EE378C: Homework #1 Solutions

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.

   Solution:

   (a) Since the hat color of player 1 is independent of those of other players, the probability that player 1 guesses his/her own hat color correctly is at most $1/2$. This probability is no smaller than the probability that everyone makes a correct guess.

   (b) One possible strategy is as follows: every player guesses his/her own hat color in the way that the number of red hats among all players is an even number. Clearly, the players win iff the total number of red hats is indeed an even number, which has probability

$$\mathbb{P}\left(\mathsf{B}(n,1/2) \text{ is even}\right) = \frac{1}{2^n}\sum_{k=0}^{\lfloor n/2\rfloor}\binom{n}{2k} = \frac{1}{2^{n+1}}\left[\sum_{k=0}^{n}\binom{n}{k} + \sum_{k=0}^{n}\binom{n}{k}(-1)^k\right] = \frac{1}{2}.$$

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 $n \to \infty$, $\|P_n^{\otimes n} - Q_n^{\otimes n}\|_{\mathrm{TV}} \to 0$ if and only if $n \cdot H^2(P_n, Q_n) \to 0$, and $\|P_n^{\otimes n} - Q_n^{\otimes n}\|_{\mathrm{TV}} \to 1$ if and only if $n \cdot H^2(P_n, Q_n) \to \infty$. Here $P^{\otimes n}$ denotes the n-fold product distribution of P.

   This problem means 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* $D_{\mathrm{KL}}(P_{Y|X}\|Q_{Y|X} \mid P_X)$ as

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

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

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

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

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

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

where P_X, P_Y denote the marginal distributions. Show that

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

where the minimum is over all possible probability distributions on $\mathcal{Y}$.
This is called the variational representation of the mutual information, and sometimes called the "golden formula". This is very useful in the Fano's method we'll 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.

Solution:

(a) Using the inequality between TV and Hellinger in the lecture note

$$\frac{1}{2} H^2(P, Q) \le \|P - Q\|_{\mathrm{TV}} \le H(P, Q) \sqrt{1 - \frac{H^2(P, Q)}{4}},$$

we see that $\|P - Q\|_{\mathrm{TV}} \to 0 \Leftrightarrow H^2(P, Q) \to 0$, and $\|P - Q\|_{\mathrm{TV}} \to 1 \Leftrightarrow H^2(P, Q) \to 2$. As a result,

$$\begin{aligned}
\|P_n^{\otimes n} - Q_n^{\otimes n}\|_{\mathrm{TV}} \to 0 &\iff H^2(P_n^{\otimes n}, Q_n^{\otimes n}) \to 0 \\
&\iff 2 \left[1 - \left(1 - \frac{H^2(P_n, Q_n)}{2} \right)^n \right] \to 0 \\
&\iff \left(1 - \frac{H^2(P_n, Q_n)}{2} \right)^n \to 1 \\
&\iff n \cdot H^2(P_n, Q_n) \to 0.
\end{aligned}$$

The other claim could be established analogously.

(b) Just expand by definition:

$$\begin{aligned}
D_{\mathrm{KL}}(P_{XY}\|Q_{XY}) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(x,y)\log\frac{P_{XY}(x,y)}{Q_{XY}(x,y)} \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(x,y)\left(\log\frac{P_X(x)}{Q_X(x)} + \log\frac{P_{Y|X=x}(y)}{Q_{Y|X=x}(y)}\right) \\
&= \sum_{x\in\mathcal{X}} P_X(x)\log\frac{P_X(x)}{Q_X(x)} + \sum_{x\in\mathcal{X}} P_X(x)\sum_{y\in\mathcal{Y}} P_{Y|X=x}(y)\log\frac{P_{Y|X=x}(y)}{Q_{Y|X=x}(y)} \\
&= D_{\mathrm{KL}}(P_X\|Q_X) + D_{\mathrm{KL}}(P_{Y|X}\|Q_{Y|X} \mid P_X).
\end{aligned}$$

(c) By definition of the mutual information, we have

$$\begin{aligned}
I(X;Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(x,y)\log\frac{P_{Y|X=x}(y)}{P_Y(y)} \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(x,y)\log\frac{P_{Y|X=x}(y)}{Q_Y(y)} + \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(x,y)\log\frac{Q_Y(y)}{P_Y(y)} \\
&= \sum_{x\in\mathcal{X}} P_X(x)\sum_{y\in\mathcal{Y}} P_{Y|X=x}(y)\log\frac{P_{Y|X=x}(y)}{Q_Y(y)} + \sum_{y\in\mathcal{Y}} P_Y(y)\log\frac{Q_Y(y)}{P_Y(y)} \\
&= D_{\mathrm{KL}}(P_{Y|X}\|Q_Y \mid P_X) - D_{\mathrm{KL}}(P_Y\|Q_Y).
\end{aligned}$$

As $D_{\mathrm{KL}}(P_Y\|Q_Y) \geq 0$, and it is zero when $Q_Y = P_Y$, the variational representation holds.

(d) Since $P(x) = \mathbb{E}_{\theta\sim\pi}[P_\theta(x)]$, we have $P(x)^2 = \mathbb{E}_{\theta,\theta'\sim\pi}[P_\theta(x)P_{\theta'}(x)]$. Consequently,

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

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 $S, Z \sim \mathsf{Bernoulli}(1/2)$. Hence, the interviewer's observation is $Y = ZX + (1-Z)S$.

Now suppose that $X \sim \mathsf{Bernoulli}(p)$ 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 $\Delta(\mathcal{M}, \mathcal{N})$.

Solution: It is easy to show that $X \sim P_X = \mathsf{Bernoulli}(p)$ and $Y \sim P_Y = \mathsf{Bernoulli}(\frac{2p+1}{4})$. Since $\mathcal{N}$ is a randomization of $\mathcal{M}$, we only need to show how well randomizations of $\mathcal{N}$ could approximate $\mathcal{M}$. The stochastic kernel K from Y to X could be represented by two free variables:

$$a \triangleq \mathsf{K}(X = 1 \mid Y = 0), \quad b \triangleq \mathsf{K}(X = 1 \mid Y = 1).$$

Then the distribution $\mathsf{K}P_Y$ is $\mathsf{Bernoulli}((3-2p)a/4 + (2p+1)b/4)$. Therefore,

$$\begin{aligned}\max_{p\in[0,1]} \|P_X - \mathsf{K}P_Y\|_{\mathrm{TV}} &= \max_{p\in[0,1]} \left| \frac{(3-2p)a + (1+2p)b}{4} - p \right| \\ &= \max\left\{ \frac{3a+b}{4}, 1 - \frac{a+3b}{4} \right\} \quad \text{(max. attained at boundary)}.\end{aligned}$$

By the equivalent representation of Le Cam's distance using randomization, $\Delta(\mathcal{M}, \mathcal{N})$ is the objective value to the following linear program:

$$\begin{aligned} \text{minimize} \quad & \xi \\ \text{subject to} \quad & \frac{3a+b}{4} \le \xi, \quad 1 - \frac{a+3b}{4} \le \xi, \quad a, b \in [0,1]. \end{aligned}$$

It is not hard to find that the optimal solution is $(a^\star, b^\star, \xi^\star) = (0, 1, 1/4)$, so $\Delta(\mathcal{M}, \mathcal{N}) = 1/4$.

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

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

where $S \subseteq [n+m]$ is a uniformly distributed random subset with size m. In other words, P is the distribution of the sequence $(x_1, \cdots, x_{n+m})$ when we generate n

iid samples $y_1, \cdots, y_n \sim Q_0$ and m iid samples $z_1, \cdots, z_m \sim P_0$, and then pool them together with a random shuffling.

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

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

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 $X \sim \mathsf{B}(m, m/(n+m))$ in terms of the convex ordering, i.e. for every convex function $f(\cdot)$ defined on $\{0, 1, \cdots, m\}$, it holds that

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

Using this result and Part (a), show that

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

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

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

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 $X_1, \cdots, X_n$ are n iid samples drawn from the multinomial model $\mathcal{M}_{n,k}$. We draw an independent $N \sim \mathsf{Poi}(n)$:

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

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.

Solution:

(a) For $S \subseteq [n+m]$ with size m, define

$$P_S(x_1, \cdots, x_{n+m}) = \prod_{i \in S} P_0(x_i) \cdot \prod_{j \in [n+m] \setminus S} Q_0(x_j).$$

Then $P = \mathbb{E}_S[P_S]$. Moreover, for $S, S' \subseteq [n+m]$, we have

$$\sum_{x \in \mathcal{X}^{n+m}} \frac{P_S(x) P_{S'}(x)}{Q(x)} = \prod_{i=1}^{n+m} \left(\sum_{x_i \in \mathcal{X}} \frac{P_{S,i}(x_i) P_{S',i}(x_i)}{Q_0(x_i)} \right),$$

where $P_{S,i} = P_0$ if $i \in S$ and Q_0 if $i \notin S$ (similarly for $P_{S',i}$). It is easy to verify that

$$\sum_{x_i \in \mathcal{X}} \frac{P_{S,i}(x_i) P_{S',i}(x_i)}{Q_0(x_i)} = \begin{cases} 1 + \chi^2(P_0, Q_0) & \text{if } i \in S \cap S' \\ 1 & \text{o.w.} \end{cases}.$$

Consequently, by Problem 2(d), it holds that

$$\chi^2(P, Q) = \mathbb{E}_{S,S'} \left[\sum_{x \in \mathcal{X}^{n+m}} \frac{P_S(x) P_{S'}(x)}{Q(x)} \right] - 1 = \mathbb{E}_{S,S'} \left[\left(1 + \chi^2(P_0, Q_0)\right)^{|S \cap S'|} \right] - 1.$$

(b) Note that $t \mapsto (1 + \chi^2(P_0, Q_0))^t$ is convex, so

$$\begin{aligned} \mathbb{E}_{S,S'} \left[\left(1 + \chi^2(P_0, Q_0)\right)^{|S \cap S'|} \right] &\le \mathbb{E} \left[\left(1 + \chi^2(P_0, Q_0)\right)^{X} \right] \\ &= \prod_{i=1}^{m} \mathbb{E} \left[\left(1 + \chi^2(P_0, Q_0)\right)^{X_i} \right] \\ &= \left(1 + \frac{m}{n+m} \chi^2(P_0, Q_0)\right)^m, \end{aligned}$$

where we have written $X = X_1 + \cdots + X_m$ with iid $X_i \sim \mathsf{Bernoulli}(m/(n+m))$.

(c) Fixing the realization of P_0, by Part (b) we have

$$\begin{aligned} \|P - Q\|_{\mathrm{TV}} &\le \sqrt{\frac{1}{2} \log(1 + \chi^2(P, Q))} \\ &\le \sqrt{\frac{m}{2} \log \left(1 + \frac{m}{n+m} \chi^2(P_0, Q_0)\right)} \le \sqrt{\frac{m^2}{2(n+m)} \chi^2(P_0, Q_0)}. \end{aligned}$$

As shown in Lecture 4, we have $\mathbb{E}[\chi^2(P_0, Q_0)] = (k-1)/n$. Consequently, by the concavity of $x \mapsto \sqrt{x}$, we have

$$\mathbb{E}[\|P - Q\|_{\mathrm{TV}}] \le \sqrt{\frac{m^2}{2(n+m)} \cdot \mathbb{E}[\chi^2(P_0, Q_0)]} < \frac{m\sqrt{k}}{n}.$$

(d) By Part (c), the final TV distance to the Poissonized model is at most $\mathbb{E}[m\sqrt{k}/(n/2)]$, where $m \triangleq (N-n)_+$ is a random variable. The proof is completed by noting that $\mathbb{E}[m] \le \mathbb{E}[(N-n)^2]^{1/2} = \sqrt{n}$.

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 $n \to \infty$. 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 $X_1, \cdots, X_n$ are n iid observations from the uniform distribution on $[k]$. Define U_n to be the number of different symbols in $X_1, \cdots, X_n$ (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 $Z_1, \cdots, Z_n \in [0, 1]$ be n possibly dependent random variables, where for each $i \in [n]$, it always holds that $\mathsf{Var}(Z_i \mid Z_1, \cdots, Z_{i-1}) \leq \sigma_i^2$ for all possible $(Z_1, \cdots, Z_{i-1})$. 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 $X_1, \cdots, X_n$ in the multinomial model (or $X_1, \cdots, X_N$ with $N \sim \mathsf{Poi}(n)$ 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 $x^m \in \mathcal{X}^m$ 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 $R^\star(\mathcal{M}_{n,k})$ (resp. $R^\star(\mathcal{P}_{n,k})$) 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

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

   (d) Show that if $k = Cn$ with a large enough C, then $R^\star(\mathcal{M}_{n,k}) \geq 0.97$.
   Hint: what is the posterior distribution of p given the samples $X_1, \cdots, X_n$? Try to show that a big loss is incurred either if x^m includes any symbol not seen in $X_1, \cdots, X_n$, or if x^m only contains symbols in $X_1, \cdots, X_n$.

(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.

Solution:

(a) Note that $U_n = \sum_{j=1}^{k} \mathbb{1}(\text{symbol } j \text{ appears in } n \text{ samples})$, so by linearity of expectation, we have

$$\mathbb{E}[U_n] = k\left[1 - \left(1 - \frac{1}{k}\right)^n\right].$$

Consequently,

$$\mathbb{E}[U_m - U_n] = k\left(1 - \frac{1}{k}\right)^n \left[1 - \left(1 - \frac{1}{k}\right)^{\sqrt{n}}\right] \geq \frac{k}{e} \cdot \frac{3\sqrt{n}}{4k} > \frac{\sqrt{n}}{4},$$

where we have used that $(1-1/k)^n \geq (1-1/k)^k \geq 1/e$ and $(1-x)^n \leq 1 - 3nx/4$ if $0 \leq nx \leq 1/2$.

(b) Define $Z_i = \mathbb{1}(\text{the } i\text{-th observation does not appear in the past observations})$, then $U_n = \sum_{i=1}^{n} Z_i$. Clearly $0 \leq Z_i \leq 1$, and

$$\mathsf{Var}(Z_i \mid Z_1, \cdots, Z_{i-1}) \leq \mathbb{E}[1 - Z_i \mid Z_1, \cdots, Z_{i-1}] \leq \frac{i-1}{k} \leq \frac{n}{k}.$$

Now the claimed result follows from Azuma's inequality and the assumption $n = \Omega(\sqrt{k})$.

(c) An estimator in the Poissonized model could be constructed as follows: if $N < m$, output any sequence in $\mathcal{X}^m$; if $N \geq m$, output the first m samples $X_1, \cdots, X_m$. In the latter scenario, Part (b) implies that with probability at most 0.01, there would be at most $\mathbb{E}[U_m] - c_1 n/\sqrt{k}$ different symbols in $X_1, \cdots, X_m$; also, the sequence always belongs to the support of p. In other words, if $N \geq m$, we have $\mathbb{E}[L(p, X^m)] \leq 0.01$; if $N < m$, we could use the trivial upper bound 1. Combining these two scenarios completes the proof.

(d) Choose $C > 0$ large enough so that $2c_1 n/\sqrt{k} < \sqrt{n}/4$. In this case, Parts (a) and (b) imply that $U_n < \mathbb{E}[U_m] - c_1 n/\sqrt{k}$ holds with probability at least 0.99.
It is easy to see that the posterior distribution of p given $X_1, \cdots, X_n$ is to pick the remaining support locations uniformly at random from the symbols not seen in $X_1, \cdots, X_n$. Note that there are at least $99k$ symbols not seen in $X_1, \cdots, X_n$, and at most k of them lie in the true support of p. Hence, if the output sequence x^m includes any symbol not seen in $X_1, \cdots, X_n$, then the first event in the loss $L(p, x^m)$ will be triggered with probability at least $98/99$. On the other hand, if x^m only contains seen symbols in $X_1, \cdots, X_n$, then with probability at least 0.99, the number of different symbols is too small and triggers the second event in the loss. Therefore, using the union bound, the Bayes risk is at least

$$1 - 0.01 - \frac{1}{99} > 0.97.$$

(e) It is known in the lecture that Le Cam's distance is always lower bounded by the difference in Bayes risks for the models. By Part (c) and Poisson CLT, we have $R^\star(\mathcal{P}_{n,k}) \le \Phi(1) + o_n(1) + 0.01 < 0.86 + o_n(1)$; in contrast, Part (d) shows that $R^\star(\mathcal{M}_{n,k}) \ge 0.97$. Therefore,

$$\Delta(\mathcal{M}_{n,k}, \mathcal{P}_{n,k}) \ge R^\star(\mathcal{M}_{n,k}) - R^\star(\mathcal{P}_{n,k}) \ge 0.11 - o_n(1),$$

which is asymptotically bounded away from zero.