UC Berkeley
Department of Electrical Engineering and Computer Sciences

EECS 126: Probability and Random Processes

Homework 12
Spring 2023

1. **Flipping Coins and Hypothesizing**

You flip a coin until you see heads. Let $0 < p < q < 1$, and suppose that your hypotheses are

$$X = \begin{cases} 0 & \text{if the bias of the coin is } p \\ 1 & \text{if the bias of the coin is } q. \end{cases}$$

You observe Y, the number of flips until you see heads. Find a decision rule $\hat{X}$ that maximizes $\mathbb{P}(\hat{X} = 1 \mid X = 1)$ subject to $\mathbb{P}(\hat{X} = 1 \mid X = 0) \le \beta$ for some $\beta \in [0, 1]$.

Hint: Remember to calculate the randomization constant γ.

Solution:

1. The likelihood ratio is

$$L(y) = \frac{\mathbb{P}(Y = y \mid X = 1)}{\mathbb{P}(Y = y \mid X = 0)} = \frac{(1-q)^{y-1}q}{(1-p)^{y-1}p}.$$

As $p < q$, this is a strictly decreasing function of y, so the Neyman–Pearson decision rule will be of the form $\mathbb{1}_{Y<t}$ for some t.

2. The threshold t should satisfy $\mathbb{P}(Y < t \mid X = 0) = 1 - (1-p)^{t-1} \le \beta$, so we should take

$$t := 1 + \left\lfloor \frac{\log(1-\beta)}{\log(1-p)} \right\rfloor.$$

3. It is possible that $1 - (1-p)^{t-1} < \beta$ for the choice of the integer threshold above, so we will need to introduce randomization. Letting $\mathbb{P}(\hat{X} = 1 \mid Y = t) = \gamma$, the probability of false alarm becomes

$$\begin{aligned} \mathbb{P}(\hat{X} = 1 \mid X = 0) &= \mathbb{P}(Y < t \mid X = 0) + \gamma\,\mathbb{P}(Y = t \mid X = 0) \\ &= 1 - (1-p)^{t-1} + \gamma p(1-p)^{t-1}, \end{aligned}$$

so to achieve $\mathsf{PFA} = \beta$, we take the randomization constant

$$\gamma = \frac{\beta - 1 + (1-p)^{t-1}}{p(1-p)^{t-1}}.$$

The final decision rule is given by, for the values of t and γ above,

$$\hat{X} = \begin{cases} 1 & \text{if } Y < t \\ \text{Bernoulli}(\gamma) & \text{if } Y = t \\ 0 & \text{if } Y > t. \end{cases}$$

2. **One Flip**

   You flip a single coin and observe its result $Y \sim \text{Bernoulli}(p)$. Suppose the hypotheses are

$$X = \begin{cases} 0 & \text{if } p = \frac{1}{3} \\ 1 & \text{if } p = \frac{2}{3}. \end{cases}$$

   a. Find the MLE of X and its associated type I and type II error rates.

   b. Plot the error curve.

   c. Derive the randomized decision rule that minimizes type II error subject to the constraint of $\beta = 0.5$ on the type I error.

   Hint: You should only need to look at the plot from part b.

Solution:

a. We see that the MLE is simply Y itself:

$$\hat{X}_{\text{MLE}} = \mathbb{1}\{p_{Y|X}(Y \mid 1) \geq p_{Y|X}(Y \mid 0)\} = \begin{cases} 0 & \text{if } Y = 0 \\ 1 & \text{if } Y = 1. \end{cases}$$

The probability of type I error is $\mathbb{P}(\hat{X} = 1 \mid X = 0) = \mathbb{P}(Y = 1 \mid X = 0) = \frac{1}{3}$, and the probability of type II error is $\mathbb{P}(\hat{X} = 0 \mid X = 1) = \frac{1}{3}$.

b. Note that $\hat{X}_{\text{MLE}} = \mathbb{1}\{L(Y) \geq 1\}$. More generally, the likelihood ratio is

$$L(y) = \begin{cases} \frac{1}{2} & \text{if } y = 0 \\ 2 & \text{if } y = 1, \end{cases}$$

so the threshold test $\hat{X}_\lambda = \mathbb{1}\{L(Y) > \lambda\}$ is

$$\hat{X}_\lambda = \begin{cases} 1 & \text{if } \lambda \leq \frac{1}{2} \\ Y & \text{if } \frac{1}{2} < \lambda \leq 2 \\ 0 & \text{if } \lambda > 2. \end{cases}$$

The error rates of the possible threshold tests are $(1, 0)$, $(\frac{1}{3}, \frac{1}{3})$, and $(0, 1)$ respectively, and the error curve is the piecewise-linear function connecting these three points.

c. For a test to actually achieve the point $(\frac{1}{2}, \frac{1}{4})$ on the error curve, we will need to take a convex combination of simple threshold tests, or introduce *randomization*. From the plot, it is clear that taking Y w.p. $\frac{3}{4}$ and 1 w.p. $\frac{1}{4}$ gives us the Neyman–Pearson optimal decision rule subject to $\mathsf{PFA} \leq \frac{1}{2}$.

3. **Exam Difficulty**

The difficulty of an EECS 126 exam, Θ, is uniformly distributed on $[0, 100]$ (continuously). Alice gets a score X that is uniformly distributed on $[0, \Theta]$, and she wants to estimate the difficulty of the exam given her score.

a. What is the MLE of Θ? What is the MAP of Θ?

b. What is the LLSE for Θ?

Solution:

a. Since the prior on Θ is uniform, the MLE and MAP estimates will be the same. Both are equal to $\hat{\Theta} = X$, as

$$\operatorname*{argmax}_{\theta} f_{X|\Theta}(x \mid \theta) = \operatorname*{argmax}_{\theta} \frac{1}{\theta} \cdot \mathbb{1}_{x \le \theta \le 100} = x.$$

b. Recall that the LLSE of Θ given X can be found as

$$\mathbb{L}(\Theta \mid X) = \mathbb{E}(\Theta) + \frac{\operatorname{cov}(\Theta, X)}{\operatorname{var}(X)}(X - \mathbb{E}(X)).$$

First, $\mathbb{E}(\Theta) = 50$ and $\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X \mid \Theta)) = \mathbb{E}(\frac{\Theta}{2}) = 25$. Let us find $\operatorname{var}(X)$ using the law of total variance and $\operatorname{cov}(\Theta, X) = \mathbb{E}(\Theta X) - \mathbb{E}(\Theta)\,\mathbb{E}(X)$:

$$\mathbb{E}(\operatorname{var}(X \mid \Theta)) = \mathbb{E}\left(\frac{\Theta^2}{12}\right) = \int_0^{100} \frac{\theta^2}{12} \cdot \frac{1}{100}\,\mathrm{d}\theta = \frac{10000}{36}.$$

$$\operatorname{var}(\mathbb{E}(X \mid \Theta)) = \operatorname{var}\left(\frac{\Theta}{2}\right) = \frac{1}{4}\frac{10000}{12} = \frac{10000}{48}.$$

$$\operatorname{var}(X) = \mathbb{E}(\operatorname{var}(X \mid \Theta)) + \operatorname{var}(\mathbb{E}(X \mid \Theta)) = \frac{70000}{144}.$$

$$\mathbb{E}(\Theta X) = \mathbb{E}(\mathbb{E}(\Theta X \mid \Theta)) = \mathbb{E}\left(\frac{\Theta^2}{2}\right) = \frac{10000}{6}.$$

$$\operatorname{cov}(\Theta, X) = \mathbb{E}(\Theta X) - \mathbb{E}(\Theta)\,\mathbb{E}(X) = \frac{1250}{3}.$$

Putting everything together, the LLSE is

$$\mathbb{L}(\Theta \mid X) = 50 + \frac{6}{7}(X - 25).$$

4. **Gaussian LLSE**

   Let X, Y, Z be i.i.d. $\mathcal{N}(0,1)$.

   a. Find $\mathbb{L}(X^2 + Y^2 \mid X + Y)$.
   b. Find $\mathbb{L}(X + 2Y \mid X + 3Y + 4Z)$.
   c. Find $\mathbb{L}((X + Y)^2 \mid X - Y)$.

   Solution:

   a. We note that

   $$\text{cov}(X^2 + Y^2, X + Y) = \mathbb{E}((X^2 + Y^2)(X + Y)) = \mathbb{E}(X^3 + X^2 Y + XY^2 + Y^3) = 0.$$

   Thus, $\mathbb{L}(X^2 + Y^2 \mid X + Y) = \mathbb{E}(X^2 + Y^2) = 2$.

   b. We find that

   $$\begin{aligned} \text{cov}(X + 2Y, X + 3Y + 4Z) &= \mathbb{E}[(X + 2Y)(X + 3Y + 4Z)] = \mathbb{E}(X^2) + 6\,\mathbb{E}(Y^2) = 7 \\ \text{var}(X + 3Y + 4Z) &= \text{var}(X) + 9\,\text{var}(Y) + 16\,\text{var}(Z) = 26 \\ \mathbb{L}(X + 2Y \mid X + 3Y + 4Z) &= \frac{7}{26}(X + 3Y + 4Z). \end{aligned}$$

   c. We observe that $\text{cov}(X + Y, X - Y) = 0$, so that the jointly Gaussian $X + Y$ and $X - Y$ are independent. Hence,

   $$\mathbb{L}((X + Y)^2 \mid X - Y) = \mathbb{E}((X + Y)^2) = \text{var}(X + Y) = 2.$$

5. **Projections**

The following exercises are from the note on the Hilbert space of random variables. See the notes for some hints.

a. Let $\mathcal{H} := \{X : X \text{ is a real-valued random variable with } \mathbb{E}(X^2) < \infty\}$. Prove that $\mathcal{H}$ is closed under addition and scalar multiplication over the real numbers $\mathbb{R}$, and prove that the function $\langle X, Y\rangle := \mathbb{E}(XY)$ is an inner product on $\mathcal{H}$. [1]

b. Let U be a subspace of a real inner product space V. We define the *projection* map P onto U as follows: for each $v \in V$, let Pv be the unique vector in U such that $v - Pv \in U^\perp$. Prove that P is a linear transformation.

c. Using part b, prove that $\mathbb{L}(X + Y \mid Z) = \mathbb{L}(X \mid Z) + \mathbb{L}(Y \mid Z)$ for all $X, Y, Z \in \mathcal{H}$.

d. Now, suppose that U is a finite-dimensional subspace, $\dim U := n$, with an orthonormal basis $\{u_i\}_{i=1}^n$. Prove that $Px = \sum_{i=1}^n \langle x, u_i\rangle u_i$ for all $x \in V$.

Solution:

a. Let $X, Y, Z \in \mathcal{H}$ and $c \in \mathbb{R}$. Then

$$\begin{aligned}\mathbb{E}((X+Y)^2) &= \mathbb{E}(X^2) + 2\,\mathbb{E}(XY) + \mathbb{E}(Y^2) \\ &\le \mathbb{E}(X^2) + 2\sqrt{\mathbb{E}(X^2)\,\mathbb{E}(Y^2)} + \mathbb{E}(Y^2)\end{aligned}$$

by the Cauchy–Schwarz inequality, and $\mathbb{E}(X^2), \mathbb{E}(Y^2) < \infty$ by hypothesis, which shows that $X + Y \in \mathcal{H}$. We also have $\mathbb{E}((cX)^2) = c^2\,\mathbb{E}(X^2) < \infty$, so $\mathcal{H}$ is closed under scalar multiplication as well.

Now, we check that $\langle X, Y\rangle = \mathbb{E}(XY)$ defines an inner product:

- $\mathbb{E}(XY) = \mathbb{E}(YX)$.
- $\mathbb{E}((X + cY)Z) = \mathbb{E}(XZ) + c\,\mathbb{E}(YZ)$ by the linearity of expectation.
- $\mathbb{E}(X^2) \ge 0$, with $\mathbb{E}(X^2) = 0$ if and only if $X = 0$. (See footnote.)

(The other properties in the definition of a vector space are familiar properties of random variables, so we have shown that $\mathcal{H}$ is a real inner product space. $\mathcal{H}$ is in particular also a *Hilbert space*, because it satisfies an analytic property called *completeness*.)

b. Let $x, y \in V$ and $c \in \mathbb{R}$. We wish to show that $P(x + cy) = Px + cPy$, and it suffices to check that $Px + cPy \in U$ and $x + cy - (Px + cPy) \in U^\perp$.

- $Px, Py \in U$ by definition of P and U is a subspace, so $Px + cPy \in U$.
- For any $u \in U$, we have $\langle u, x + y - Px - cPy\rangle = \langle u, x - Px\rangle + c\langle u, y - Py\rangle = 0$ by $x - Px, y - Py \in U^\perp$.

By definition of P, this shows that $Px + cPy = P(x + cy)$.

c. Let $V = \mathcal{H}$. Then $X \mapsto \mathbb{L}(X \mid Z)$ is the projection map onto the subspace $U = \operatorname{span}\{1, Z\}$, and we have shown that projections are linear in part b.

d. For $x \in V$, we check that $\sum_{i=1}^n \langle x, u_i\rangle u_i \in U$ and $x - \sum_{i=1}^n \langle x, u_i\rangle u_i \in U^\perp$.

[1] *Remark.* It is possible for $X \neq 0$ to have $\mathbb{E}(X^2) = 0$, e.g. if $X = 0$ with probability 1. To fix this, we can take almost-sure equivalence classes of random variables, where X and Y are equivalent if $\mathbb{P}(X = Y) = 1$. You may cite this construction when checking that $X \neq 0$ implies $\mathbb{E}(X^2) > 0$.

- $u_1, \ldots, u_n$ belong to the subspace U, so the linear combination $\sum_{i=1}^{n} \langle x, u_i \rangle u_i$ belongs to U as well.
- We want to show that $\langle u, (x - \sum_{i=1}^{n} \langle x, u_i \rangle u_i) \rangle = 0$ for all $u \in U$. By the linearity of an inner product, it suffices to show the claim for any basis vector u_j, $j = 1, \ldots, n$:

$$\left\langle u_j, x - \sum_{i=1}^{n} \langle x, u_i \rangle u_i \right\rangle = \langle u_j, x \rangle - \sum_{i=1}^{n} \langle x, u_i \rangle \langle u_j, u_i \rangle = \langle u_j, x \rangle - \langle x, u_j \rangle = 0.$$

  We used the fact that $\{u_i\}_{i=1}^{n}$ is an orthonormal basis of U, where $\langle u_i, u_i \rangle = 1$ for all $i = 1, \ldots, n$ and $\langle u_i, u_j \rangle = 0$ for all $i \neq j$.

6. **Sufficient Statistics**

Suppose $X_1, \dots, X_n$ are i.i.d. samples drawn from a probability distribution parameterized by θ. (We are in the non-Bayesian setting, so θ is deterministic but unknown).

A statistic $T(X_1, \dots, X_n)$ is a *sufficient statistic* for θ if for all t, the conditional distribution of $(X_1, \dots, X_n)$ given $T = t$ does not depend on θ. Intuitively, $T(X_1, \dots, X_n)$ "captures all there is to know about θ from the sample $X_1, \dots, X_n$."

a. Let $X_1, \dots, X_n$ be drawn i.i.d. from a Poisson distribution with mean μ. Show that $T = \sum_{i=1}^n X_i$ is a sufficient statistic for μ.

b. Let T be a sufficient statistic for θ, and let $\hat\theta$ be an estimator for θ with $\mathrm{var}(\hat\theta) < \infty$. Show that in mean-squared error sense, $\mathbb{E}[\hat\theta \mid T]$ is at least as good as $\hat\theta$ at estimating θ:

$$\mathbb{E}[(\mathbb{E}[\hat\theta \mid T] - \theta)^2] \le \mathbb{E}[(\hat\theta - \theta)^2].$$

Hint: Consider expanding the decomposition $\mathbb{E}[(\hat\theta - \theta)^2] = \mathbb{E}[((\hat\theta - \mathbb{E}[\hat\theta]) + (\mathbb{E}[\hat\theta] - \theta))^2]$.
Remark. Since $\mathbb{E}[\hat\theta \mid T]$ is a function of T, the result above suggests we should be looking for estimators of θ that are functions of sufficient statistics.

Solution:

a. We note that $T \sim \mathrm{Poisson}(n\mu)$ has $\mathbb{P}(T = t) = \frac{e^{-n\mu}(n\mu)^t}{t!}$, and

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n, T = t) = \prod_{i=1}^n \frac{e^{-\mu}\mu^{x_i}}{x_i!} \cdot \mathbb{1}_{t = \sum_{i=1}^n x_i} = \frac{e^{-n\mu}\mu^t}{\prod_{i=1}^n x_i!}\,\mathbb{1}_{t = \sum_{i=1}^n x_i}.$$

Then the conditional distribution is

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n \mid T = t) = \frac{t!}{n^t \prod_{i=1}^n x_i!}\,\mathbb{1}_{t = \sum_{i=1}^n x_i},$$

which has no dependence on μ.

b. Fron the hint, observe that

$$\begin{aligned}
\mathbb{E}[(\hat\theta - \theta)^2] &= \mathbb{E}[((\hat\theta - \mathbb{E}[\hat\theta]) + (\mathbb{E}[\hat\theta] - \theta))^2] \\
&= \mathbb{E}[(\hat\theta - \mathbb{E}[\hat\theta])^2] + \cancel{2\,\mathbb{E}[\hat\theta - \mathbb{E}[\hat\theta]](\mathbb{E}[\hat\theta] - \theta)} + \mathbb{E}[(\mathbb{E}[\hat\theta] - \theta)^2] \\
&= \mathrm{var}(\hat\theta) + (\mathbb{E}[\hat\theta] - \theta)^2.
\end{aligned}$$

This is commonly known as the *bias-variance decomposition* in machine learning contexts. Similarly, we have

$$\begin{aligned}
\mathbb{E}[(\mathbb{E}[\hat\theta \mid T] - \theta)^2] &= \mathrm{var}(\mathbb{E}[\hat\theta \mid T] - \theta) + (\mathbb{E}[\mathbb{E}[\hat\theta \mid T] - \theta])^2 \\
&= \mathrm{var}(\mathbb{E}[\hat\theta \mid T]) + (\mathbb{E}[\hat\theta] - \theta)^2
\end{aligned}$$

by the law of iterated expectation. From the law of total variance, $\mathrm{var}(\hat\theta) \ge \mathrm{var}(\mathbb{E}[\hat\theta \mid T])$, which proves the claim. This result is known as the **Rao–Blackwell theorem**.