Foundations of Data Science
Gastpar & Urbanke

Ecole Polytechnique Fédérale, Lausanne: Fall 2023
November 20, 2023

Problem Set 6 **(Graded)** —*Due Tuesday, December 5, before class starts*

For the Exercise Sessions on Nov 21 and Nov 28

Last name	First name	SCIPER Nr	Points

Problem 1: Property Testing: Variance

A colleague claims to have implemented an algorithm which outputs i.i.d. samples distributed according to a discrete distribution P that has unit variance. Your task is to design a statistic to test whether this is indeed true.

Let Δ_k be the set of probability distributions on the alphabet $\mathcal{X} = \{1, \cdots, k\}$. Assume that $P \in \mathcal{P} \cup \mathcal{Q}$ with $\mathcal{P} := \{P \in \Delta_k : P \text{ has variance } 1\}$ and $\mathcal{Q} := \{P \in \Delta_k : P \text{ has variance} \in [0, 1-\epsilon] \cup [1+\epsilon, \infty)\}$, where $0 < \epsilon < 1$. You are given n samples $\{X_i\}_{i=1}^n$, where the X_i are independent copies sampled according to P.

Remark: For the following three questions we do not ask you to write down a proof (or explicit calculation) that your proposed solution works.

a) We say that an estimator $e : S^n \mapsto \Pi$ on a sample S^n of length n (ϵ, δ)-learns a parameter $p \in \Pi$ if for any $(\epsilon, \delta) \in (0,1)^2$, given sufficiently many samples n, we have that $\mathbb{P}(\{|e(S^n) - p| > \epsilon\}) < \delta$. Give a brief explanation (one sentence, no calculations) why the empirical estimator of the second moment $\hat{\mu}_{X^2} := \frac{1}{n}\sum_{i=1}^n X_i^2$ can (ϵ, δ)-learn the second moment in our setting.

b) First, assume that a genie tells you that X has zero mean. Design a simple test statistic and give a threshold in order to check for the above mentioned unit variance property.

Hint: Use the claim in a).

c) Now consider the more general case where X can have arbitrary mean. Again, design a simple test statistic and give a threshold.

Hint: You can assume that $\hat{\mu}_X^2$ (ϵ, δ)-learns $\mathbb{E}[X]^2$.

Solution 1. a) The empirical mean of squares $\hat{\mu}_{X^2}$ concentrates strongly around the true second moment due to the boundedness of X^2 together with Hoeffding's inequality.

Alternatively: finite expectation of distribution P + (weak) law of large numbers.

b) The approach is analogous to the property test for uniformity as presented in the lecture. We first note that the variance is equal to the second moment. Compute the empirical mean of X^2, i.e., $\hat{\mu}_{X^2} := \frac{1}{n}\sum_{i=1}^n X_i^2$ with sufficiently many samples such that we have that $|\hat{\mu}_{X^2} - var(X)| < \epsilon/2$ with high probability due to Hoeffding's inequality (which applies due to the boundedness of X^2). Then consider the statistic $T(S) := |\hat{\mu}_{X^2} - 1|$, compare it to the threshold $\tau = \epsilon/2$ (with ϵ as given in the definition of $\mathcal{Q}$) and choose $\mathcal{P}$ if we are below the threshold and $\mathcal{Q}$ else. If $P \in \mathcal{P}$, by assumption $|\hat{\mu}_{X^2} - 1| < \epsilon/2$ with high probability and we successfully pick $\mathcal{P}$. If $P \in \mathcal{Q}$, then by assumption $|var(X) - 1| > \epsilon$ and furthermore $|\hat{\mu}_{X^2} - var(X)| < \epsilon/2$. Due to the triangle inequality we have that $|\hat{\mu}_{X^2} - 1| \geq |var(X) - 1| - |\hat{\mu}_{X^2} - var(X)| > \epsilon/2$ with high probability and we correctly select $\mathcal{Q}$. (*Remark: we did not expect the above derivation in the case* $P \in \mathcal{Q}$ *from you.*)

c) We additionally compute the empirical mean of X, i.e., $\hat{\mu}_X := \frac{1}{n}\sum_{i=1}^{n} X_i$.

We consider the statistic $T(S) := |\hat{\mu}_{X^2} - \hat{\mu}_X^2 - 1|$, compare it to the threshold $\tau = \epsilon/2$ (with ϵ as given in the definition of $\mathcal{Q}$) and choose $\mathcal{P}$ if we are below the threshold and $\mathcal{Q}$ else. The proof that this strategy works is analogous to the on presented in b).

Remark: the following is a rigorous justification as to why learnability of the first two moments implies learnability of the variance. Again, we did not expect this from you. Assume that we pick sufficiently many samples such that we $(\epsilon/4, \delta/2)$-learn the first two moments. By applying the triangle inequality and union bound we get that

$$\begin{aligned}
\mathbb{P}(\{|\hat{\mu}_{X^2} - \hat{\mu}_X^2 - var(X)| > \epsilon/2\}) &\leq \mathbb{P}(\{|\hat{\mu}_{X^2} - \mu_{X^2}| + |\hat{\mu}_X^2 - \mu_X^2| > \epsilon/2\}) && (1)\\
&\leq \mathbb{P}(\{|\hat{\mu}_{X^2} - \mu_{X^2}| > \epsilon/4\} \cup \{|\hat{\mu}_X^2 - \mu_X^2| > \epsilon/4\}) && (2)\\
&\leq \mathbb{P}((\{|\hat{\mu}_{X^2} - \mu_{X^2}| > \epsilon/4\}) + \mathbb{P}(\{|\hat{\mu}_X^2 - \mu_X^2| > \epsilon/4\}) && (3)\\
&\leq \delta/2 + \delta/2 = \delta && (4)
\end{aligned}$$

hence $\hat{\mu}_{X^2} - \hat{\mu}_X^2$ $(\epsilon/2, \delta)$-learns $var(X)$.

Problem 2: MMSE Estimation

Consider the scenario where $p(x|d) = de^{-dx}$, for $x \geq 0$ (and zero otherwise), that is, the observed data x is distributed according to an exponential with mean $1/d$. Moreover, the desired variable d itself is also exponentially distributed, with mean $1/\mu$.

(a) Find the MMSE estimator of d given x, and calculate the corresponding mean-squared error incurred by this estimator.

(b) Find the MAP estimator of d given x.

Solution 2. *(a)* Since d is exponentially distributed random variable with mean $1/\mu$, we have

$$p(d) = \mu e^{-\mu d} \tag{5}$$

Then the probability $p(x,d) = p(d)p(x|d) = \mu e^{-\mu d} de^{-dx} = \mu d e^{-(\mu+x)d}$. Thus, we have

$$p_X(x) = \int_d p(x,d) = \frac{\mu}{(\mu+x)^2} \tag{6}$$

Given x, the probability of $p(d|x) = (\mu+x)^2 de^{-(\mu+x)d}$, which is Gamma distribution $\Gamma(2, \mu+x)$.

The MMSE estimator of d given x, $\hat{d}_{MMSE}(x)$, satisfies

$$\hat{d}_{MMSE}(x) = \mathbb{E}[d|X=x] = \frac{2}{\mu+x} \tag{7}$$

and the mean-squared error is

$$
\begin{aligned}
\mathcal{E} &= \mathbb{E}_D[(d-\hat{d}_{MMSE})^2] && (8)\\
&= \mathbb{E}_X[\mathbb{E}_D[(d-\hat{d}_{MMSE}(x))^2|X=x]] && (9)\\
&= \int \mathbb{E}_D[(d-\hat{d}_{MMSE}(x))^2|X=x]p_X(x)dx && (10)\\
&= \int \mathbb{E}_D[(d^2-2d\hat{d}_{MMSE}(x)+\hat{d}^2_{MMSE}(x)|X=x]p_X(x)dx && (11)\\
&= \int (\mathbb{E}_D[d^2|X=x]-\hat{d}^2_{MMSE}(x))p_X(x)dx && (12)\\
&= \int \left(\frac{6}{(\mu+x)^2}-\frac{4}{(\mu+x)^2}\right)p_X(x)dx && (13)\\
&= \int \frac{2\mu}{(\mu+x)^4}dx && (14)\\
&= \frac{2}{3\mu^2} && (15)
\end{aligned}
$$

where $\mathbb{E}_D[d\hat{d}_{MMSE}(x)|X=x]=\hat{d}^2_{MMSE}(x)$ and $\mathbb{E}_D[d^2|X=x]=\mathrm{Var}(D|X)+\mathbb{E}[D|X]=\frac{6}{(\mu+x)^2}$ is because $p(d|x)$ is Gamma distribution.

(b) MAP estimator is

$$
\begin{aligned}
\hat{d}_{MAP}(x) &= \arg\max_d p(d|x) && (16)\\
&= \arg\max_d (\mu+x)^2 de^{-(\mu+x)d} && (17)\\
&= \arg\max_d de^{-(\mu+x)d} && (18)\\
&= \frac{1}{\mu+x} && (19)
\end{aligned}
$$

as $\frac{\partial}{\partial d}de^{-(\mu+x)d}=0$, when $d=\frac{1}{\mu+x}$.

Problem 3: Parameter Estimation and Fisher Information

The Fisher information $J(\Theta)$ for the family $f_\theta(x), \theta \in \mathbf{R}$ is defined by

$$J(\theta)=\mathbb{E}_\theta\left(\frac{\partial f_\theta(X)/\partial\theta}{f_\theta(X)}\right)^2=\int\frac{(f_\theta^{'})^2}{f_\theta}$$

Find the Fisher information for the following families:

(a) $f_\theta(x)=N(0,\theta)=\frac{1}{\sqrt{2\pi\theta}}e^{-\frac{x^2}{2\theta}}$

(b) $f_\theta(x)=\theta e^{-\theta x}, x\geq 0$

(c) What is the Cramèr Rao lower bound on $\mathbb{E}_\theta(\hat{\theta}(X)-\theta)^2$, where $\hat{\theta}(X)$ is an unbiased estimator of θ for (a) and (b)?

Solution 3. (a) Since $f_\theta(x)=N(0,\theta)=\frac{1}{\sqrt{2\pi\theta}}e^{-\frac{x^2}{2\theta}}$, we have

$$f_\theta'=-\frac{1}{2}\frac{1}{\sqrt{2\pi\theta^3}}e^{-\frac{x^2}{2\theta}}+\frac{x^2}{2\theta^2}\frac{1}{\sqrt{2\pi\theta}}e^{-\frac{x^2}{2\theta}} \qquad (20)$$

and

$$\frac{f'_\theta}{f_\theta} = \left(-\frac{1}{2\theta} + \frac{x^2}{2\theta^2}\right). \tag{21}$$

Therefore the Fisher information,

$$\begin{aligned} J(\theta) &= \mathbb{E}_\theta \left(\frac{f'_\theta}{f_\theta}\right)^2 &(22)\\ &= \mathbb{E}_\theta \left(\frac{1}{4\theta^2} - 2\frac{1}{2\theta}\frac{x^2}{2\theta^2} + \frac{x^4}{4\theta^4}\right) &(23)\\ &= \frac{1}{4\theta^2} - \frac{1}{\theta}\frac{\theta}{2\theta^2} + \frac{3\theta^2}{4\theta^4} &(24)\\ &= \frac{1}{2\theta^2}, &(25)\end{aligned}$$

where for $X \sim N(0,\theta)$, $\mathbb{E}[X^2] = \theta$ and $\mathbb{E}[X^4] = 3\theta^2$.

(b) Since $f_\theta(x) = \theta e^{-\theta x}, x \geq 0$, we have $\ln f_\theta = \ln\theta - \theta x$, and

$$\frac{f'_\theta}{f_\theta} = \frac{\partial \ln f_\theta}{\partial\theta} = \frac{1}{\theta} - x, \tag{26}$$

and therefore

$$\begin{aligned} J(\theta) &= \mathbb{E}_\theta \left(\frac{d\ln f_\theta}{d\theta}\right)^2 &(27)\\ &= \mathbb{E}_\theta \left(\frac{1}{\theta^2} - 2\frac{1}{\theta}x + x^2\right) &(28)\\ &= \frac{1}{\theta^2} - 2\frac{1}{\theta}\frac{1}{\theta} + \frac{2}{\theta^2} &(29)\\ &= \frac{1}{\theta^2} &(30)\end{aligned}$$

where for $X \sim \text{Exp}(\theta)$, $\mathbb{E}[X] = \frac{1}{\theta}$ and $\mathbb{E}[X^2] = \text{Var}(X) + \mathbb{E}[X]^2 = \frac{2}{\theta^2}$.

(c) The Cramèr-Rao lower bound is the reciprocal of the Fisher information, and is therefore $2\theta^2$ and θ^2 for parts (a) and (b) respectively.

Problem 4: Missing Data

We are given real-valued data with a single missing sample :

$$X_1, X_2, X_3, X_4, X_5, X_6, ?, X_8, X_9, \ldots \tag{31}$$

where we assume that the data is wide-sense stationary with autocorrelation function $R_X[k] = \alpha^{|k|}$, where $0 < \alpha < 1$. We would like to find a meaningful estimate for the missing sample X_7.

1. As a starting point, let us consider the estimate $\hat{X}_7 = wX_6$, where w is a real number. Find the value of w so as to minimize the mean-squared error $\mathbb{E}[(X_7 - \hat{X}_7)^2]$, and determine the incurred mean-squared error.
2. Now, consider the estimate $\hat{X}_7 = w_1X_6 + w_2X_8$. Again, find the values of w_1 and w_2 so as to minimize the mean-squared error $\mathbb{E}[(X_7 - \hat{X}_7)^2]$, and determine the incurred mean-squared error.

Solution 4. 1. $\hat{X}_7 = \alpha X_6$ and $\mathcal{E} = 1 - \alpha^2$. The corresponding Wiener filter has

$$R_x = \alpha^0 = 1 \qquad\qquad r_{dx} = \mathbb{E}[X_6 X_7] = \alpha$$

2. $\hat{X}_7 = \frac{\alpha}{1+\alpha^2}(X_6 + X_8)$ and $\mathcal{E} = \frac{1-\alpha^2}{1+\alpha^2}$ (better than Part 1). The corresponding Wiener filter has

$$R_x = \begin{bmatrix} 1 & \alpha^2 \\ \alpha^2 & 1 \end{bmatrix} \qquad\qquad r_{dx} = \begin{bmatrix} \alpha \\ \alpha \end{bmatrix}$$

Problem 5: Tweedie's Formula

For the special case where $X = D + N$, where N is Gaussian noise of mean zero and variance σ^2, *Tweedie's formula* says that the conditional mean (that is, the MMSE estimator) can be expressed as

$$\mathbb{E}\left[D \mid X = x\right] = x + \sigma^2 \ell'(x), \tag{32}$$

where

$$\ell'(x) = \frac{d}{dx} \log f_X(x), \tag{33}$$

where $f_X(x)$ denotes the marginal PDF of X. In this exercise, we derive this formula.

(a) Assume that $f_{X|D}(x|d) = e^{\alpha dx - \psi(d)} f_0(x)$ for some functions $\psi(d)$ and $f_0(x)$ and some constant α (such that $f_{X|D}(x|d)$ is a valid PDF for every value of d). Define

$$\lambda(x) = \log \frac{f_X(x)}{f_0(x)}, \tag{34}$$

where $f_X(x)$ is the marginal PDF of X, i.e., $f_X(x) = \int f_{X|D}(x|\delta) f_D(\delta) d\delta$. With this, establish that

$$\mathbb{E}\left[D \mid X = x\right] = \frac{1}{\alpha} \frac{d}{dx} \lambda(x). \tag{35}$$

(b) Show that the case where $X = D + N$, where N is Gaussian noise of mean zero and variance σ^2, is indeed of the form required in Part *(a)* by finding the corresponding $\psi(d), f_0(x)$, and α. Show that in this case, we have

$$\frac{f_0'(x)}{f_0(x)} = -\frac{x}{\sigma^2}, \tag{36}$$

and use this fact in combination with Part *(a)* to establish Tweedie's formula.

Solution 5. This formula is due to M. C. K. Tweedie, "Functions of a statistical variate with given means, with special reference to Laplacian distributions," *Proc. Camb. Phil. Soc.,* Vol. 43 (1947), pp.41-49.

(a) Simply plugging in, we find

$$\frac{d}{dx}\lambda(x) = \frac{d}{dx} \log \left(\frac{\int f_{X|D}(x|\delta) f_D(\delta) d\delta}{f_0(x)} \right) \tag{37}$$

$$= \frac{d}{dx} \log \left(\frac{\int e^{\alpha\delta x - \psi(\delta)} f_0(x) f_D(\delta) d\delta}{f_0(x)} \right) \tag{38}$$

$$= \frac{d}{dx} \log \int e^{\alpha\delta x - \psi(\delta)} f_D(\delta) d\delta \tag{39}$$

$$= \frac{1}{\int e^{\alpha\delta x - \psi(\delta)} f_D(\delta) d\delta} \int \alpha\delta e^{\alpha\delta x - \psi(\delta)} f_D(\delta) d\delta \tag{40}$$

But since we know that

$$\int e^{\alpha\delta x-\psi(\delta)} f_0(x) f_D(\delta) d\delta = f_X(x), \tag{41}$$

we can rewrite

$$\begin{aligned}
\frac{d}{dx}\lambda(x) &= \frac{f_0(x)}{f_X(x)} \int \alpha\delta e^{\alpha\delta x-\psi(\delta)} f_D(\delta) d\delta \qquad (42)\\
&= \alpha \int \delta \underbrace{\frac{e^{\alpha\delta x-\psi(\delta)} f_0(x) f_D(\delta)}{f_X(x)}}_{f_{D|X}(d|x)} d\delta \qquad (43)\\
&= \alpha\mathbb{E}\left[D|\, X = x\right] \qquad (44)
\end{aligned}$$

as claimed.

(b) In this case, we have

$$f_{X|D}(x|d) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} e^{-\frac{d^2}{2\sigma^2}} e^{\frac{1}{\sigma^2}xd}. \tag{45}$$

Pattern matching with the desired form

$$f_{X|D}(x|d) = e^{\alpha dx-\psi(d)} f_0(x), \tag{46}$$

it is quickly verified that $\alpha = 1/\sigma^2$, and

$$f_0(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}, \tag{47}$$

and thus,

$$f_0'(x) = -\frac{1}{\sqrt{2\pi}\sigma} \frac{2x}{2\sigma^2} e^{-\frac{x^2}{2\sigma^2}}, \tag{48}$$

giving the claimed result.

Putting things together, we have

$$\begin{aligned}
\mathbb{E}\left[D|\, X = x\right] &= \frac{1}{\alpha}\frac{d}{dx}\lambda(x) = \sigma^2 \left(\frac{d}{dx}\log f_X(x) - \frac{d}{dx}\log f_0(x)\right) \qquad (49)\\
&= \sigma^2 \left(\frac{d}{dx}\log f_X(x) - \frac{f_0'(x)}{f_0(x)}\right) \qquad (50)\\
&= \sigma^2 \left(\frac{d}{dx}\log f_X(x) + \frac{x}{\sigma^2}\right) \qquad (51)\\
&= x + \sigma^2 \frac{d}{dx}\log f_X(x), \qquad (52)
\end{aligned}$$

which is the claimed formula.