EE378C: Homework #2 Solutions

Due on Wednesday, May 5, 2021

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

1. This problem concerns the local asymptotic minimax theorem applied to the entropy estimation example covered in Lecture 7. Recall the following setup: the learner draws n iid samples $X_1, \cdots, X_n$ from a discrete distribution $P = (p_1, \cdots, p_k)$, and aims to estimate the entropy $H(P) = \sum_{i=1}^k -p_i \log p_i$. With a slight abuse of notation, we also use P to denote the free parameter $(p_1, \cdots, p_{k-1})$, which belongs to the parameter set $\mathcal{P}_k = \{(p_1, \cdots, p_{k-1}) \in \mathbb{R}_+^{k-1} : \sum_{i=1}^{k-1} p_i \le 1\}$ with a non-empty interior in $\mathbb{R}^{k-1}$.

   (a) For a fixed P in the interior of $\mathcal{P}_k$, find the expression of the Fisher information $I(P)$ and the inverse Fisher information $I(P)^{-1}$ in the above model with $n = 1$. *Hint: the following Woodbury matrix identity might be useful: for invertible A, C,*

   $$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$

   (b) Use the local asymptotic minimax theorem to show that for any P_0 in the interior of $\mathcal{P}_k$ and any sequence of estimators $\widehat{H}_n$ based on n samples, it holds that

   $$\lim_{C\to\infty} \liminf_{n\to\infty} n \cdot \sup_{P\in\mathcal{P}_k : \|P-P_0\|_2 \le C/\sqrt{n}} \mathbb{E}_P[(\widehat{H}_n - H(P))^2] \ge \mathsf{Var}_{X\sim P_0}(\log P_0(X)),$$

   where for $P = (p_1, \cdots, p_k)$, the variance is defined as

   $$\mathsf{Var}_{X\sim P}(\log P(X)) \triangleq \sum_{i=1}^k p_i \log^2 p_i - \left(\sum_{i=1}^k p_i \log p_i\right)^2.$$

   (c) Find a suitable P_0 in (b) to conclude that

   $$\liminf_{n\to\infty} n \cdot \inf_{\widehat{H}_n} \sup_{P\in\mathcal{P}_k} \mathbb{E}_P[(\widehat{H}_n - H(P))^2] \ge c \cdot \log^2 k,$$

   where $c > 0$ is an absolute constant independent of (n, k).

Solution:

(a) The log-likelihood function is

$$\ell_P(x) = \sum_{i=1}^{k-1} \mathbb{1}(x=i)\log p_i + \mathbb{1}(x=k)\log\left(1 - \sum_{i=1}^{k-1} p_i\right).$$

Consequently, the score function is

$$[\dot{\ell}_P(x)]_i = \frac{\partial}{\partial p_i}\ell_P(x) = \frac{\mathbb{1}(x=i)}{p_i} - \frac{\mathbb{1}(x=k)}{1-\sum_{j=1}^{k-1} p_j}, \quad i \in [k-1],$$

and the Fisher information matrix is

$$[I(P)]_{i,j} = \mathbb{E}_{X\sim P}[[\dot{\ell}_P(x)]_i[\dot{\ell}_P(x)]_j] = \frac{\mathbb{1}(i=j)}{p_i} + \frac{1}{p_k}, \quad i,j \in [k-1].$$

In other words, $I(P) = \mathrm{diag}(p_1^{-1}, p_2^{-1}, \cdots, p_{k-1}^{-1}) + p_k^{-1}\mathbf{1}\mathbf{1}^\top$. Applying the Woodbury identity to $A = \mathrm{diag}(p_1^{-1}, p_2^{-1}, \cdots, p_{k-1}^{-1}), U = \mathbf{1}, C = p_k^{-1}, V = \mathbf{1}^\top$ gives

$$I(P)^{-1} = \mathrm{diag}(P) - PP^\top.$$

(b) In the application of the local asymptotic minimax theorem, we have $\ell(t) = t^2$, and $\psi(P) = \sum_{i=1}^{k-1} -p_i \log p_i - (1 - \sum_{i=1}^{k-1} p_i)\log(1 - \sum_{i=1}^{k-1} p_i)$. Note that

$$\nabla\psi(P) = (\log(p_k/p_1), \cdots, \log(p_k/p_{k-1}))^\top,$$

we have

$$\begin{aligned}
\nabla\psi(P)^\top I(P)^{-1}\nabla\psi(P) &= \sum_{i=1}^{k-1} p_i(\log p_k - \log p_i)^2 - \left(\sum_{i=1}^{k-1} p_i(\log p_k - \log p_i)\right)^2 \\
&= \sum_{i=1}^{k-1} p_i \log^2 p_i + (1-p_k)\log^2 p_k + 2(H(P) + p_k\log p_k)\log p_k \\
&\qquad - (H(P) + \log p_k)^2 \\
&= \sum_{i=1}^{k} p_i \log^2 p_i - H(P)^2,
\end{aligned}$$

as desired.

(c) Choose $P_0 = (1/[3(k-1)], \cdots, 1/[3(k-1)], 2/3)$, then

$$\begin{aligned}
\mathsf{Var}_{X\sim P_0}(\log P_0(X)) &= \frac{1}{3}\log^2(3(k-1)) + \frac{2}{3}\log^2(3/2) - \left(\frac{\log[3(k-1)] + 2\log(3/2)}{3}\right)^2 \\
&= \frac{2\log^2[2(k-1)]}{9} \geq \frac{2}{9}\log^2 k.
\end{aligned}$$

Therefore, we can choose $c = 2/9$.

2. In this problem we analyze the reduction scheme from Lecture 6 that lets us approximately map

- $\mathrm{Bern}(1/2) \to \mathcal{N}(0,1)$
- $\mathrm{Bern}(1) \to \mathcal{N}(\mu,1)$

Formally, we observe a bit $B \in \{0,1\}$ and use it to create a distribution that approximates a Gaussian random variable. Consider the following algorithm $RK(B)$:

- If $B = 1$, sample $Z \sim \mathcal{N}(\mu, 1)$. Output Z with probability 1.

- If $B = 0$, set $T = 0$ and $Z = 0$.
  While $T \leq N$, sample $Y_T \sim \mathcal{N}(0,1)$.
  With probability $\max\left\{0, 1 - \frac{\mathcal{N}(\mu,1)(Y_T)}{2\mathcal{N}(0,1)(Y_T)}\right\}$, set $Z = Y_T$ and break.
  Otherwise, increment T by 1.
  Output Z.

Here $\mathcal{N}(a,1)(x)$ represents the pdf of the distribution $\mathcal{N}(a,1)$ at x. Denote the distribution of the output of this algorithm when $B \sim \text{Bern}(x)$ as $RK(\text{Bern}(x))$.

(a) Compute $\|RK(1) - \mathcal{N}(\mu,1)\|_{\text{TV}}$.

(b) Define $S = \{x \in \mathbb{R} : 2\mathcal{N}(0,1)(x) \geq \mathcal{N}(\mu,1)(x)\}$ and a distribution φ supported on S specified by the pdf

$$\varphi(x) = \frac{\mathcal{N}(0,1)(x) - \frac{1}{2}\mathcal{N}(\mu,1)(x)}{p} \cdot \mathbb{1}(x \in S),$$

where $p := \mathbb{P}_{X\sim\mathcal{N}(0,1)}[X \in S] - \frac{1}{2}\mathbb{P}_{X\sim\mathcal{N}(\mu,1)}[X \in S]$ is the normalizing constant. Show that $\|RK(0) - \varphi\|_{\text{TV}} = (1-p)^N$.
Hint: Show that if $P_{X|A}$ denotes the conditional distribution of X given $X \in A$, then $\|P_X - P_{X|A}\|_{\text{TV}} = P_X(A^c)$.

(c) Show that

$$\|RK(\text{Bern}(1/2)) - \mathcal{N}(0,1)\|_{\text{TV}} \leq \frac{1}{2}(1-p)^N + p - \frac{1}{2}.$$

Note: as $\mu \to 0$ and $N \to \infty$, one can show that $p \to 1/2$, which means that we have achieved the desired approximate reduction.

Solution:

(a) Clearly $RK(1)$ is $\mathcal{N}(\mu,1)$, so the TV distance is 0.

(b) First we show the auxiliary result in the hint. Clearly, the symmetric difference between $\{x : P_X(x) \geq P_{X|A}(x)\}$ and $\{x : x \in A^c\}$ has P_X-probability zero, so

$$\|P_X - P_{X|A}\|_{\text{TV}} = P_X(A^c) - P_{X|A}(A^c) = P_X(A^c).$$

Next, observe that if Y_T is outputted at any iteration, the distribution of $Z = Y_T$ is precisely ϕ. Consequently, $\|RK(0) - \phi\|_{\text{TV}}$ is equal to the probability that the loop does not break for N rounds. The probability of breaking the loop at each iteration is

$$\begin{aligned}\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\max\left\{0, 1 - \frac{\mathcal{N}(\mu,1)(x)}{2\mathcal{N}(0,1)(x)}\right\}\right] &= \int_0^\infty \max\left\{0, \mathcal{N}(0,1)(x) - \frac{1}{2}\mathcal{N}(\mu,1)(x)\right\} dx \\ &= p.\end{aligned}$$

The result follows from the fact that the above events are mutually independent for each iteration.

(c) Using the triangle inequality for the TV distance:

$$\begin{aligned}\|RK(\text{Bern}(1/2)) - \mathcal{N}(0,1)\|_{\text{TV}} &= \left\|\frac{RK(0)+\mathcal{N}(\mu,1)}{2} - \mathcal{N}(0,1)\right\|_{\text{TV}} \\ &\le \left\|\frac{\varphi+\mathcal{N}(\mu,1)}{2} - \mathcal{N}(0,1)\right\|_{\text{TV}} + \frac{1}{2}\|RK(0)-\varphi\|_{\text{TV}}.\end{aligned}$$

By Part (b), the second term is upper bounded by $(1-p)^N/2$. As for the first term, note that

$$\begin{aligned} p &= \int_0^\infty \max\left\{0, \mathcal{N}(0,1)(x) - \frac{1}{2}\mathcal{N}(\mu,1)(x)\right\} dx \\ &> \int_0^\infty \Big(\mathcal{N}(0,1)(x) - \frac{1}{2}\mathcal{N}(\mu,1)(x)\Big) dx = \frac{1}{2},\end{aligned}$$

therefore it is straightforward to verify

$$\left\{x : \frac{\varphi+\mathcal{N}(\mu,1)}{2}(x) \ge \mathcal{N}(0,1)(x)\right\} = \{x : x \in S^c\}.$$

Consequently,

$$\begin{aligned}\left\|\frac{\varphi+\mathcal{N}(\mu,1)}{2} - \mathcal{N}(0,1)\right\|_{\text{TV}} &= \frac{1}{2}\mathbb{P}_{X\sim\mathcal{N}(\mu,1)}(X\in S^c) - \mathbb{P}_{X\sim\mathcal{N}(0,1)}(X\in S^c) \\ &= \mathbb{P}_{X\sim\mathcal{N}(0,1)}(X\in S) - \mathbb{P}_{X\sim\mathcal{N}(\mu,1)}(X\in S) - \frac{1}{2} \\ &= p - \frac{1}{2}.\end{aligned}$$

3. In class we see how the two-point method is used to establish the lower bound for the *expected* loss $\mathbb{E}_\theta[L(\theta,T)]$. In some scenarios we are also interested in the *high probability* upper bound of the following form: $L(\theta,T) < \varepsilon$ with probability at least $1-\delta$. In this problem we show how to adapt the two-point method to proving lower bounds for the high probability result, i.e. show that $L(\theta,T) \ge \varepsilon$ with probability at least δ under $X \sim P_\theta$ for some $\theta \in \Theta$. In particular, we are interested in the risk dependence on the error probability δ.

   (a) Find a loss function $L_0(\theta,a)$ (which may depend on L and ε), such that

   $$\sup_{\theta\in\Theta} \mathbb{E}_\theta[L_0(\theta,T(X))] \le \delta$$

   if and only if the estimator T satisfies $L(\theta,T) < \varepsilon$ with probability at least $1-\delta$ for every $\theta \in \Theta$.

   (b) Consider a Bernoulli model $X_1,\cdots,X_n \sim \text{Bern}(p)$ with unknown $p \in [0,1]$ and loss $L(p,a) = |p-a|$. By applying the two-point method to the loss function L_0, argue that there exists an estimator T with $L(\theta, T) < \varepsilon$ with probability at least $1-\delta$ for every $\theta \in \Theta$, where $\varepsilon, \delta \in (0, 1/4)$, *only if*

$$n \geq c \cdot \frac{\log(1/\delta)}{\varepsilon^2},$$

for some absolute constant $c > 0$ independent of (ε, δ). In other words, given n samples, any estimator suffers a loss at least $\Omega(\sqrt{\log(1/\delta)/n})$ with probability at least δ for the worst-case $p \in [0, 1]$.

Hint: recall the following relationship between TV *and* KL*:*

$$\|P - Q\|_{\mathrm{TV}} \leq 1 - \frac{1}{2}\exp(-D_{\mathrm{KL}}(P\|Q)).$$

(c) Now consider the uniformity testing problem covered in Lecture 8. Show that if the test error is required to be at most $\delta \in (0, 1/4)$ under both $H_0 : P = \mathsf{Unif}([k])$ and $H_1 : \|P - \mathsf{Unif}([k])\|_{\mathrm{TV}} \geq \varepsilon$, the number of samples required is at least

$$n = \Omega\left(\frac{1}{\varepsilon^2}\sqrt{k \log\left(\frac{1}{\delta}\right)}\right).$$

Note: the dependence on δ in (b) and (c), albeit different, is both tight.

Solution:

(a) $L_0(\theta, a) = \mathbb{1}(L(\theta, a) > \varepsilon)$.

(b) Choose $p_0 = 1/2 - \varepsilon$ and $p_1 = 1/2 + \varepsilon$. Then by definition of L_0 in Part (a), the separation condition holds with $\Delta = 1$. Consequently, two-point method gives

$$\begin{aligned}
\inf_T \sup_{p \in [0,1]} \mathbb{E}_\theta[L_0(p, T(X))] &\geq \frac{1}{2}\left(1 - \|\mathrm{Bern}(p_0)^{\otimes n} - \|\mathrm{Bern}(p_1)^{\otimes n}\|_{\mathrm{TV}}\right) \\
&\geq \frac{1}{4}\exp\left(-D_{\mathrm{KL}}(\mathrm{Bern}(p_0)^{\otimes n}\|\mathrm{Bern}(p_1)^{\otimes n})\right) \\
&= \frac{1}{4}\exp\left(-nD_{\mathrm{KL}}(\mathrm{Bern}(p_0)\|\mathrm{Bern}(p_1))\right) \\
&\geq \frac{1}{4}\exp\left(-n\chi^2(\mathrm{Bern}(p_0), \mathrm{Bern}(p_1))\right) \\
&\geq \frac{1}{4}\exp(-8n(p_0 - p_1)^2),
\end{aligned}$$

where the last inequality is due to $p_0, p_1 \in [1/4, 3/4]$ as $\varepsilon < 1/4$. By assumption, the minimax risk is upper bounded by δ, so $n \geq \log(1/(4\delta))/(32\varepsilon^2)$, as desired.

(c) In class we have shown that for the null distribution P_0 and a mixture distribution P_1, it holds that

$$\chi^2(P_1, P_0) = O\left(\frac{n^2\varepsilon^4}{k}\right).$$

Therefore,

$$1 - \|P_1 - P_0\|_{\mathrm{TV}} \geq \frac{1}{2}\exp(-\chi^2(P_1, P_0)) = \frac{1}{2}\exp\left(-O\left(\frac{n^2\varepsilon^4}{k}\right)\right).$$

The LHS is the smallest sum of test errors under H_0 and H_1, and by assumption is upper bounded by 2δ. This gives $n = \Omega(\sqrt{k\log(1/\delta)}/\varepsilon^2)$, as desired.

4. Consider a multi-armed bandit problem with K arms and two possible scenarios: the reward of arm $i \in [K]$ follows the distribution μ_i in the first scenario, and the distribution ν_i in the second scenario; the rewards across different times are independent. Now consider a generic policy $\pi = (\pi_1, \cdots, \pi_T)$, where for each time t, the action $\pi_t \in [K]$ depends causally on the historic observations $(\pi_1, r_{1,\pi_1}, \pi_2, r_{2,\pi_2}, \cdots, \pi_{t-1}, r_{t,\pi_{t-1}})$, where $r_{t,i} \sim \mu_i$ or ν_i denotes the random reward of arm i at time t. Let $P_{\mu,\pi}^T$ be the probability distribution of all observations under policy π and the first scenario, and $P_{\nu,\pi}^T$ is defined similarly under the second scenario. Moreover, for any $i \in [K]$, let $N_i = \sum_{t=1}^T \mathbb{1}(\pi_t = i)$ be the number of times that arm i is pulled. Show that

$$D_{\mathrm{KL}}(P_{\mu,\pi}^T \| P_{\nu,\pi}^T) = \sum_{i=1}^K \mathbb{E}_{P_{\mu,\pi}^T}[N_i] \cdot D_{\mathrm{KL}}(\mu_i \| \nu_i).$$

Solution: Using the chain rule of KL divergence, we have

$$D_{\mathrm{KL}}(P_{\mu,\pi}^T \| P_{\nu,\pi}^T) = \sum_{t=1}^T D_{\mathrm{KL}}(P_\mu(\pi_t, r_{t,\pi_t} \mid \mathcal{H}_t) \| P_\nu(\pi_t, r_{t,\pi_t} \mid \mathcal{H}_t) \mid P_\mu(\mathcal{H}_t)),$$

where $\mathcal{H}_t = (\pi_1, r_{1,\pi_1}, \pi_2, r_{2,\pi_2}, \cdots, \pi_{t-1}, r_{t,\pi_{t-1}})$ is the history available at the beginning of time t. Further write each term as

$$\begin{aligned}&D_{\mathrm{KL}}(P_\mu(\pi_t, r_{t,\pi_t} \mid \mathcal{H}_t) \| P_\nu(\pi_t, r_{t,\pi_t} \mid \mathcal{H}_t) \mid P_\mu(\mathcal{H}_t)) \\ &= D_{\mathrm{KL}}(P_\mu(\pi_t \mid \mathcal{H}_t) \| P_\nu(\pi_t \mid \mathcal{H}_t) \mid P_\mu(\mathcal{H}_t)) + D_{\mathrm{KL}}(P_\mu(r_{t,\pi_t} \mid \mathcal{H}_t, \pi_t) \| P_\nu(r_{t,\pi_t} \mid \mathcal{H}_t, \pi_t) \mid P_\mu(\mathcal{H}_t, \pi_t)).\end{aligned}$$

Since π_t only depends on the history $\mathcal{H}_t$, we have $P_\mu(\pi_t \mid \mathcal{H}_t) = P_\nu(\pi_t \mid \mathcal{H}_t)$, and the first term is zero. As for the second term, note that $P_\mu(r_{t,\pi_t} \mid \mathcal{H}_t, \pi_t = i) = \mu_i$ for each $i \in [K]$. Consequently,

$$\begin{aligned}&D_{\mathrm{KL}}(P_\mu(r_{t,\pi_t} \mid \mathcal{H}_t, \pi_t) \| P_\nu(r_{t,\pi_t} \mid \mathcal{H}_t, \pi_t) \mid P_\mu(\mathcal{H}_t, \pi_t)) \\ &= \sum_{\mathcal{H}_t} \sum_{i=1}^K D_{\mathrm{KL}}(\mu_i \| \nu_i) \cdot P_\mu(\mathcal{H}_t, \pi_t = i) = \sum_{i=1}^K D_{\mathrm{KL}}(\mu_i \| \nu_i) \cdot \mathbb{E}_{P_{\mu,\pi}^T}[\mathbb{1}(\pi_t = i)].\end{aligned}$$

Summing over $t \in [T]$ completes the proof.

5. Consider the following Gaussian sequence model $X_i = \theta_i + Z_i$ for $i \in [p]$, where the parameter vector $\theta = (\theta_1, \cdots, \theta_p)$ could take any value in $\mathbb{R}^p$, and $Z_1, \cdots, Z_p \sim \mathcal{N}(0,1)$ are iid standard normal noises. Consider the target of estimating $\theta_{\max} = \max_{i\in[p]} \theta_i$, with the loss function $L(\theta, T) = (T - \theta_{\max})^2$.

(a) Show that there exists an absolute constant $c > 0$ independent of p such that

$$\inf_T \sup_{\theta\in\mathbb{R}^p} \mathbb{E}_\theta[(T - \theta_{\max})^2] \geq c \cdot \log p.$$

(b) Propose an estimator T such that

$$\sup_{\theta\in\mathbb{R}^p} \mathbb{E}_\theta[(T - \theta_{\max})^2] \leq C \cdot \log p,$$

for some absolute constant $C < \infty$ independent of p.

Note: a careful analysis could give the tight constant $1/2$*:*

$$\inf_T \sup_{\theta\in\mathbb{R}^p} \mathbb{E}_\theta[(T - \theta_{\max})^2] = \left(\frac{1}{2} + o_p(1)\right) \cdot \log p.$$

Solution:

(a) Consider the following two hypotheses: $H_0 : \theta = 0$ and $H_1 : \theta \sim \mathsf{Unif}(\{\tau e_1, \cdots, \tau e_p\})$, where $e_1, \cdots, e_p$ are canonical vectors in $\mathbb{R}^p$, and $\tau > 0$ is a parameter to be determined later. The separation condition holds with $\Delta = \tau^2/2$, and

$$\chi^2(P_1, P_0) = \mathbb{E}[\exp(\theta^\top \theta') - 1] = \frac{\exp(\tau^2) - 1}{p},$$

where θ' is an independent of θ following the distribution under H_1. Consequently, if $\tau = \sqrt{(1-\varepsilon)\log p}$ with any $\varepsilon > 0$, we have $\chi^2(P_1, P_0) \to 0$. Using

$$\|P_1 - P_0\|_{\text{TV}} \leq \sqrt{\frac{D_{\text{KL}}(P_1\|P_0)}{2}} \leq \sqrt{\frac{\chi^2(P_1, P_0)}{2}},$$

we have $\|P_1 - P_0\|_{\text{TV}} \to 0$ as well. Consequently, the two-point method gives

$$\inf_T \sup_{\theta\in\mathbb{R}^p} \mathbb{E}_\theta[(T - \theta_{\max})^2] \geq \frac{(1-\varepsilon)\log p}{4}$$

for every $\varepsilon > 0$ as $p \to \infty$.

(b) A natural estimator is $T = X_{\max} = \max_{i\in[p]} X_i$. Using the Gaussian concentration property that $\mathbb{E}[\|Z\|_\infty^2] \leq (2 + o_p(1)) \log p$, we conclude that

$$\mathbb{E}_\theta[(X_{\max} - \theta_{\max})^2] \leq \mathbb{E}_\theta[\|X - \theta\|_\infty^2] = \mathbb{E}[\|Z\|_\infty^2] \leq (2 + o_p(1)) \log p.$$