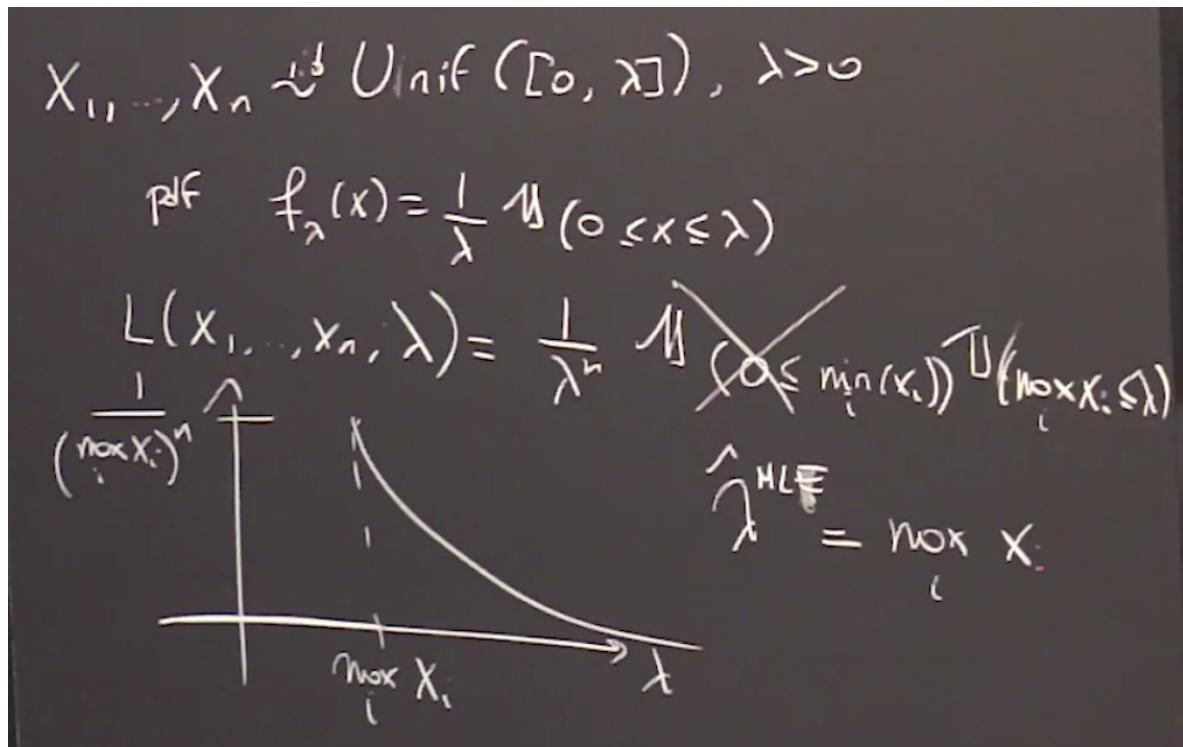


## Unit 3 cont'd Methods of Estimation

once lambda is less than the maximum of  $x_i$  then it is 0



this model does not satisfy the regulatory conditions - i.e. the derivative is not defined throughout so we can't set it to 0 to find the MLE  
(can't take derivative below  $\max x_i$ )

## Concept Check: Maximum Likelihood Estimator for a Uniform Statistical Model

1/1 point (graded)

Let  $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Unif}[0, \theta^*]$  where  $\theta^*$  is an unknown parameter. We constructed the associated statistical model  $(\mathbb{R}_{\geq 0}, \{\text{Unif}[0, \theta]\}_{\theta > 0})$  (where  $\mathbb{R}_{\geq 0}$  denotes the nonnegative reals).

For any  $\theta > 0$ , the density of  $\text{Unif}[0, \theta]$  is given by  $f(x) = \frac{1}{\theta} \mathbf{1}(x \in [0, \theta])$ . Recall that

$$\mathbf{1}(x \in [0, \theta]) = \begin{cases} 1 & \text{if } x \in [0, \theta] \\ 0 & \text{otherwise.} \end{cases}$$

Hence we can use the product formula and compute the likelihood to be

$$L_n(x_1, \dots, x_n, \theta) = \prod_{i=1}^n \left( \frac{1}{\theta} \mathbf{1}(x_i \in [0, \theta]) \right) = \frac{1}{\theta^n} \mathbf{1}(x_i \in [0, \theta] \ \forall 1 \leq i \leq n).$$

For the fixed values  $(1, 3, 2, 2.5, 5, 0.1)$  (think of these as observations of random variables  $X_1, \dots, X_6$ ), what value of  $\theta$  maximizes  $L_6(1, 3, 2, 2.5, 5, 0.1, \theta)$ ?

5

✓ Answer: 5

### Solution:

Observe that

$$L_6(1, 3, 2, 2.5, 5, 0.1, \theta) = \frac{1}{\theta^6} \mathbf{1}(\{1, 3, 2, 2.5, 5, 0.1\} \subset [0, \theta]).$$

If  $\theta < \max\{1, 3, 2, 2.5, 5, 0.1\}$ , then we have  $\{1, 3, 2, 2.5, 5, 0.1\} \not\subset [0, \theta]$ . By the definition of the indicator function, this means  $L_6(1, 3, 2, 2.5, 5, 0.1, \theta) = 0$  for  $\theta < \max\{1, 3, 2, 2.5, 5, 0.1\} = 5$ . Hence, when maximizing  $L_6(1, 3, 2, 2.5, 5, 0.1, \theta)$ , we need to consider  $\theta \in [5, \infty)$ . Restricted to this interval, we observe that

$$L_6(1, 3, 2, 2.5, 5, 0.1, \theta) = \frac{1}{\theta^6}.$$

The above is a decreasing function on  $[5, \infty)$ , so the maximum is attained when  $\theta = \max\{1, 3, 2, 2.5, 5, 0.1\} = 5$ .

**Remark:** In general, the maximum likelihood estimator for  $\theta^*$  in this uniform statistical model is

$$\widehat{\theta}_n^{MLE} = \max_{1 \leq i \leq n} X_i.$$

## MLE for a Loaded Die: Likelihood

1/1 point (graded)

You have a loaded (i.e. possibly unfair) six-sided die with the probability that it shows a "3" equal to  $\eta$  and the probability that it shows any other number equal to  $(1 - \eta)/5$ .

Let  $X$  be a random variable representing a roll of this die. You roll this die  $n$  times, and record your data set, consisting of the values of the faces as  $X_1, X_2, X_3, \dots, X_n$ .

Let the outcome of a set of  $n$  rolls of the die be modeled by the i.i.d. random variable sequence  $(X_1, \dots, X_n)$ . We model the  $i$ 'th roll as  $X_i$  where  $X_i = j$  if the top face of the die shows a " $j$ ".

You roll the die  $n$  times and observe a sequence of outcomes  $x_1, \dots, x_n$  which contains exactly  $k$  outcomes  $x_i = 3$ . What is the likelihood function  $L_n(x_1, \dots, x_n, \eta)$  for the entire sequence of outcomes?

(Enter **eta** for  $\eta$ .)

eta<sup>k</sup>\*((1-eta)/5)<sup>(n-k)</sup>

✓ Answer: eta<sup>k</sup>((1-eta)/5)<sup>(n-k)</sup>

$$\eta^k \cdot \left(\frac{1-\eta}{5}\right)^{n-k}$$

### Solution:

Denote by  $p_\eta(x)$  the pmf of  $X_i$ . The probability that  $X_i$  takes on a value 3 is equal to  $\eta$  and the probability that  $X_i$  takes on any other value is  $(1 - \eta)/5$ . Therefore, using the probability of observing a particular sequence of outcomes, we obtain the likelihood function as

$$L_n(x_1, \dots, x_n, \eta) = \prod_{i=1}^n p_\eta(x_i)$$
$$= \eta^k \left(\frac{1-\eta}{5}\right)^{n-k}.$$

Note that the above does not contain a combinatorial expression as we are interested in the probability of observing a particular sequence of die rolls and not in the probability of obtaining a certain number of 3's in  $n$  rolls.

## MLE for a Loaded Die: MLE

1/1 point (graded)

Find the ML estimator  $\hat{\eta}_n^{\text{MLE}}$ .

k/n

✓ Answer: k/n

$$\frac{k}{n}$$

STANDARD NOTATION

### Solution:

Since we are looking for the  $\operatorname{argmax}_{\eta \in [0,1]} L_n(x_1, \dots, x_n, \eta)$ , we can ignore any scaling constant in  $L_n(x_1, \dots, x_n, \eta)$ . Hence, we will maximize  $\tilde{L}_n(x_1, \dots, x_n, \eta) = \eta^k(1 - \eta)^{n-k}$ .

Taking the derivative of  $\tilde{L}_n(x_1, \dots, x_n, \eta)$  with respect to  $\eta$  and setting it to 0, we get

$$k(1 - \eta) = (n - k)\eta$$
$$\implies \hat{\eta}_n^{\text{MLE}} = \frac{k}{n}.$$

**Remark:** The function  $\tilde{L}_n(x_1, \dots, x_n, \eta) = \eta^k(1 - \eta)^{n-k}$  whose maximizer is  $\hat{\eta}_n^{\text{MLE}}$  is the same as the likelihood function for a Bernoulli experiment with parameter  $\eta$ , even though each roll of a die has 6 potential outcomes.

## Review: Definition of MLE

1/1 point (graded)

Let  $\{E, (\mathbf{P}_\theta)_{\theta \in \Theta}\}$  be a statistical model associated with a sample of i.i.d. random variables  $X_1, X_2, \dots, X_n$ . Assume that there exists  $\theta^* \in \Theta$  such that  $X_i \sim \mathbf{P}_{\theta^*}$ .

Recall that the **Kullback-Leibler (KL) divergence** between two distributions  $\mathbf{P}_{\theta^*}$  and  $\mathbf{P}_\theta$ , with pdfs  $p_{\theta^*}$  and  $p_\theta$  respectively, is defined as

$$\text{KL}(\mathbf{P}_{\theta^*}, \mathbf{P}_\theta) = \mathbb{E}_{\theta^*} \left[ \ln \left( \frac{p_{\theta^*}(X)}{p_\theta(X)} \right) \right],$$

and a consistent, up to a constant, estimator of  $\theta \mapsto \text{KL}(\mathbf{P}_{\theta^*}, \mathbf{P}_\theta)$  is

$$\widehat{\text{KL}}_n(\mathbf{P}_{\theta^*}, \mathbf{P}_\theta) - \text{constant} = -\frac{1}{n} \sum_{i=1}^n \ln p_\theta(X_i).$$

Which of the following represents the maximum likelihood estimator of  $\theta^*$ ? (Choose all that apply).

$\text{argmin}_{\theta \in \Theta} \widehat{\text{KL}}_n(\mathbf{P}_{\theta^*}, \mathbf{P}_\theta)$

$\text{argmax}_{\theta \in \Theta} \sum_{i=1}^n \ln p_\theta(X_i)$

$\text{argmax}_{\theta \in \Theta} \ln \left( \prod_{i=1}^n p_\theta(X_i) \right)$

$\text{argmax}_{\theta \in \Theta} \ln(L_n(X_1, X_2, \dots, X_n; \theta))$

✓

### Solution:

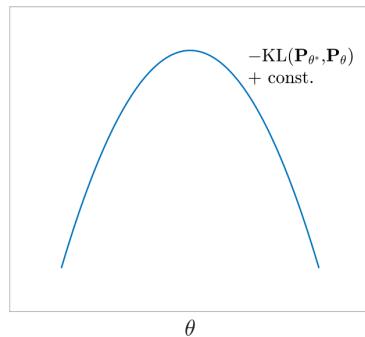
Recall the **maximum likelihood estimator** can be defined as the

$$\hat{\theta}_n^{MLE} = \text{argmin}_{\theta \in \Theta} \widehat{\text{KL}}_n(\mathbf{P}_{\theta^*}, \mathbf{P}_\theta).$$

In other words, the maximum likelihood estimator is the (unique)  $\theta$  that minimizes  $\widehat{\text{KL}}(\mathbf{P}_{\theta^*}, \mathbf{P}_\theta)$  over the parameter space  $\theta \in \Theta$ . (The minimizer of the KL divergence is unique due to it being strictly convex in the space of distributions once  $\mathbf{P}_{\theta^*}$  is fixed.) All choices are equivalent to this definition:

$$\begin{aligned} \hat{\theta}_n^{MLE} &= \text{argmin}_{\theta \in \Theta} \widehat{\text{KL}}_n(\mathbf{P}_{\theta^*}, \mathbf{P}_\theta) = \text{argmin}_{\theta \in \Theta} \left( \text{Constant} - \frac{1}{n} \sum_{i=1}^n \ln p_\theta(X_i) \right) \\ &= \text{argmax}_{\theta \in \Theta} \left( \frac{1}{n} \sum_{i=1}^n \ln p_\theta(X_i) \right) \quad (\text{drop additive constant and negative sign}) \\ &= \text{argmax}_{\theta \in \Theta} \left( \sum_{i=1}^n \ln p_\theta(X_i) \right) \quad (\text{drop positive scaling factor}) \\ &= \text{argmax}_{\theta \in \Theta} \left( \ln \left( \prod_{i=1}^n p_\theta(X_i) \right) \right) \quad (\text{log property}) \\ &= \text{argmax}_{\theta \in \Theta} \ln(L_n(X_1, X_2, \dots, X_n; \theta)) \quad (\text{definition of likelihood}). \end{aligned}$$

**Note:** In the following video, at around the 3:20 mark, the plot of  $-\text{KL}(\mathbf{P}_{\theta^*}, \mathbf{P}_\theta)$ , with  $\theta^*$  fixed and as a function of  $\theta$ , is presented incorrectly as a convex curve while it should be concave. This error propagates until the end of the video and we request you to keep the following picture in mind instead:



the 2 dotted plot should shift down and converge to theta as the KL divergence reduces

also in that respect the x-axis difference should also be converging ( $\theta^*$  to  $\theta$ )



### Consistency of MLE

Given i.i.d samples  $X_1, \dots, X_n \sim \mathbf{P}_{\theta^*}$  and an associated statistical model  $(E, \{\mathbf{P}_\theta\}_{\theta \in \Theta})$ , the maximum likelihood estimator  $\hat{\theta}_n^{\text{MLE}}$  of  $\theta^*$  is a **consistent** estimator under mild regularity conditions (e.g. continuity in  $\theta$  of the pdf  $p_\theta$  almost everywhere), i.e.

$$\hat{\theta}_n^{\text{MLE}} \xrightarrow[p]{n \rightarrow \infty} \theta^*.$$

Note that this is true even if the parameter  $\theta$  is a vector in a higher dimensional parameter space  $\Theta$ , and  $\hat{\theta}_n^{\text{MLE}}$  is a multivariate random variable, e.g. if  $\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} \in \mathbb{R}^2$  for a Gaussian statistical model.

### Multivariate Random Variables

A **multivariate random variable**, or a **random vector**, is a vector-valued function whose components are (scalar) random variables on the same underlying probability space. More specifically, a random vector  $\mathbf{X} = (X^{(1)}, \dots, X^{(d)})^T$  of dimension  $d \times 1$  is a vector-valued function from a probability space  $\Omega$  to  $\mathbb{R}^d$ :

$$\mathbf{X} : \Omega \rightarrow \mathbb{R}^d$$

$$\omega \mapsto \begin{pmatrix} X^{(1)}(\omega) \\ X^{(2)}(\omega) \\ \vdots \\ X^{(d)}(\omega) \end{pmatrix}$$

where each  $X^{(k)}$  is a (scalar) random variable on  $\Omega$ . We will often (but not always) use the bracketed superscript  $(k)$  to denote the  $k$ -th component of a random vector, especially when the subscript is already used to index the samples.

The **probability distribution** of a random vector  $\mathbf{X}$  is the **joint distribution** of its components  $X^{(1)}, \dots, X^{(d)}$ .

The **cumulative distribution function (cdf)** of a random vector  $\mathbf{X}$  is defined as

$$F : \mathbb{R}^d \rightarrow [0, 1]$$

$$\mathbf{x} \mapsto \mathbf{P}(X^{(1)} \leq x^{(1)}, \dots, X^{(d)} \leq x^{(d)}).$$

### Convergence in Probability in Higher Dimension

To make sense of the consistency statement  $\hat{\theta}_n^{\text{MLE}} \xrightarrow[p]{n \rightarrow \infty} \theta^*$  where the MLE  $\hat{\theta}_n^{\text{MLE}}$  is a random vector, we need to know what convergence in probability means in higher dimensions. But this is no more than convergence in probability for **each component**.

Let  $\mathbf{X}_1, \mathbf{X}_2, \dots$  be a sequence of random vectors of size  $d \times 1$ , i.e.  $\mathbf{X}_i = \begin{pmatrix} X_i^{(1)} \\ \vdots \\ X_i^{(d)} \end{pmatrix}$ .

Let  $\mathbf{X} = \begin{pmatrix} X^{(1)} \\ \vdots \\ X^{(d)} \end{pmatrix}$  be another vector of size  $d \times 1$ .

Then

$$\mathbf{X}_n \xrightarrow[p]{n \rightarrow \infty} \mathbf{X} \iff X_n^{(k)} \xrightarrow[p]{n \rightarrow \infty} X^{(k)} \text{ for all } 1 \leq k \leq d.$$

In other words, the sequence  $\mathbf{X}_1, \mathbf{X}_2, \dots$  **converges in probability** to  $\mathbf{X}$  if and only if each component sequence  $X_1^{(k)}, X_2^{(k)}, \dots$  converges in probability to  $X^{(k)}$ .

Hence, for example, in the Gaussian model  $((-\infty, \infty), \{\mathcal{N}(\mu, \sigma^2)\}_{(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0}})$ , consistency of the MLE  $\hat{\theta}_n^{\text{MLE}} = \left( \begin{array}{c} \widehat{\mu} \\ \widehat{\sigma^2} \end{array} \right)$  means that  $\widehat{\mu}$  and  $\widehat{\sigma^2}$  are consistent estimators of  $\mu^*$  and  $(\sigma^2)^*$ , respectively.

**Remark:** You can check that this condition is equivalent to the following definition of convergence in probability, which is a straightforward generalization of the 1-dimensional case:

$$P(\{\omega \in \Omega : |\mathbf{X}_n(\omega) - \mathbf{X}(\omega)| < \epsilon\}) \xrightarrow{n \rightarrow \infty} 1 \quad \text{for any } \epsilon > 0.$$

### Consistency of the MLE of a Uniform Model

0/1 point (graded)

Let  $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Unif}[0, \theta^*]$  where  $\theta^*$  is an unknown parameter. We construct the associated statistical model  $(\mathbb{R}_{\geq 0}, \{\text{Unif}[0, \theta]\}_{\theta > 0})$

Consider the maximum likelihood estimator  $\hat{\theta}_n^{\text{MLE}} = \max_{i=1, \dots, n} X_i$ .

Which of the following are true about  $\hat{\theta}_n^{\text{MLE}}$ . (Choose all that apply.)

$\max_{i=1, \dots, n} X_i$  is a consistent estimator ✓

For any  $0 < \epsilon \leq \theta^*$ ,  $P\left(|\max_{i=1, \dots, n} X_i - \theta^*| \geq \epsilon\right) \rightarrow 0$  as  $n \rightarrow \infty$  ✓

For any  $0 < \epsilon \leq \theta^*$ ,  $P\left(|\max_{i=1, \dots, n} X_i - \theta^*| \geq \epsilon\right) \rightarrow c$  as  $n \rightarrow \infty$ , where  $c > 0$  is a constant

For any  $0 < \epsilon \leq \theta^*$ ,  $P\left(|\max_{i=1, \dots, n} X_i - \theta^*| \geq \epsilon\right) = \left(\frac{\theta^* - \epsilon}{\theta^*}\right)^n$  ✓

✗

#### Solution:

Choices 1, 2, and 4 are true because of the following proof for consistency of this ML estimator. Let  $0 < \epsilon \leq \theta^*$ :

$$\begin{aligned} P\left(\left|\max_{i=1, \dots, n} X_i - \theta^*\right| \geq \epsilon\right) &= P\left(\theta^* - \max_{i=1, \dots, n} X_i \geq \epsilon\right) \\ &= P\left(\max_{i=1, \dots, n} X_i \leq \theta^* - \epsilon\right) \\ &= \left(\frac{\theta^* - \epsilon}{\theta^*}\right)^n \rightarrow 0 \text{ as } n \rightarrow \infty. \end{aligned}$$

Choice 3 is not true because if a sequence (the relevant sequence here is  $\left(\frac{\theta^* - \epsilon}{\theta^*}\right)^n$ ) converges to a limit, then the limit is unique.

## Review: Covariance

1/2 points (graded)

If  $X$  and  $Y$  are random variables with respective means  $\mu_X$  and  $\mu_Y$ , then recall the **covariance** of  $X$  and  $Y$  (written  $\text{Cov}(X, Y)$ ) is defined to be

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)].$$

Alternatively, one can show that this is equivalent to  $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ .

For each of the following statements, indicate whether it is true or false.

" $\text{Cov}(X, X) = \text{Var}(X)$ ".

True

False



"Like the variance, the covariance between an arbitrary pair of RVs  $X$  and  $Y$  is always non-negative."

True

False ✓



### Solution:

- **True.**  $\text{Cov}(X, X) = \mathbb{E}[(X - \mu_X)^2] = \text{Var}(X)$ .
- **False.** Consider  $(X, Y)$  which is distributed uniformly over the set  $\{(1, -1), (-1, 1)\}$ . The marginal distributions of both  $X$  and  $Y$  are uniform over  $\{\pm 1\}$ , so  $\mu_X = \mu_Y = 0$ . On the other hand,  $\mathbb{E}[XY] = -1$ , so  $\text{Cov}(X, Y) = -1$ .

## Alternate Formula for Covariance

1/1 point (graded)

Let  $X$  and  $Y$  are random variables with respective means  $\mu_X$  and  $\mu_Y$ . Is it true that  $\mathbb{E}[(X)(Y - \mu_Y)] = \text{Cov}(X, Y)$ ?

True

False



### Solution:

Indeed,  $\mathbb{E}[(X)(Y - \mu_Y)] = \text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y)]$ . That is, it is sufficient to center one random variable around its mean when computing the covariance between two random variables. This can be seen from the following:

$$\begin{aligned}\mathbb{E}[(X)(Y - \mu_Y)] &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].\end{aligned}$$

## Bilinearity of Covariance

1/1 point (graded)

Let  $X, Y, Z$  be random variables and  $a, b$  be constants. Indicate whether the following statement is true or false.

"Covariance is bilinear, i.e.  $\text{Cov}(aX + bY, Z) = a\text{Cov}(X, Z) + b\text{Cov}(Y, Z)$ ."

*Hint:* Use the result from the problem immediately above.

True

False



### Solution:

**True.** This can be seen by using the trick of centering only  $Z$  when computing covariance. First, note that  $\text{Cov}(aX, Z) = \mathbb{E}[(aX)(Z - \mu_Z)] = a\mathbb{E}[(X)(Z - \mu_Z)] = a\text{Cov}(X, Z)$ . Then,

$$\begin{aligned}\text{Cov}(aX + bY, Z) &= \mathbb{E}[(aX + bY)(Z - \mu_Z)] \\ &= \mathbb{E}[(aX)(Z)] - \mu_Z \mathbb{E}[aX] + \mathbb{E}[(bY)(Z)] - \mu_Z \mathbb{E}[bY] \\ &= a\text{Cov}(X, Z) + b\text{Cov}(Y, Z)\end{aligned}$$

## Example of Covariance I

0/1 point (graded)

Let  $A = X + Y$  and  $B = X - Y$ . Let  $\mu_X = \mathbb{E}[X]$ ,  $\mu_Y = \mathbb{E}[Y]$ ,  $\tau_X = \text{Var}(X)$ ,  $\tau_Y = \text{Var}(Y)$  and  $c = \text{Cov}(X, Y)$ . In terms of  $\mu_X$ ,  $\mu_Y$ ,  $\tau_X$ ,  $\tau_Y$ , and  $c$ , what is  $\text{Cov}(A, B)$ ?

(Enter **mu\_X** for  $\mu_X$ , **tau\_X** for  $\tau_X$ .)

c - (tau\_X + mu\_X^2 - ta

✖ Answer: tau\_X-tau\_Y

$c - (\tau_X + \mu_X^2 - \tau_Y - \mu_Y^2)$

**STANDARD NOTATION**

### Solution:

Expand out the definition of covariance using bi-linearity (see the solution to the previous question):

$$\begin{aligned}\text{Cov}(A, B) &= \text{Cov}(X + Y, X - Y) \\ &= \text{Cov}(X + Y, X) - \text{Cov}(X + Y, Y) \\ &= \text{Cov}(X, X) + \text{Cov}(Y, X) - \text{Cov}(X, Y) - \text{Cov}(Y, Y) \\ &= \text{Var}(X) - \text{Var}(Y) \\ &= \tau_X - \tau_Y.\end{aligned}$$

## Review: Independence

2/3 points (graded)

Let  $X$  be a random variable that takes on values 0 and 1 with **equal probability**. You decide to communicate this random variable to your friend through a medium that "adds" (defined below in individual sub-problems) a random noise  $Y$  that takes on values 0 and 1 with probabilities  $\alpha_0$  and  $\alpha_1$ , respectively. Let  $Y$  be **independent** of  $X$  and let  $Z$  denote the result of this "addition", which is what your friend receives.

**Note:** This  $Z = X + Y$  model for random variables can be viewed as a problem on exploring functions of random variables. However, the problem is practically motivated by many real-world examples. For example, in a communication system the  $X$  is usually what a sender sends over the "medium" and  $Z$  is what the receiver receives. The "medium", which is also called channel, could range anything from a wired line to a wireless link connecting a sender and a receiver. The noise  $Y$  added by the medium is independent of what the sender sends over the channel.

In each of the following scenarios, indicate whether the statement in quotes is true or false.

**Scenario 1:** You treat your random variable as a binary digit (bit) and the medium adds (XORs) noise that takes on binary values 0 and 1 with probabilities  $\alpha_0 = \frac{1}{2}$  and  $\alpha_1 = \frac{1}{2}$ , respectively. That is  $Z = X \text{ XOR } Y$ . "In this scenario,  $X$  and  $Z$  are independent random variables."

**Note:** The XOR of two bits is a function that takes in two bits and produces an output of 1 if the two input bits are different and produces an output of 0 otherwise.

True

False

✓

**Scenario 2:** You treat your random variable as a binary digit (bit) and the medium adds (XORs) noise that takes on binary value 0 and 1 with probabilities  $\alpha_0 = \frac{1}{3}$  and  $\alpha_1 = \frac{2}{3}$ , respectively. That is  $Z = X \text{ XOR } Y$ . "In this scenario,  $X$  and  $Z$  are independent random variables."

True

False ✓

✗

**Scenario 3:** You treat your random variable as an integer and the medium adds (integer addition) noise that takes on integer values 0 and 1 with probabilities  $\alpha_0 = \frac{1}{2}$  and  $\alpha_1 = \frac{1}{2}$ , respectively. That is  $Z = X + Y$ . "In this scenario,  $X$  and  $Z$  are independent random variables."

True

False

✓

In the first two scenarios,  $Z$  takes on binary values 0 and 1 whose probabilities can be computed as follows (recall that  $X$  and  $Y$  are independent):

$$\begin{aligned} p_Z(z=0) &= p_X(x=0) \cdot p_Y(y=0) + p_X(x=1) \cdot p_Y(y=1) \\ &= \frac{1}{2}\alpha_0 + \frac{1}{2}\alpha_1 \\ &= \frac{1}{2}, \\ p_Z(z=1) &= 1 - p_Z(z=0) = \frac{1}{2} \end{aligned}$$

Furthermore, in the first two scenarios the joint pmf can be computed as

$$p_{X,Z}(x,z) = p_X(x)p_Z(z|x) = \begin{cases} \frac{1}{2}p_{Y|X}(y=0|x) = \frac{1}{2}\alpha_0 & \text{if } x=z \\ \frac{1}{2}p_{Y|X}(y=1|x) = \frac{1}{2}\alpha_1 & \text{if } x \neq z. \end{cases}$$

- Scenario 1:** For scenario 1, the above joint pmf resolves to a value of  $\frac{1}{4}$  for all values of  $x$  and  $z$ . Hence,  $X$  and  $Z$  are independent in scenario 1.
- Scenario 2:** For scenario 2,  $X$  and  $Z$  are not independent. For example, when  $x=0, z=0$ ,  $p_{X,Z}(x,z) = \frac{1}{2}\alpha_0 = \frac{1}{6}$ , which is not equal to  $p_X(x=0)p_Z(z=0) = \frac{1}{4}$ .

**Scenario 3:** In this case,  $X$  takes on integer values 0 and 1 and  $Z$  takes on integer values 0, 1, 2.  $X$  and  $Z$  are not independent. This can be seen by computing, for example,  $p_{X,Z}(x=0, z=0)$  and  $p_X(x=0)p_Z(z=0)$ :

$$\begin{aligned} p_{X,Z}(x=0, z=0) &= p_X(x=0)p_{Z|X}(z=0|x=0) \\ &= \frac{1}{2}p_Y(y=0) = \frac{1}{4} \\ p_X(x=0)p_Z(z=0) &= \frac{1}{2} \left[ p_{Z|X}(z=0|x=0) \frac{1}{2} + p_{Z|X}(z=0|x=1) \frac{1}{2} \right] \\ &= \frac{1}{2} \left[ \frac{1}{4} + 0 \right] = \frac{1}{8} \end{aligned}$$

## Independence, Estimation

1/1 point (graded)

Problem setup as above.

Your friend decides to perform ML estimation of what you might have transmitted based on what they received.

Concretely, your friend wishes to obtain an ML estimate  $\hat{x}$  upon observing  $z = x$  add  $y$ . They obtain a likelihood function  $L(z, x)$ , where  $z$  is the observed realization of random variable  $Z$  and  $x$  can be treated as the unknown parameter. Assume that your friend knows that  $X$  is equally likely to take on either 0 or 1 (in all scenarios).

Under which scenario(s) would your friend's ML estimate  $\hat{X}$  be equal to  $X$  with a probability more than 0.5?

**Hint:** Think heuristically, assuming you obtained the correct answers for the previous problem.

Scenario 1

Scenario 2

Scenario 3



**Solution:**

**Scenario 1:** In scenario 1,  $Z$  is independent of  $X$ . This implies that even though  $Z$  is a function of what you send over the medium, your friend would not do better with an ML estimator than guessing randomly what the input  $X$  could have been. This can be seen from the following likelihood function:

$$L(z, x) = p_{Z|X}(z | x) = p_Z(z).$$

Upon observing  $Z$ , the above expression is maximized to have a value of  $p_Z(z)$  with either  $\hat{x} = 0$  or  $\hat{x} = 1$  both being optimal irrespective of the received  $z$ .

**Scenarios 2 and 3:** In scenario 2 (and 3),  $Z$  is not independent of  $X$ . This implies that the dependence should intuitively make an ML estimator perform better than randomly guessing what might have been transmitted. We show that the ML estimator  $\hat{X}$  is equal to  $X$  with a probability more than 0.5 in the following for scenario 2, and the proof for scenario 3 is obtained in a similar fashion.

The likelihood function is

$$L(z, x) = p_{Z|X}(z | x) = \frac{p_{X|Z}(x | z) p_Z(z)}{p_X(x)},$$

where we can ignore  $p_X(x)$  as it is equal to  $\frac{1}{2}$  for any  $x$  and also ignore  $p_Z(z)$  because it does not depend upon parameter  $x$ . Therefore, we need to perform the following optimization:

$$\max_{x \in \{0,1\}} p_{X|Z}(x | z)$$

upon observing  $z$ . If  $z = 0$  is observed, then  $\hat{x} = 1$  maximizes the above expression with a value of  $\frac{2}{3}$ . Similarly, if  $z = 1$  is observed,  $\hat{x} = 0$  maximizes the above expression with a value of  $\frac{2}{3}$ .

Therefore, the ML estimator for scenario 2 is:

$$\hat{x} = \begin{cases} 1 & \text{if } z = 0 \\ 0 & \text{if } z = 1. \end{cases}$$

The probability of error  $P[\hat{X} \neq X]$  for this estimator is  $\frac{1}{3}$ .

## Covariance and Independence

2/2 points (graded)

For each of the following statements, indicate whether it is true or false.

" $X, Y$  are independent  $\implies \text{Cov}(X, Y) = 0$ ."

*Hint:  $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$  whenever  $X, Y$  are independent.)*

True

False



" $\text{Cov}(X, Y) = 0 \implies X, Y$  are independent."

*Hint: If this were false, there should be an easy counterexample. Is there an easy example where  $\mathbb{E}[XY] = 0$  and  $\mathbb{E}[Y] = 0$  but  $X, Y$  are not independent?*

True

False



**Solution:**

- **True.** If  $X, Y$  are independent,  $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] = 0$ .
- **False.** Consider  $X$  which is Bernoulli(1/2). Let  $Y$  be a random variable which is always 0 if  $X = 0$ , and uniformly distributed over  $\{\pm 1\}$  if  $X = 1$ . Notice that  $\mathbb{E}[Y] = \frac{1}{2}0 + \frac{1}{4}1 + \frac{1}{4}(-1) = 0$ . On the other hand,  $\mathbb{E}[XY] = (0 \cdot 0) \cdot \frac{1}{2} + (1 \cdot 1) \frac{1}{4} + (1 \cdot -1) \frac{1}{4} = 0$ . However,  $X$  and  $Y$  are not independent.

slide 30, why  $X+Y$  is  $2X$  with  $P(1/2)$  and 0 with  $P(1/2)$

A handwritten note on a dark background. It shows two equations:  $Y = R \cdot X$  and  $X + Y = (I + R)X$ . There is a large checkmark above the first equation.

## Sample Covariance

1/4 points (graded)

Let  $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n) \stackrel{\text{iid}}{\sim} (X, Y)$  with  $\mathbb{E}[X] = \mu_X$ ,  $\mathbb{E}[Y] = \mu_Y$ , and  $\mathbb{E}[XY] = \mu_{XY}$ . That is, each random variable pair  $(X_i, Y_i)$  has the same distribution as the random variable pair  $(X, Y)$ , and the pairs are independent of one another.

Estimating the covariance of  $X$  and  $Y$  is a useful exercise because non-zero covariance implies statistical dependence of  $X$  and  $Y$ . In this problem, we study one way to obtain an unbiased estimator for  $\text{Cov}(X, Y)$ .

Consider the following estimator of the covariance:

$$\widetilde{S}_{XY} = \frac{1}{n} \left( \sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n) \right),$$

where  $\bar{X}_n$  and  $\bar{Y}_n$  denote the sample means estimators of  $\mu_X$  and  $\mu_Y$ .

What is  $\mathbb{E}\left[\frac{(\sum_{i=1}^n X_i)(\sum_{j=1}^n Y_j)}{n}\right]$ ? Provide an expression in terms of  $n$ ,  $\mu_X$ ,  $\mu_Y$ , and  $\mu_{XY}$ .

(Enter  $\text{mu\_}\{XY\}$  for  $\mu_{XY}$ ,  $\text{mu\_}X$  for  $\mu_X$ , and  $\text{mu\_}Y$  for  $\mu_Y$ .)

$$\mathbb{E}\left[\frac{(\sum_{i=1}^n X_i)(\sum_{j=1}^n Y_j)}{n}\right] = \boxed{(\text{mu\_}\{XY\}-\text{mu\_}X*\text{mu\_}Y)/n} \times$$

Answer:  $(1/n)*(n*\text{mu\_}XY + n*(n-1)*\text{mu\_}X*\text{mu\_}Y)$

$$\frac{\mu_{XY} - \mu_X \cdot \mu_Y}{n}$$

Is  $\widetilde{S}_{XY}$  an unbiased estimator of  $\text{Cov}(X, Y)$ ?

Yes

No



If your answer to the above question is "Yes", then type "1" in the following box. Otherwise, find a scaling factor  $c$  such that

$$\widehat{S}_{XY} = c \cdot \widetilde{S}_{XY}$$

is an unbiased estimator of  $\text{Cov}(X, Y)$ . Provide your answer in terms of  $n$ ,  $\mu_X$ ,  $\mu_Y$ , and  $\mu_{XY}$ .

(Enter  $\text{mu\_}\{XY\}$  for  $\mu_{XY}$ ,  $\text{mu\_}X$  for  $\mu_X$ , and  $\text{mu\_}Y$  for  $\mu_Y$ .)

$$c = \boxed{n} \quad \times \text{ Answer: } (n/(n-1))$$

n

First,

$$\begin{aligned}\mathbb{E} \left[ \frac{(\sum_{i=1}^n X_i)(\sum_{j=1}^n Y_j)}{n} \right] &= \frac{1}{n} \mathbb{E} \left[ \sum_{i=1}^n X_i Y_i + \sum_{i=1}^n \sum_{j=1, j \neq i}^n X_i Y_j \right] \\ &= [\mu_{XY} + (n-1)\mu_X\mu_Y],\end{aligned}$$

where we have used the property that  $X_i$  and  $Y_j$  are independent whenever  $i \neq j$ . Then,

$$\begin{aligned}\mathbb{E}[\widetilde{S}_{XY}] &= \frac{1}{n} \mathbb{E} \left[ \sum_{i=1}^n (X_i - \bar{X}_n) (Y_i - \bar{Y}_n) \right] \\ &= \frac{1}{n} \mathbb{E} \left[ \sum_{i=1}^n X_i Y_i - \frac{\sum_{i=1}^n X_i}{n} \sum_{j=1}^n Y_j - \frac{\sum_{i=1}^n Y_i}{n} \sum_{j=1}^n X_j + \frac{\sum_{i=1}^n X_i \sum_{j=1}^n Y_j}{n} \right] \\ &= \frac{1}{n} \mathbb{E} \left[ \sum_{i=1}^n X_i Y_i - \frac{\sum_{i=1}^n X_i \sum_{j=1}^n Y_j}{n} \right].\end{aligned}$$

Using the result in the first part of the problem, we get

$$\begin{aligned}\mathbb{E}[\widetilde{S}_{XY}] &= \frac{1}{n} [n\mu_{XY} - (\mu_{XY} + (n-1)\mu_X\mu_Y)] \\ &= \frac{n-1}{n} [\mu_{XY} - \mu_X\mu_Y] \\ &= \frac{n-1}{n} \text{Cov}(X, Y).\end{aligned}$$

From the above, we can see that the estimator is biased because  $\mathbb{E}[\widetilde{S}_{XY}] \neq \text{Cov}(X, Y)$ .

However, the bias can be fixed by multiplying  $\widetilde{S}_{XY}$  by  $\frac{n}{n-1}$  to obtain the following unbiased estimator of  $\text{Cov}(X, Y)$ :

$$\widehat{S}_{XY} = \frac{1}{n-1} \left[ \sum_{i=1}^n (X_i - \bar{X}_n) (Y_i - \bar{Y}_n) \right].$$