Name: ______________________________

Question:	1	2	3	Total
Points:	8	12	15	35
Score:				

E1 244 - Detection & Estimation Theory (2019) Midterm Exam

Instructions

- The total time for this test is 1.5 hours.
- Write your name at the top of this cover sheet.
- Attach your solution sheets to this cover sheet and return everything including rough work.
- No class notes, calculators or electronic aids are permitted.
- Academic dishonesty will not be tolerated.

Useful formulas and definitions:

- **Exponential probability distribution.** The exponential probability distribution with mean $\beta > 0$ is defined by the probability density function $p(x) = \frac{1}{\beta}e^{-\frac{x}{\beta}}$ if $x \geq 0$ and $p(x) = 0$ if $x < 0$.
- **Gaussian probability distribution.** The Gaussian probability distribution $\mathcal{N}(\mu, \sigma^2)$ is defined by the probability density function $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{x^2}{2\sigma^2}}$, $x \in \mathbb{R}$.
- $\log 2 = 0.69$.

1. **(8 points) Traffic Intensity Testing**
Vehicles arrive at the Yeshwanthpur traffic junction in Bengaluru with independent inter-arrival times distributed according to the exponential distribution with mean either $\frac{1}{2}$s. or 1s. depending on whether the type of traffic is 'heavy' or 'light', respectively. Suppose you measure the first 10 vehicle inter-arrival times (in s.) to be 0.06, 0.24, 0.17, 0.07, 1.80, 0.25, 1.37, 0.42, 0.48, 0.69. Assuming that both the types of traffic are equally likely *a priori*, what would you guess about the traffic type to minimize the probability of incorrect guessing and why?

Solution: We first compute the optimal Bayes decision rule for the hypotheses

$$\begin{aligned} &H_0 \,(\text{'heavy'}) : Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} p_0(y) = 2e^{-2y}\mathbb{1}\{y \geq 0\} \\ &\quad \text{vs.} \\ &H_1 \,(\text{'light'}) : Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} p_1(y) = e^{-y}\mathbb{1}\{y \geq 0\}, \end{aligned}$$

with uniform costs and equal priors.

This must be the likelihood ratio test given by (assuming observations are non-negative) $\delta(\underline{y}) = 1$ if and only if

$$\begin{aligned} 0 &\leq \log \frac{p_1(\underline{y})}{p_0(\underline{y})} \\ &= \log \prod_{i=1}^{n} \frac{e^{-y_i}}{2e^{-2y_i}} \\ &= \sum_{i=1}^{n} \log \frac{e^{-y_i}}{2e^{-2y_i}} = -n \log 2 + \sum_{i=1}^{n} y_i. \end{aligned}$$

For the given data, $\sum_{i=1}^{n} y_i = 5.55 < 10 \log 2 = 6.9$, so this test returns $\delta(\underline{y}) = 0$ which corresponds to the guess H_0 **('heavy' traffic)**.

2. **Signal Detection**
Consider the signal model

$$Y_k = \sqrt{\theta}\,(s_k R_k + N_k)\,, \quad 1 \leq k \leq n,$$

where $s_1, \ldots, s_n$ is a known signal sequence, $\theta \geq 0$ is a constant and $R_1, \ldots, R_n$, $N_1, \ldots, N_n$ are i.i.d. $\mathcal{N}(0, 1)$ random variables.

(a) **(6 points)** Consider the hypothesis pair

$$\begin{aligned} &H_0 : \theta = \frac{a}{2} \\ &\quad \text{vs.} \\ &H_1 : \theta = a, \end{aligned}$$

where $a > 0$ is a known constant. Describe the structure of the Neyman-Pearson optimal detector for given false-alarm level $\alpha \in (0,1)$. Derive its receiver operating characteristic (ROC) in terms of the standard normal cdf Φ for the signal sequence $s_1 = \cdots = s_n = 1$.

Solution: The Neyman-Pearson optimal test statistic is the likelihood ratio

$$L_n = \frac{p_1(Y_1,\ldots,Y_n)}{p_0(Y_1,\ldots,Y_n)}.$$

Under H_0, each $Y_k \sim \mathcal{N}(0, as_k^2/2 + a/2) = \mathcal{N}(0, \sigma_{0k}^2)$ independently, while under H_1, each $Y_k \sim \mathcal{N}(0, as_k^2 + a) = \mathcal{N}(0, \sigma_{1k}^2)$ independently, so

$$L_n = \prod_{k=1}^n \frac{\frac{1}{\sigma_{1k}} e^{-\frac{Y_k^2}{2\sigma_{1k}^2}}}{\frac{1}{\sigma_{0k}} e^{-\frac{Y_k^2}{2\sigma_{0k}^2}}} = \prod_{k=1}^n \frac{1}{\sqrt{2}}\, e^{-Y_k^2\left(\frac{1}{2\sigma_{1k}^2} - \frac{1}{2\sigma_{0k}^2}\right)} = \prod_{k=1}^n \frac{1}{\sqrt{2}}\, e^{-\frac{Y_k^2}{2a(s_k^2+1)}\left(\frac{1}{1} - \frac{1}{1/2}\right)}$$
$$= 2^{-n/2} \exp\left(\frac{1}{2a}\sum_{k=1}^n \frac{Y_k^2}{s_k^2+1}\right).$$

Note that thresholding this is equivalent to thresholding the statistic $T(\underline{Y}) = \sum_{k=1}^n \frac{Y_k^2}{s_k^2+1}$, which is the sum of weighted squares of independent Gaussian random variables under each hypothesis. The optimal Neyman-Pearson detector uses the threshold η that solves $\mathbb{P}_0[T(\underline{Y}) = \eta] = \alpha$, since the distribution of $T(\underline{Y})$ admits a valid probability density function.

When all the $s_k = 1$, $1 \le k \le n$, a valid test statistic is $\tilde{T}(\underline{Y}) = \sum_{k=1}^n Y_k^2$, which is a sum of iid chi-square (square of zero-mean Gaussian) random variables under each hypothesis. Specifically, under H_0, each $Y_k \sim \mathcal{N}(0,a) \Leftrightarrow Y_k/\sqrt{a} \sim \mathcal{N}(0,1)$, so

$$\mathbb{P}\left[Y_k^2 \le x\right] = \mathbb{P}\left[\frac{Y_k}{\sqrt{a}} \in \left[-\sqrt{\frac{x}{a}}, \sqrt{\frac{x}{a}}\right]\right] = \Phi\left(\sqrt{\frac{x}{a}}\right) - \Phi\left(-\sqrt{\frac{x}{a}}\right).$$

Thus the density of each Y_k is

$$f_{Y_k}^{(0)}(x) = \frac{1}{\sqrt{ax}}\,\Phi'\left(\sqrt{\frac{x}{a}}\right), \quad x > 0$$

under H_0, and similarly

$$f_{Y_k}^{(1)}(x) = \frac{1}{\sqrt{2ax}}\,\Phi'\left(\sqrt{\frac{x}{2a}}\right), \quad x > 0$$

under H_1. From these expressions, the ROC can be worked out using the n-fold convolution of probability densities.

(b) **(6 points)** Consider the hypothesis pair

$$H_0 : \theta = \frac{a}{2}$$
$$\text{vs.}$$
$$H_1 : \theta > \frac{a}{2}.$$

Does a Uniformly Most Powerful (UMP) detector exist (why/why not)? What about a Locally Most Powerful (LMP) detector (why/why not)?

Solution: A calculation similar to that in part (a) shows that the optimal Neyman-Pearson test for $H_0 : \theta = a/2$ vs. $H_1 : \theta = \theta' > a/2$ thresholds the quantity

$$L_n = \prod_{k=1}^{n} \frac{\frac{1}{\sigma_{1k}} e^{-\frac{Y_k^2}{2\sigma_{1k}^2}}}{\frac{1}{\sigma_{0k}} e^{-\frac{Y_k^2}{2\sigma_{0k}^2}}} = \prod_{k=1}^{n} \sqrt{\frac{a}{2\theta'}}\, e^{-Y_k^2\left(\frac{1}{2\sigma_{1k}^2} - \frac{1}{2\sigma_{0k}^2}\right)} = \prod_{k=1}^{n} \sqrt{\frac{a}{2\theta'}}\, e^{-\frac{Y_k^2}{2(s_k^2+1)}\left(\frac{1}{\theta'} - \frac{1}{a/2}\right)}$$
$$= \left(\frac{a}{2\theta'}\right)^{n/2} \exp\left(\frac{1}{2}\left(-\frac{1}{\theta'} + \frac{1}{a/2}\right)\sum_{k=1}^{n} \frac{Y_k^2}{s_k^2+1}\right),$$

or, equivalently, the statistic

$$\sum_{k=1}^{n} \frac{Y_k^2}{s_k^2+1}.$$

But this does not depend on θ', so a Uniformly Most Powerful (UMP) detector exists.

3. **Independence Testing**

Suppose you want to determine whether two random variables A and B are independent, by observing pairs $(A_1, B_1), (A_2, B_2), \ldots$ independently sampled from a joint distribution. More specifically, consider testing the hypotheses

$H_0 : A_i \sim \mathcal{N}(0,1), B_i = A_i + Z_i, Z_i \sim \mathcal{N}(0,1), A_i \perp\!\!\!\perp Z_i, i = 1, 2, \ldots, n$
(not independent)

vs.

$H_1 : A_i \sim \mathcal{N}(0,1), B_i \sim \mathcal{N}(0,2), A_i \perp\!\!\!\perp B_i, i = 1, 2, \ldots, n$
(independent),

where $X \perp\!\!\!\perp Y$ denotes that X, Y are independent random variables, and any random variables with different indices i are assumed independent. Note that the marginal distributions of A_i and B_i are the same under both the hypotheses.

(a) **(3 points)** Find the log-likelihood ratio for the i-th observed pair (A_i, B_i) as a function of A_i and B_i.

Solution: The log-likelihood ratio for the i-th sample pair is

$$
\begin{aligned}
\log \frac{p_1(A_i, B_i)}{p_0(A_i, B_i)} &= \log \frac{p_1(A_i)p_1(B_i|A_i)}{p_0(A_i)p_0(B_i|A_i)} \\
&= \log \frac{p_1(B_i)}{p_0(B_i|A_i)} \quad \text{(since } A_i \perp\!\!\!\perp B_i \text{ under } H_1\text{)} \\
&= \log \frac{\frac{1}{\sqrt{2}\sqrt{2\pi}} e^{-\frac{B_i^2}{2\cdot 2}}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{(B_i - A_i)^2}{2}}} \\
&= \log \frac{1}{\sqrt{2}} + \frac{1}{2}\left((B_i - A_i)^2 - \frac{B_i^2}{2}\right) \\
&= -\frac{1}{2}\log 2 + \frac{1}{2}\left(\frac{B_i^2}{2} - 2B_iA_i + A_i^2\right).
\end{aligned}
$$

(b) **(6 points)** Under each hypothesis, find the expectation of the log-likelihood ratio $\log L_n$ for the <u>whole sequence</u> of observations $(A_1, B_1), (A_2, B_2), \ldots, (A_n, B_n)$.

Solution: Under H_0:

$$
\begin{aligned}
\mathbb{E}_0[\log L_n] &= \mathbb{E}_0\left[\log \prod_{i=1}^n \frac{p_1(A_i, B_i)}{p_0(A_i, B_i)}\right] \\
&= \sum_{i=1}^n \mathbb{E}_0\left[\log \frac{p_1(A_i, B_i)}{p_0(A_i, B_i)}\right] \\
&= \sum_{i=1}^n \mathbb{E}_0\left[-\frac{1}{2}\log 2 + \frac{1}{2}\left(\frac{B_i^2}{2} - 2B_iA_i + A_i^2\right)\right] \quad \text{(from the previous part)} \\
&= -\frac{n}{2}\log 2 + \frac{n}{2}\left(\frac{2}{2} - 2\cdot 1 + 1\right) \quad \text{(since } \mathbb{E}_0[A_iB_i] = 1\text{)} \\
&= -\frac{n}{2}\log 2,
\end{aligned}
$$

while under H_1,

$$
\begin{aligned}
\mathbb{E}_1[\log L_n] &= \mathbb{E}_1\left[\log \prod_{i=1}^n \frac{p_1(A_i, B_i)}{p_0(A_i, B_i)}\right] \\
&= \sum_{i=1}^n \mathbb{E}_1\left[-\frac{1}{2}\log 2 + \frac{1}{2}\left(\frac{B_i^2}{2} - 2B_iA_i + A_i^2\right)\right] \quad \text{(from the previous part)} \\
&= -\frac{n}{2}\log 2 + \frac{n}{2}\left(\frac{2}{2} - 2\cdot 0 + 1\right) \quad \text{(since } \mathbb{E}_1[A_iB_i] = 0\text{)} \\
&= -\frac{n}{2}\log 2 + n.
\end{aligned}
$$

(c) **(6 points)** Consider the hypothesis test that outputs H_1 if the log-likelihood ratio ($\log L_n$) of n observations exceeds a given threshold $\eta \in \mathbb{R}$, and H_0 otherwise. Suppose the number of pairs of observations (n) is extremely large (think 'almost ∞'). Argue, using the law of large numbers, why you would expect this test to detect the true hypothesis reliably, i.e., with very low error.

Solution: The quantity $\frac{1}{n}\log L_n = \frac{1}{n}\sum_{i=1}^{n}\log\frac{p_1(A_i,B_i)}{p_0(A_i,B_i)}$ is the average of the n iid random variables $\log\frac{p_1(A_i,B_i)}{p_0(A_i,B_i)}$, $i = 1,\ldots,n$. So by the law of large numbers, for n large enough and under any hypothesis H_j ($j = 0, 1$),

$$\log L_n = n \cdot \frac{1}{n}\log L_n \approx n \cdot \mathbb{E}_j\left[\log\frac{p_1(A_1,B_1)}{p_0(A_1,B_1)}\right] = \begin{cases} -\frac{n}{2}\log 2, & j = 0 \\ n\left(1 - \frac{1}{2}\log 2\right), & j = 1. \end{cases}$$

Since $-\frac{1}{2}\log 2 < 0 < 1 - \frac{1}{2}\log 2$, the argument above shows that with very high probability, $\log L_n$ i very negative or very positive if the true hypothesis is H_0 or H_1, respectively, which is why one can expect the test to perform reliably.