = \sum_{x \in \mathcal{X}} P_X(x) \cdot D_{\text{KL}}(P_{Y|X=x} \| Q_{Y|X=x}).$$

Prove the following chain rule: for two joint distributions $P_{XY} = P_X \times P_{Y|X}$ and $Q_{XY} = Q_X \times Q_{Y|X}$, it holds that

$$D_{\text{KL}}(P_{XY} \| Q_{XY}) = D_{\text{KL}}(P_X \| Q_X) + D_{\text{KL}}(P_{Y|X} \| Q_{Y|X} | P_X).$$

This is a more general chain rule for KL divergence covered in the lecture.

- (c) For a pair of discrete random variables (X, Y) with the joint distribution P_{XY} , the *mutual information* $I(X; Y)$ is defined as

$$I(X; Y) = D_{\text{KL}}(P_{XY} \| P_X \times P_Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{XY}(x, y) \log \frac{P_{XY}(x, y)}{P_X(x)P_Y(y)},$$

where P_X, P_Y denote the marginal distributions. Show that

$$I(X; Y) = \min_{Q_Y} D_{\text{KL}}(P_{Y|X} \| Q_Y | P_X),$$

where the minimum is over all possible probability distributions on \mathcal{Y} .

This is known as the variational representation of the mutual information, and sometimes called the “golden formula”. It is very useful in Fano’s method we will cover later.

- (d) Let P and Q be two probability distributions on the space \mathcal{X} , and $P = \int P_\theta \pi(d\theta)$ be a mixture distribution. In other words, $P(x) = \mathbb{E}_{\theta \sim \pi}[P_\theta(x)]$ for all $x \in \mathcal{X}$, and π is a probability distribution of θ . Show that

$$\chi^2(P, Q) = \mathbb{E}_{\theta, \theta' \sim \pi} \left[\sum_{x \in \mathcal{X}} \frac{P_\theta(x) P_{\theta'}(x)}{Q(x)} \right] - 1,$$

where θ' is an independent copy of θ .

This problem shows that the χ^2 -divergence enjoys a nice behavior under mixtures of the first argument. This is the key observation of the Ingster-Suslina method.

3. Consider the following randomized response technique that is used in sampling surveys when a sensitive question is asked that a person perhaps does not want to answer truthfully (e.g. using drugs). The technique is as follows: the respondent flips an unbiased coin first. If the coin turns up “heads”, he answers the question truthfully. Otherwise, he answers whether or not his social security number is even. The result of the coin toss is not revealed to the interviewer.

Let $X = 1$ if the interviewed person has the sensitive property, and $X = 0$ otherwise. Similarly, let Z be the indicator function on whether the coin turns up “heads”, and S be the indicator function on whether the social security number is even. It is natural to assume that X, S, Z are independent, and $S, Z \sim \text{Bernoulli}(1/2)$. Hence, the interviewer’s observation is $Y = ZX + (1 - Z)S$.

Now suppose that $X \sim \text{Bernoulli}(p)$ with the parameter set $p \in [0, 1]$. Let \mathcal{M} be the statistical model with the direct observation X , and \mathcal{N} be the associated statistical model with the indirect observation Y . Compute Le Cam’s distance $\Delta(\mathcal{M}, \mathcal{N})$.

4. As in Lecture 4, let $\mathcal{M}_{n,k}$ be the multinomial model with n samples and support size k , and $\mathcal{P}_{n,k}$ be the Poissonized model. In the lecture we have shown that $\Delta(\mathcal{M}_{n,k}, \mathcal{P}_{n,k}) = O(k^{1/2}/n^{1/4})$. In fact, a better characterization is possible:

$$\Delta(\mathcal{M}_{n,k}, \mathcal{P}_{n,k}) = \Theta \left(\min \left\{ 1, \sqrt{k/n} \right\} \right).$$

This problem is devoted to proving the improved upper bound.

- (a) Let P_0, Q_0 be two distributions on \mathcal{X} . For positive integers n and m , consider the following two distributions on \mathcal{X}^{n+m} : Q is the product distribution $Q_0^{\otimes(n+m)}$, and

$$P(x_1, \dots, x_{n+m}) = \mathbb{E}_S \left[\prod_{i \in S} P_0(x_i) \cdot \prod_{j \in [n+m] \setminus S} Q_0(x_j) \right],$$

where $S \subseteq [n+m]$ is a uniformly distributed random subset with size m . In other words, P is the distribution of the sequence (x_1, \dots, x_{n+m}) when we generate n iid samples $y_1, \dots, y_n \sim Q_0$ and m iid samples $z_1, \dots, z_m \sim P_0$, and then pool them together with a random shuffling.

Use the result of Problem 2(d), show that

$$\chi^2(P, Q) = \mathbb{E}_{S, S'} \left[\left(1 + \chi^2(P_0, Q_0)\right)^{|S \cap S'|} - 1 \right],$$

where S' is an independent copy of S , and $|A|$ denotes the cardinality of set A .

- (b) The random variable $|S \cap S'|$ follows a hypergeometric distribution, which is a bit hard to analyze. Luckily, there is a result saying that $|S \cap S'|$ is dominated by a Binomial random variable $X \sim \mathbf{B}(m, m/(n+m))$ in terms of the convex ordering, i.e. for every convex function $f(\cdot)$ defined on $\{0, 1, \dots, m\}$, it holds that

$$\mathbb{E}[f(|S \cap S'|)] \leq \mathbb{E}[f(X)].$$

Using this result and Part (a), show that

$$\chi^2(P, Q) \leq \left(1 + \frac{m}{n+m} \chi^2(P_0, Q_0)\right)^m - 1.$$

- (c) Now choose Q_0 to be the discrete distribution (p_1, \dots, p_k) , and P_0 to be a *random* distribution $(\hat{p}_1, \dots, \hat{p}_k)$, which is the empirical distribution of n *external* samples drawn from Q_0 . Based on Part (b), argue that

$$\mathbb{E}[\|P - Q\|_{\text{TV}}] \leq \frac{m\sqrt{k}}{n},$$

where the expectation is taken with respect to the randomness in P_0 .

- (d) Now we are ready to present the new randomization scheme from multinomial to Poissonized model. Suppose that X_1, \dots, X_n are n iid samples drawn from the multinomial model $\mathcal{M}_{n,k}$. We draw an independent $N \sim \text{Poi}(n)$:

- if $N \leq n$, we simply output (X_1, \dots, X_N) ;
- if $N > n$, we compute the empirical distribution P_0 from the first $n/2$ samples $(X_1, \dots, X_{n/2})$, generate $m \triangleq N - n$ fake samples $Y_1, \dots, Y_m \sim P_0$, and shuffle the pool $(X_{n/2+1}, \dots, X_n, Y_1, \dots, Y_m)$ uniformly at random to arrive at $(Z_1, \dots, Z_{n/2+m})$. Finally, we output $(X_1, \dots, X_{n/2}, Z_1, \dots, Z_{n/2+m})$.

Show that this scheme achieves a TV distance to the Poissonized model at most $O(\sqrt{k/n})$.

Note: the randomization from Poissonized model to the multinomial model could be done symmetrically.

5. This problem is devoted to a partial converse of Problem 4. Specifically, we will show that when $k = Cn$ with a large constant $C > 0$, the distance $\Delta(\mathcal{M}_{n,k}, \mathcal{P}_{n,k})$ does not converge to zero as $n \rightarrow \infty$. Throughout this problem, fix $m = n + \sqrt{n}$ and assume m is an integer; also assume that $k \geq 2n \geq \sqrt{k}$.

- (a) Suppose X_1, \dots, X_n are n iid observations from the uniform distribution on $[k]$. Define U_n to be the number of different symbols in X_1, \dots, X_n (for example, when $n = 5$ and the observations are $(1, 2, 1, 4, 1)$, then $U_5 = 3$). Similarly, define U_m to be the number of different symbols in m iid observations. Find $\mathbb{E}[U_n], \mathbb{E}[U_m]$, and argue that $\mathbb{E}[U_m - U_n] \geq \sqrt{n}/4$.
- (b) Using Azuma's inequality, argue that there is an absolute constant $c_1 < \infty$ independent of (n, k) such that

$$\mathbb{P}\left(U_n - \mathbb{E}[U_n] \geq c_1 \cdot \frac{n}{\sqrt{k}}\right) \leq 0.01, \quad \mathbb{P}\left(U_m - \mathbb{E}[U_m] \leq -c_1 \cdot \frac{n}{\sqrt{k}}\right) \leq 0.01.$$

You could use the following version of Azuma's inequality. Let $Z_1, \dots, Z_n \in [0, 1]$ be n possibly dependent random variables, where for each $i \in [n]$, it always holds that $\text{Var}(Z_i \mid Z_1, \dots, Z_{i-1}) \leq \sigma_i^2$ for all possible (Z_1, \dots, Z_{i-1}) . Then for $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^n (Z_i - \mathbb{E}[Z_i])\right| \geq \varepsilon\right) \leq 2 \exp\left(-\frac{\varepsilon^2}{2(\sum_{i=1}^n \sigma_i^2 + \varepsilon/3)}\right).$$

Now we consider a prior distribution on the discrete distribution p , as well as a proper loss function, to evaluate and compare the Bayes risks in both models.

Suppose that there are $100k$ symbols in total, and we choose p to be the uniform distribution on a uniformly random subset of $\mathcal{X} = [100k]$ with size k . Upon observing the samples X_1, \dots, X_n in the multinomial model (or X_1, \dots, X_N with $N \sim \text{Poi}(n)$ in the Poissonized model), the action space \mathcal{A} is chosen to be the set \mathcal{X}^m of all length- m sequences. The loss function $L(p, x^m)$ with the true distribution p and the action $x^m \in \mathcal{X}^m$ is defined as

$$L(p, x^m) = \mathbb{1}(\text{some symbol in } x^m \text{ does not belong to the support of } p, \\ \text{or there are at most } \mathbb{E}[U_m] - c_1 n / \sqrt{k} \text{ different symbols in } x^m),$$

with constant c_1 defined in Part (b). Let $R^*(\mathcal{M}_{n,k})$ (resp. $R^*(\mathcal{P}_{n,k})$) be the Bayes risk under the above prior and loss in the multinomial (resp. Poissonized) model.

- (c) Show that under the Poissonized model, we have

$$R^*(\mathcal{P}_{n,k}) \leq \mathbb{P}(\text{Poi}(n) < m) + 0.01.$$

- (d) Show that if $k = Cn$ with a large enough C , then $R^*(\mathcal{M}_{n,k}) \geq 0.97$.

Hint: what is the posterior distribution of p given the samples X_1, \dots, X_n ? Try to show that a big loss is incurred either if x^m includes any symbol not seen in X_1, \dots, X_n , or if x^m only contains symbols in X_1, \dots, X_n .

- (e) Based on (c) and (d), conclude that when $k = Cn$, Le Cam's distance $\Delta(\mathcal{M}_{n,k}, \mathcal{P}_{n,k})$ will be bounded away from zero.