Homework 3 Simple Solution

TA: Heng-Chien Liou*

1. (Binary hypothesis testing) [16]

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. Bernoulli p random variables, that is,

$$\mathsf{Pr}\{X_i = 1\} = 1 - \mathsf{Pr}\{X_i = 0\} = p.$$

Based on the observations so far, the goal is of a decision maker to determine which of the following two hypotheses is true:

$$\begin{aligned} \mathcal{H}_0 &: \quad p = p_0 \\ \mathcal{H}_1 &: \quad p = p_1 \end{aligned}$$

where $0 < p_0 < p_1 \le 1/2$.

a) (Warm-up) Consider the problem of making the decision based on X_1.
Draw the optimal $(\pi_{1|0}, \pi_{0|1})$ trade-off curve. [4]

b) Suppose the decision maker waits until an 1 appears and makes the decision based on the whole observed sequence. Sketch the optimal $(\pi_{1|0}, \pi_{0|1})$ trade-off curve. [4]

c) Now suppose the decision maker waits until in total n 1's appear and makes the decision based on the whole observed sequence. Let $\varpi^*_{0|1}(n, \epsilon)$ denote the minimum type-II error probability subject to the constraint that the type-I error probability is not greater than ϵ, $0 < \epsilon < 1$. Does $\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{\varpi^*_{0|1}(n,\epsilon)}$ exist? If so, find it. Otherwise, show that the limit does not exist. [8]

Solution:

a) By Neyman-Pearson theorem, the optimal test is randomized LRT. Note that the likelihood ratio can only take two values: $\frac{p_1}{p_0}, \frac{1-p_1}{1-p_0}$. Therefore, discuss the range of τ we get

$$\begin{cases} \pi_{1|0} = 1, \pi_{0|1} = 0, & 0 \le \tau < \frac{1-p_1}{1-p_0} \\ \pi_{1|0} = p_0 + \gamma(1-p_0), \pi_{0|1} = (1-\gamma)(1-p_1) = \frac{1-p_1}{1-p_0}(1-\pi_{1|0}), & \tau = \frac{1-p_1}{1-p_0} \\ \pi_{1|0} = p_0, \pi_{0|1} = 1 - p_1, & \frac{1-p_1}{1-p_0} < \tau < \frac{p_1}{p_0} \\ \pi_{1|0} = \gamma p_0, \pi_{0|1} = (1-\gamma)p_1 + (1-p_1) = 1 - \frac{p_1}{p_0}\pi_{1|0}, & \tau = \frac{p_1}{p_0} \\ \pi_{1|0} = 0, \pi_{0|1} = 1, & \tau > \frac{p_1}{p_0}. \end{cases}$$

*with contribution by Chen-Hao Hsiao

We can then draw the trade-off curve using the equations derived above.

b) Note that our observation can only be $1, 01, 001, 0001, \cdots$, let L be the length of the observation, we have

$$\begin{aligned} \mathcal{H}_0: &\quad L \sim \text{Geo}(p_0) \\ \mathcal{H}_1: &\quad L \sim \text{Geo}(p_1) \end{aligned}$$

Similar to a), we can discuss the range of τ and get:

$$\begin{cases} \pi_{1|0} = 0, \pi_{0|1} = 1, \ \tau > \frac{p_1}{p_0} \\ \pi_{1|0} = \sum\limits_{i=1}^{n-1} (1-p_0)^{i-1} p_0 + \gamma (1-p_0)^{n-1} p_0, \\ \pi_{0|1} = \sum\limits_{i=n+1}^{\infty} (1-p_1)^{i-1} p_1 + (1-\gamma)(1-p_1)^{n-1} p_1, \ \tau = \frac{(1-p_1)^{n-1} p_1}{(1-p_0)^{n-1} p_0} \\ \pi_{1|0} = \sum_{i=1}^{n} (1-p_0)^{i-1} p_0, \pi_{0|1} = \sum_{i=n+1}^{\infty} (1-p_1)^{i-1} p_1, \ \frac{(1-p_1)^{n-1} p_1}{(1-p_0)^{n-1} p_0} > \tau > \frac{(1-p_1)^{n} p_1}{(1-p_0)^{n} p_0}. \end{cases}$$

And we can draw the trade-off curve using the equations derived above.

c) The observation can be viewed as n i.i.d. geometric random variables. To see this, for any realization of observation, insert a "—" symbol in front of the sequence, also insert a "—" right after a "1". For example, if $n = 4$ and the realization is 010001101, we write it as $|01|0001|1|01|$. Apperantly, the length of the subsequence between two — is a geometric random variable. Hence, in this subproblem, we are testing $\text{Geo}(p_0)^{\otimes n}$ and $\text{Geo}(p_1)^{\otimes n}$. By Chernoff-Stein lemma,

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{\varpi^*_{0|1}(n,\epsilon)} = \mathrm{D}\left(\text{Geo}(p_0) \| \text{Geo}(p_1)\right) = \log \frac{p_0}{p_1} + \left(\frac{1-p_0}{p_0}\right) \log \frac{1-p_0}{1-p_1}.$$

Grading Policy:

a) Specify the optimal trade-off curve [2]; invoke Neyman-Pearson Theorem or directly argue optimality [2].

b) Formulate the problem as detecting geometric random variables (or directly used the mass function) [2] and invoke Neyman-Pearson's to specify the curve [2].

c) Formulate the problem as a hypothesis testing with n instances [3], Chernoff-Stein lemma [2], and calculation [3].

2. (Asymptotic behavior of posterior probability [12])

Consider a binary hypothesis testing problem

$$\begin{cases} \mathcal{H}_0: \ X_i \overset{\text{i.i.d.}}{\sim} \mathsf{P}_0, \ i = 1, 2, \ldots, n \\ \mathcal{H}_1: \ X_i \overset{\text{i.i.d.}}{\sim} \mathsf{P}_1, \ i = 1, 2, \ldots, n \end{cases}.$$

Under a Bayes setup, the unknown binary parameter Θ is assumed to be *random* and follow a prior distribution defined by the *prior probabilities*

$$\pi_0^{(0)} := \mathsf{Pr}\left\{\Theta = 0\right\}, \ \pi_1^{(0)} := \mathsf{Pr}\left\{\Theta = 1\right\}.$$

Let the *posterior probabilities* be the conditional distribution of Θ given $X^n = x^n$:

$$\pi_0^{(n)}(x^n) = \mathsf{Pr}\{\Theta = 0 | X^n = x^n\}, \ \pi_1^{(n)}(x^n) = \mathsf{Pr}\{\Theta = 1 | X^n = x^n\}.$$

a) Derive the expressions of $\pi_0^{(n)}(x^n)$ and $\pi_1^{(n)}(x^n)$ in terms of $\pi_0^{(0)}, \pi_1^{(0)}, \mathsf{P}_0, \mathsf{P}_1$. [4]

b) Consider $\pi_0^{(n)}(X^n)$ and $\pi_1^{(n)}(X^n)$ as random variables, because they are functions of the random sequence X^n. Use the Strong Law of Large Numbers to show that if $\mathcal{H}_0$ is true, then with probability 1,

$$\pi_0^{(n)}(X^n) \to 1, \ -\frac{1}{n}\log \pi_1^{(n)}(X^n) \to \mathrm{D}(\mathsf{P}_0 \| \mathsf{P}_1) \quad \text{as } n \to \infty. \quad [8]$$

Solution:

a) Denote $x^n = (x_1, x_2, ..., x_i, ..., x_n)$

$$\begin{aligned}
\pi_0^{(n)}(X^n) &= \mathsf{Pr}\{\Theta = 0 | X^n = x^n\} = \frac{\mathsf{Pr}\{\Theta = 0, X^n = x^n\}}{\mathsf{Pr}\{X^n = x^n\}} \\
&= \frac{\pi_0^{(0)}\mathsf{Pr}\{X^n = x^n | \Theta = 0\}}{\pi_0^{(0)}\mathsf{Pr}\{X^n = x^n | \Theta = 0\} + \pi_1^{(0)}\mathsf{Pr}\{X^n = x^n | \Theta = 1\}} \\
&= \frac{\pi_0^{(0)}\prod_{i=1}^n \mathsf{P}_0(x_i)}{\pi_0^{(0)}\prod_{i=1}^n \mathsf{P}_0(x_i) + \pi_1^{(0)}\prod_{i=1}^n \mathsf{P}_1(x_i)} \\
\pi_1^{(n)}(X^n) &= \frac{\pi_1^{(0)}\prod_{i=1}^n \mathsf{P}_1(x_i)}{\pi_0^{(0)}\prod_{i=1}^n \mathsf{P}_0(x_i) + \pi_1^{(0)}\prod_{i=1}^n \mathsf{P}_1(x_i)}
\end{aligned}$$

b) Recall the basic Bayes' theorem.

$$\begin{aligned}
\pi_0^{(n)}(X^n) &= \frac{\pi_0^{(0)}\prod_{i=1}^n \mathsf{P}_0(x_i)}{\pi_0^{(0)}\prod_{i=1}^n \mathsf{P}_0(x_i) + \pi_1^{(0)}\prod_{i=1}^n \mathsf{P}_1(x_i)} \\
&= \frac{\pi_0^{(0)}2^{-n\left(-\frac{1}{n}\sum_{i=1}^n \log \mathsf{P}_0(x_i)\right)}}{\pi_0^{(0)}2^{-n\left(-\frac{1}{n}\sum_{i=1}^n \log \mathsf{P}_0(x_i)\right)} + \pi_1^{(0)}2^{-n\left(-\frac{1}{n}\sum_{i=1}^n \log \mathsf{P}_1(x_i)\right)}} \\
&= \frac{1}{1 + \left(\frac{\pi_1^{(0)}}{\pi_0^{(0)}}\right)2^{-n\left(\frac{1}{n}\sum_{i=1}^n \log \frac{\mathsf{P}_0(x_i)}{\mathsf{P}_1(x_i)}\right)}} = \frac{1}{1 + c2^{-n\left(\frac{1}{n}\sum_{i=1}^n \log \frac{\mathsf{P}_0(x_i)}{\mathsf{P}_1(x_i)}\right)}},
\end{aligned}$$

where we let $c = \pi_1^{(0)}/\pi_0^{(0)}$. For any $\epsilon \in (0, \mathrm{D}(\mathsf{P}_0 \| \mathsf{P}_1))$, there is an $N_\epsilon \in \mathbb{N}$ such that, for every $n \geq N_\epsilon$

$$\mathsf{Pr}\left\{\left|\frac{1}{n}\sum_{i=1}^n \log \frac{\mathsf{P}_0(X_i)}{\mathsf{P}_1(X_i)} - \mathrm{D}(\mathsf{P}_0 \| \mathsf{P}_1)\right| < \epsilon\right\} = 1$$

$$\mathsf{Pr}\left\{\frac{1}{1+c2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)-\epsilon)}} < \pi_0^{(n)}\left(X^n\right) < \frac{1}{1+c2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)+\epsilon)}}\right\} = 1$$

Hence, with probability 1, $\pi_0^{(n)}\left(X^n\right) \to 1$ as $n \to \infty$.

Following this derivation, for every $n \geq N_\epsilon$, with probability 1,

$$-\frac{1}{n}\log\left(\frac{c2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)-\epsilon)}}{1+c2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)-\epsilon)}}\right) < -\frac{1}{n}\log\pi_1^{(n)}\left(X^n\right) < -\frac{1}{n}\log\left(\frac{c2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)+\epsilon)}}{1+c2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)+\epsilon)}}\right)$$

Furthermore,

$$\begin{aligned}
-\frac{1}{n}\log\left(\frac{c2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)+\epsilon)}}{1+c2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)+\epsilon)}}\right) &= \mathrm{D}\left(\mathsf{P}_0\|\mathsf{P}_1\right)+\epsilon-\frac{\log c}{n}+\frac{1}{n}\log\left(1+c2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)+\epsilon)}\right)\\
&\leq \mathrm{D}\left(\mathsf{P}_0\|\mathsf{P}_1\right)+\epsilon-\frac{\log c}{n}+\frac{1}{n}\log\left(1+c2^{-N_\epsilon(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)+\epsilon)}\right)\\
&= \mathrm{D}\left(\mathsf{P}_0\|\mathsf{P}_1\right)+\epsilon+O\left(\frac{1}{n}\right)\\
-\frac{1}{n}\log\left(\frac{c2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)-\epsilon)}}{1+c2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)-\epsilon)}}\right) &= \mathrm{D}\left(\mathsf{P}_0\|\mathsf{P}_1\right)-\epsilon-\frac{\log c}{n}+\frac{1}{n}\log\left(1+c2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)-\epsilon)}\right)\\
&\geq \mathrm{D}\left(\mathsf{P}_0\|\mathsf{P}_1\right)-\epsilon-\frac{\log c}{n}+\frac{1}{n}\log\left(1+c2^{-N_\epsilon(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)-\epsilon)}\right)\\
&= \mathrm{D}\left(\mathsf{P}_0\|\mathsf{P}_1\right)-\epsilon+O\left(\frac{1}{n}\right)
\end{aligned}$$

Hence, with probability 1,

$$\mathrm{D}\left(\mathsf{P}_0\|\mathsf{P}_1\right)-\epsilon \leq \lim_{n\to\infty} -\frac{1}{n}\log\pi_1^{(n)}\left(X^n\right) \leq \mathrm{D}\left(\mathsf{P}_0\|\mathsf{P}_1\right)+\epsilon,$$

where $\epsilon \in (0, \mathrm{D}\left(\mathsf{P}_0\|\mathsf{P}_1\right))$ arbitrarily. Hence,

$$\mathsf{Pr}\left\{\lim_{n\to\infty} -\frac{1}{n}\log\pi_1^{(n)}\left(X^n\right) = \mathrm{D}\left(\mathsf{P}_0\|\mathsf{P}_1\right)\right\} = 1$$

Grading Policy:

a) $\pi_0^{(n)}\left(x^n\right)$ [2] and $\pi_1^{(n)}\left(x^n\right)$ [2].

b) Application of SLLN and argumnet for almost sure convergence [4], convergence for $\pi_0^{(n)}\left(X^n\right)$ [2], convergence for $-(1/n)\log\pi_1^{(n)}\left(X^n\right)$ [2].

3. Minimizing information divergence) [22]

a) Let $\mathcal{P}(\mathbb{N})$ denote the collection of all probability distributions over $\mathbb{N}$ and $\mathrm{G}(p) \in \mathcal{P}(\mathbb{N})$ be a geometric distribution with parameter $p \in (0,1)$:

$$X \sim \mathrm{G}(p) \iff \mathsf{Pr}\{X = n\} = (1-p)p^{n-1},\ n \in \mathbb{N} = \{1,2,\ldots\}.$$

Under the constraint that $\mathsf{P} \in \mathcal{P}(\mathbb{N})$ and $\mathsf{E}_{X\sim\mathsf{P}}[X] = \sum_{x=1}^{\infty} x\mathsf{P}(x) = \mu > 1$, find the minimum value of $\mathrm{D}(\mathsf{P}\|\mathrm{G}(p))$ and a minimizing distribution. [12]

b) For m discrete probability distributions $\mathsf{P}_1, \mathsf{P}_2, \dots, \mathsf{P}_m$ with the same support $\mathcal{X}$, consider the following minimization problem:

$$\min_{\mathsf{Q}\in\mathcal{P}(\mathcal{X})} \sum_{i=1}^{m} \mathrm{D}(\mathsf{P}_i\|\mathsf{Q}),$$

where $\mathcal{P}(\mathcal{X})$ denotes the collection of probability distributions over $\mathcal{X}$. Find a minimizer to the above problem. [10]

Solution:

a) Expanding the Kullback-Leibler divergence over countable alphabet gives:

$$\begin{aligned}
\mathrm{D}(\mathsf{P}\|\mathrm{G}(p)) &= \sum_{x=1}^{\infty} \mathsf{P}(x)\log\frac{\mathsf{P}(x)}{(1-p)p^{x-1}} \\
&= \sum_{x=1}^{\infty} \mathsf{P}(x)\log\mathsf{P}(x) - \log(1-p) - \log p\sum_{x=1}^{\infty}\mathsf{P}(x)(x-1) \\
&= -\mathrm{H}(\mathsf{P}) - \log(1-p) - (\mu-1)\log p \\
&\geq -\mu\mathsf{h_b}\left(\frac{1}{\mu}\right) - \log(1-p) - (\mu-1)\log p,
\end{aligned}$$

where the inequality comes from Problem 1 of Homework 2, and $\mathsf{P}^* = \mathrm{G}\left(1-\mu^{-1}\right)$ is the minimizing distribution.

b) Let $\overline{\mathsf{P}} = \frac{1}{m}\sum_{i=1}^{m}\mathsf{P}_i$, $\overline{\mathsf{P}} \in \mathcal{P}(\mathcal{X})$.

$\forall \mathsf{Q} \in \mathcal{P}(\mathcal{X})$,

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}\mathrm{D}(\mathsf{P}_i\|\mathsf{Q}) - \frac{1}{m}\sum_{i=1}^{m}\mathrm{D}\left(\mathsf{P}_i\|\overline{\mathsf{P}}\right) \\
=&\frac{1}{m}\sum_{i=1}^{m}\sum_{x\in\mathcal{X}}\mathsf{P}_i(x)\log\frac{\overline{\mathsf{P}}(x)}{\mathsf{Q}(x)} \\
=&\sum_{x\in\mathcal{X}}\left(\frac{1}{m}\sum_{i=1}^{m}\mathsf{P}_i(x)\right)\log\frac{\overline{\mathsf{P}}(x)}{\mathsf{Q}(x)} \\
=&\sum_{x\in\mathcal{X}}\overline{\mathsf{P}}(x)\log\frac{\overline{\mathsf{P}}(x)}{\mathsf{Q}(x)} \\
=&\mathrm{D}\left(\overline{\mathsf{P}}\|\mathsf{Q}\right) \geq 0.
\end{aligned}$$

Hence, $\overline{\mathsf{P}}$ is a minimizer.

Grading Policy

a) Fine the tight lower bound [8] and find minimizer to achieve minimum [4].

b) Specify the minimizer [3] and justify it [7].