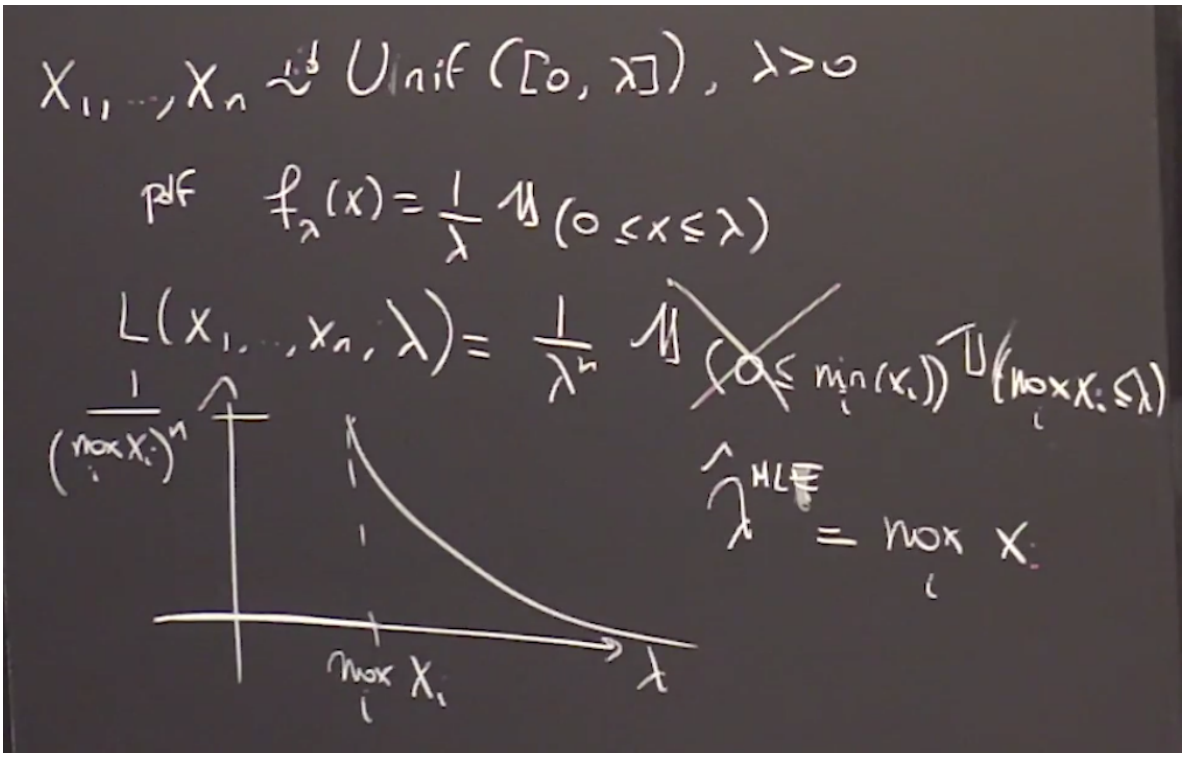
Unit 3 cont'd Methods of Estimation

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 $X_1, \dots, X_n \overset{iid}{\sim} \text{Unif}[0, \theta^*]$ where θ^* 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 θ 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

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

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 θ^* 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 η 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 $X_1, X_2, X_3, \ldots X_n$.

Let the outcome of a set of n rolls of the die be modeled by the i.i.d. random variable sequence $(X_1, \ldots, X_n)$. 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 $x_1, \ldots, x_n$ which contains exactly k outcomes $x_i = 3$. What is the likelihood function $L_n(x_1, \ldots, x_n, \eta)$ for the entire sequence of outcomes?

(Enter **eta** for η.)

eta^k*((1-eta)/5)^(n-k)

✔ **Answer:** eta^k*((1-eta)/5)^(n-k)

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

Solution:

Denote by $p_\eta(x)$ the pmf of X_i. The probability that X_i takes on a value 3 is equal to η 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

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

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 $\operatorname{argmax}_{\eta \in [0,1]} L_n(x_1, \ldots, x_n, \eta)$, we can ignoring any scaling constant in $L_n(x_1, \ldots, x_n, \eta)$. Hence, we will maximize $\widetilde{L}_n(x_1, \ldots, x_n, \eta) = \eta^k (1-\eta)^{n-k}$.

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

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

Remark: The function $\widetilde{L}_n(x_1, \ldots, x_n, \eta) = \eta^k (1-\eta)^{n-k}$ whose maximizer is $\hat{\eta}_n^{\text{MLE}}$ is the same as the likelihood function for a Bernoulli experiment with parameter η, 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 $X_1, X_2, \ldots, X_n$. 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_{θ^*} and p_θ respectively, is defined as

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

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

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

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

- [x] $\mathrm{argmin}_{\theta\in\Theta}\, \widehat{\mathrm{KL}}_n\left(\mathbf{P}_{\theta^*}, \mathbf{P}_\theta\right)$
- [x] $\mathrm{argmax}_{\theta\in\Theta} \sum_{i=1}^{n} \ln p_\theta(X_i)$
- [x] $\mathrm{argmax}_{\theta\in\Theta} \ln\left(\prod_{i=1}^{n} p_\theta(X_i)\right)$
- [x] $\mathrm{argmax}_{\theta\in\Theta} \ln\left(L_n(X_1, X_2, \ldots, X_n; \theta)\right)$

✔

Solution:

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

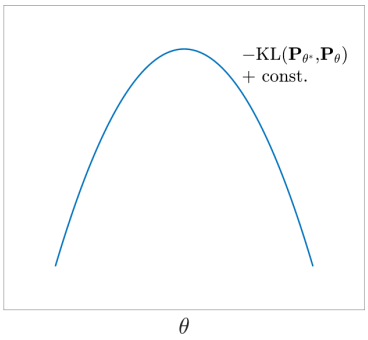
$$\widehat{\theta}_n^{MLE} = \mathrm{argmin}_{\theta\in\Theta}\, \widehat{\mathrm{KL}}_n\left(\mathbf{P}_{\theta^*}, \mathbf{P}_\theta\right).$$

In other words, the maximum likelihood estimator is the (unique) θ that minimizes $\widehat{\mathrm{KL}}\left(\mathbf{P}_{\theta^*}, \mathbf{P}_\theta\right)$ 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:

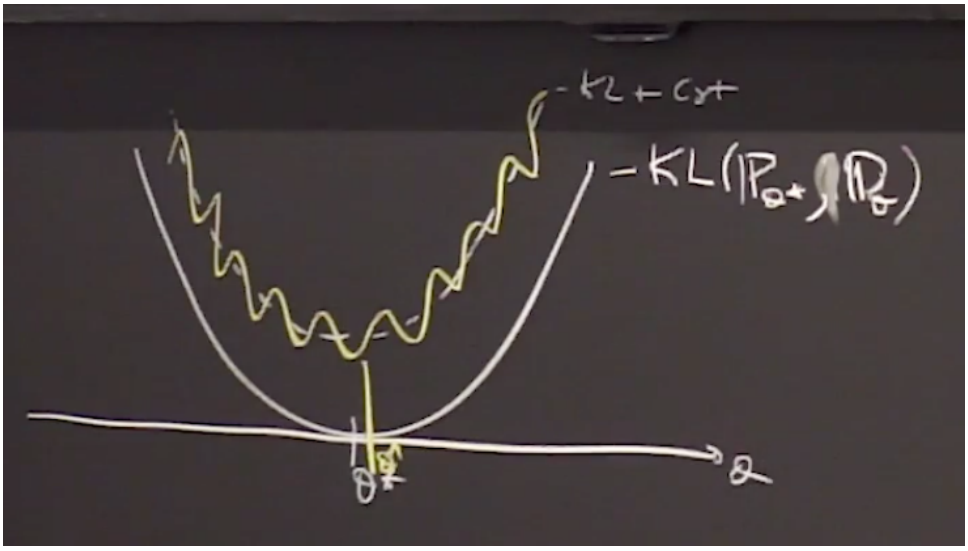
$$\begin{aligned}
\widehat{\theta}_n^{MLE} = \mathrm{argmin}_{\theta\in\Theta}\, \widehat{\mathrm{KL}}_n\left(\mathbf{P}_{\theta^*}, \mathbf{P}_\theta\right) &= \mathrm{argmin}_{\theta\in\Theta}\left(\text{Constant} - \frac{1}{n}\sum_{i=1}^{n} \ln p_\theta(X_i)\right) \\
&= \mathrm{argmax}_{\theta\in\Theta}\left(\frac{1}{n}\sum_{i=1}^{n}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

slide 28

Note: In the following video, at around the 3:20 mark, the plot of $-\mathrm{KL}\left(\mathbf{P}_{\theta^*}, \mathbf{P}_{\theta}\right)$, with θ^* fixed and as a function of θ, 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 $\left(E, \{\mathbf{P}_\theta\}_{\theta \in \Theta}\right)$, the maximum likelihood estimator $\widehat{\theta}_n^{\text{MLE}}$ of θ^* is a **consistent** estimator under mild regularity conditions (e.g. continuity in θ of the pdf p_θ almost everywhere), i.e.

$$\widehat{\theta}_n^{\text{MLE}} \xrightarrow[(p)]{n \to \infty} \theta^*.$$

Note that this is true even if the parameter θ is a vector in a higher dimensional parameter space Θ, and $\widehat{\theta}_n^{\text{MLE}}$ 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 $\mathbf{X} = \left(X^{(1)}, \dots, X^{(d)}\right)^T$ of dimension $d \times 1$ is a vector-valued function from a probability space Ω to $\mathbb{R}^d$:

$$\begin{aligned} \mathbf{X} : \Omega &\to \mathbb{R}^d \\ \omega &\mapsto \begin{pmatrix} X^{(1)}(\omega) \\ X^{(2)}(\omega) \\ \vdots \\ X^{(d)}(\omega) \end{pmatrix} \end{aligned}$$

where each $X^{(k)}$ is a (scalar) random variable on Ω. 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

$$\begin{aligned} F : \mathbb{R}^d &\to [0,1] \\ \mathbf{x} &\mapsto \mathbf{P}(X^{(1)} \le x^{(1)}, \dots, X^{(d)} \le x^{(d)}). \end{aligned}$$

Convergence in Probability in Higher Dimension

To make sense of the consistency statement $\widehat{\theta}_n^{\text{MLE}} \xrightarrow[(p)]{n \to \infty} \theta^*$ where the MLE $\widehat{\theta}_n^{\text{MLE}}$ 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

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

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

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

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

Consistency of the MLE of a Uniform Model

0/1 point (graded)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Unif}[0, \theta^*]$ where θ^* 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 $\widehat{\theta}_n^{\text{MLE}} = \max_{i=1,\ldots,n} X_i$.

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

- [x] $\max_{i=1,\ldots,n} X_i$ is a consistent estimator ✔
- [x] For any $0 < \epsilon \leq \theta^*$, $\mathbf{P}\left(|\max_{i=1,\ldots,n} X_i - \theta^*| \geq \epsilon\right) \to 0$ as $n \to \infty$ ✔
- [x] For any $0 < \epsilon \leq \theta^*$, $\mathbf{P}\left(|\max_{i=1,\ldots,n} X_i - \theta^*| \geq \epsilon\right) \to c$ as $n \to \infty$, where $c > 0$ is a constant
- [ ] For any $0 < \epsilon \leq \theta^*$, $\mathbf{P}\left(|\max_{i=1,\ldots,n} X_i - \theta^*| \geq \epsilon\right) = \left(\frac{\theta^* - \epsilon}{\theta^*}\right)^n$ ✔

✖

Solution:

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

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

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 μ_X and μ_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 μ_X and μ_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:

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

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 μ_X, μ_Y, τ_X, τ_Y, and c, what is $\text{Cov}(A, B)$?

(Enter **mu_X** for μ_X, **tau_X** for τ_X.)

c - (tau_X + mu_X^2 - ta ✖ **Answer:** tau_X-tau_Y

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

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 α_0 and α_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 $Z = X$ add 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

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

- **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 $p_X(x=0)\, p_Z(z=0)$:

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

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 $z = x \text{ add } y$. 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 $\widehat{X}$ 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
- [x] Scenario 2
- [x] 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:

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

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 $\widehat{X}$ 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

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

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:

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

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 $P[\widehat{X} \neq X]$ 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 $\implies$ Cov $(X, Y) = 0$."

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

- (●) True
- ( ) False

✔

"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** ?

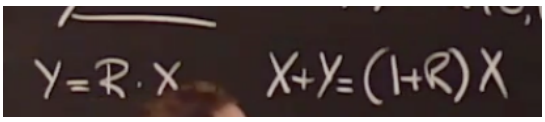
- ( ) True
- (●) False

✔

Solution:

- **True.** If X, Y are independent, 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)



Sample Covariance

1/4 points (graded)

Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \overset{\text{iid}}{\sim} (X, Y)$ 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:

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

where $\overline{X}_n$ and $\overline{Y}_n$ denote the sample means estimators of μ_X and μ_Y.

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

(Enter **mu_{XY}** for μ_{XY}, **mu_X** for μ_X, and **mu_Y** for μ_Y.)

$\mathbb{E}\left[\frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{j=1}^{n} Y_j\right)}{n}\right] =$ (mu_{XY}-mu_X*mu_Y)/n ✖

Answer: (1/n)*(n*mu_XY + n*(n-1)*mu_X*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, μ_X, μ_Y, and μ_{XY}.

(Enter **mu_{XY}** for μ_{XY}, **mu_X** for μ_X, and **mu_Y** for μ_Y.)

$c =$ n ✖ **Answer:** (n/(n-1))

$$n$$

First,

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

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

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

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

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

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

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

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

8. Covariance Matrices

Exercises due Jul 1, 2020 08:59 JST

Bookmark this page

Note: Now is a good time to review the matrix exercises in Homework 0.

Note on Notation: In this course, we assume all vectors to be column vectors. Therefore, while

$$\mathbf{X} = \begin{bmatrix} X^{(1)} \\ X^{(2)} \\ \vdots \\ X^{(d)} \end{bmatrix},$$

we sometimes write it as $\mathbf{X} = \left(X^{(1)}, \ldots, X^{(d)}\right)^T$ to be more compact in representation.

Example of Covariance II

1/4 points (graded)

Let X, Y be random variables such that

- X takes the values ± 1 each with probability 0.5
- (Conditioned on X) Y is chosen uniformly from the set $\{-3X - 1, -3X, -3X + 1\}$.

(Round all answers to 2 decimal places.)

What is $\mathrm{Cov}(X, X)$ (equivalent to $\mathrm{Var}(X)$)?

$\mathrm{Cov}(X, X) =$ 1 ✔ **Answer:** 1.0

What is $\mathrm{Cov}(Y, Y)$ (equivalent to $\mathrm{Var}(Y)$)?

$\mathrm{Cov}(Y, Y) =$ -1.5 ✖ **Answer:** 9.67

What is $\mathrm{Cov}(X, Y)$?

$\mathrm{Cov}(X, Y) =$ ✖ **Answer:** -3.00

What is $\mathrm{Cov}(Y, X)$?

$\mathrm{Cov}(Y, X) =$ ✖ **Answer:** -3.00

Solution:

Observe that $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ are both zero, since X is uniformly distributed over $\{\pm 1\}$ and Y is uniformly distributed over the set $\{-4, -3, -2, 2, 3, 4\}$.

- $\mathrm{Cov}(X, X)$ is the variance of X, which equals $\mathbb{E}[X^2] - \mathbb{E}[X]^2 = p + (1 - p) = 1$.
- $\mathrm{Cov}(Y, Y)$ is the variance of Y, which equals $\mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \frac{16+9+4+4+9+16}{6} = \frac{29}{3} \approx 9.67$.
- $\mathrm{Cov}(X, Y)$ and $\mathrm{Cov}(Y, X)$ are always equal, by the symmetry of the definition. Observe that the joint density of (X, Y) is uniform over the pairs $(1, -4), (1, -3), (1, -2), (-1, 2), (-1, 3), (-1, 4)$. Thus, either covariance can be computed as $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \frac{-4-3-2-2-3-4}{6} = -3$.

Covariance Matrix

4/4 points (graded)

Given random variables $X^{(1)}, X^{(2)}, \ldots, X^{(d)}$, one can write down the **covariance matrix** Σ, where $\Sigma_{i,j} = \text{Cov}(X^{(i)}, X^{(j)})$.

Let $X^{(1)}, X^{(2)}$ be random variables such that

- $X^{(1)}$ takes the values ± 1 each with probability 0.5
- (Conditioned on $X^{(1)}$) $X^{(2)}$ is chosen uniformly from the set $\{-3X^{(1)} - 1, -3X^{(1)}, -3X^{(1)} + 1\}$.

What is the covariance matrix Σ?

$\Sigma_{1,1} =$ 1 ✔ **Answer:** 1.0 $\Sigma_{1,2} =$ -3.00 ✔ **Answer:** -3.00

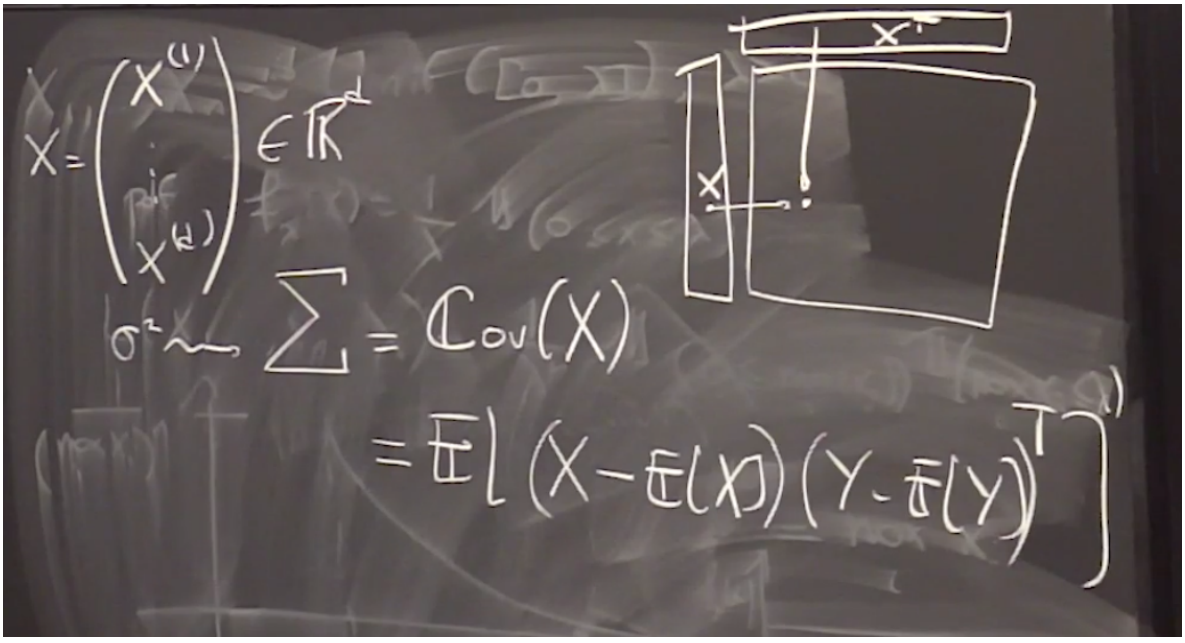
$\Sigma_{2,1} =$ -3.00 ✔ **Answer:** -3.00 $\Sigma_{2,2} =$ 9.67 ✔ **Answer:** 9.67

Solution:

Using the answer to the previous question, the 2×2 covariance matrix Σ evaluates to

$$\Sigma = \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix} = \begin{pmatrix} 1 & -3 \\ -3 & \frac{29}{3} \end{pmatrix}$$

slide 32
want a matrix of covariances so have to take the outer product
this is a vector multiplied by the transpose of another vector
(taking a dot product would just give one number)



Here is a compact formula for the covariance matrix using vector notation.

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

Let $\mu \triangleq \mathbb{E}[\mathbf{X}]$ denote the **entry-wise** mean, i.e.

$$\mathbb{E}[\mathbf{X}] = \begin{pmatrix} \mathbb{E}[X^{(1)}] \\ \vdots \\ \mathbb{E}[X^{(d)}] \end{pmatrix}.$$

Consider the vector outer product (refer to Homework 0) $(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T$, which is a random $d \times d$ matrix. Then the **covariance matrix** Σ can be written as

$$\Sigma = \mathbb{E}[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T].$$

Covariance Matrix: Properties I

0/1 point (graded)

Let $\mathbf{X}$ be a random vector and let $\mathbf{Y} = \mathbf{X} + \boldsymbol{B}$, where $\boldsymbol{B}$ is a constant vector. Let $\mu_{\mathbf{X}}$ be the mean vector of $\mathbf{X}$ and let $\Sigma_{\mathbf{X}}$ be the covariance matrix of $\mathbf{X}$. Select from the following all statements that are correct.

- [ ] The covariance matrix of $\mathbf{Y}$ could potentially be equal to $\Sigma_{\mathbf{X}}$ only under some conditions imposed on $\boldsymbol{B}$
- [ ] The covariance matrix of $\mathbf{Y}$ is the same as $\Sigma_{\mathbf{X}}$ for all vectors $\boldsymbol{B}$ ✔
- [x] The covariance matrix of $\mathbf{Y}$ has the same size as the matrix $\Sigma_{\mathbf{X}}$ ✔
- [x] The covariance matrix of $\mathbf{Y}$ is the same as $\Sigma_{\mathbf{X}}$ if and only if vector $\boldsymbol{B}$ is equal to 0

✖

Solution:

Choices 2 and 3 are correct. Let the covariance matrix of $\mathbf{Y}$ be denoted $\Sigma_{\mathbf{Y}}$. Note that $\mathbb{E}[\mathbf{X} + \boldsymbol{B}] = \mu_{\mathbf{X}} + \boldsymbol{B}$ for any vector $\boldsymbol{B}$.

$$\Sigma_{\mathbf{Y}} = \mathbb{E}\left[(\mathbf{X} + \boldsymbol{B} - \mu_{\mathbf{X}} - \boldsymbol{B})(\mathbf{X} + \boldsymbol{B} - \mu_{\mathbf{X}} - \boldsymbol{B})^T\right] = \Sigma_{\mathbf{X}}$$

Since choice 2 is correct, choices 1 and 4 that impose certain conditions on $\boldsymbol{B}$ are technically incorrect as we do not require that $\boldsymbol{B}$ satisfiy some conditions for $\Sigma_{\mathbf{Y}}$ to be the same as $\Sigma_{\mathbf{X}}$.

Covariance Matrix: Properties II

0/1 point (graded)

Let $\mathbf{X}$ be a random vector of size $d \times 1$ and let $\mathbf{Y} = A\mathbf{X} + B$, where A is a constant matrix of size $n \times d$ and B is a constant vector of size $n \times 1$. Let $\mu_{\mathbf{X}}$ be the mean vector of $\mathbf{X}$ and let $\Sigma_{\mathbf{X}}$ be the covariance matrix of $\mathbf{X}$. Let $\mu_{\mathbf{Y}}$ be the mean vector of $\mathbf{Y}$ and let $\Sigma_{\mathbf{Y}}$ be the covariance matrix of $\mathbf{Y}$.

Select from the following all statements that are correct.

- [x] $\Sigma_{\mathbf{Y}}$ is the same as covariance matrix of $A\mathbf{X}$ ✔
- [x] $\Sigma_{\mathbf{Y}}$ is of size $n \times n$ ✔
- [ ] $\Sigma_{\mathbf{Y}} = A^2\Sigma_{\mathbf{X}}$
- [x] $\Sigma_{\mathbf{Y}} = A\Sigma_{\mathbf{X}}A^T$ ✔
- [ ] $\Sigma_{\mathbf{Y}} = A^T\Sigma_{\mathbf{X}}A$

Solution:

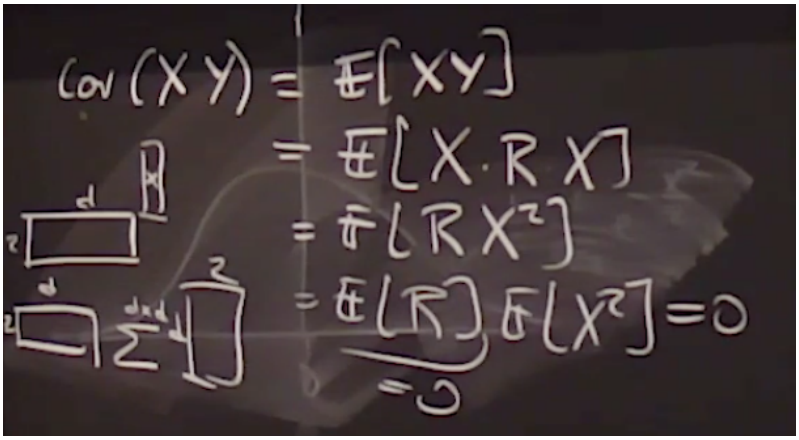
As $\mathbf{Y}$ is an $n \times 1$ random vector, $\Sigma_{\mathbf{Y}}$ is of size $n \times n$.

From the previous problem we know that $\Sigma_{\mathbf{Y}}$ is the same as the covariance matrix of $A\mathbf{X}$. Therefore, it suffices to find this matrix, which we denote $\Sigma_{A\mathbf{X}}$.

$$\begin{aligned}
\Sigma_{A\mathbf{X}} &= \mathbb{E}\left[(A\mathbf{X} - A\mu_X)(A\mathbf{X} - A\mu_X)^T\right] \\
&= \mathbb{E}\left[A(\mathbf{X} - \mu_X)\left(\mathbf{X}^T A^T - \mu_X^T A^T\right)\right] \\
&= \mathbb{E}\left[A(\mathbf{X} - \mu_X)(\mathbf{X} - \mu_X)^T A^T\right] \\
&= A\mathbb{E}\left[(\mathbf{X} - \mu_X)(\mathbf{X} - \mu_X)^T\right] A^T \\
&= A\Sigma_{\mathbf{X}}A^T.
\end{aligned}$$

Therefore, choices 1, 2, and 4 are correct.

Choices 3 and 5 are not correct in general (even if A is a square matrix) because matrix multiplication is not commutative.



Effect of Linear Transformations of Covariance Matrix

4/4 points (graded)

Let $\mathbf{X} = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}$ be a random vector with covariance Matrix $\Sigma_{\mathbf{X}} = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}$.

Let $\mathbf{Y} = M\mathbf{X}$, where $M = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$.

Observe that $Y^{(1)} = X^{(1)} - X^{(2)}$ and $Y^{(2)} = X^{(1)} + X^{(2)}$. What is the new covariance matrix Σ_Y?

$(\Sigma_{\mathbf{Y}})_{1,1} =$ 1 ✔ **Answer:** 1 $(\Sigma_{\mathbf{Y}})_{1,2} =$ 0 ✔ **Answer:** 0.0

$(\Sigma_{\mathbf{Y}})_{2,1} =$ 0 ✔ **Answer:** 0.0 $(\Sigma_{\mathbf{Y}})_{2,2} =$ 3 ✔ **Answer:** 3

Solution:

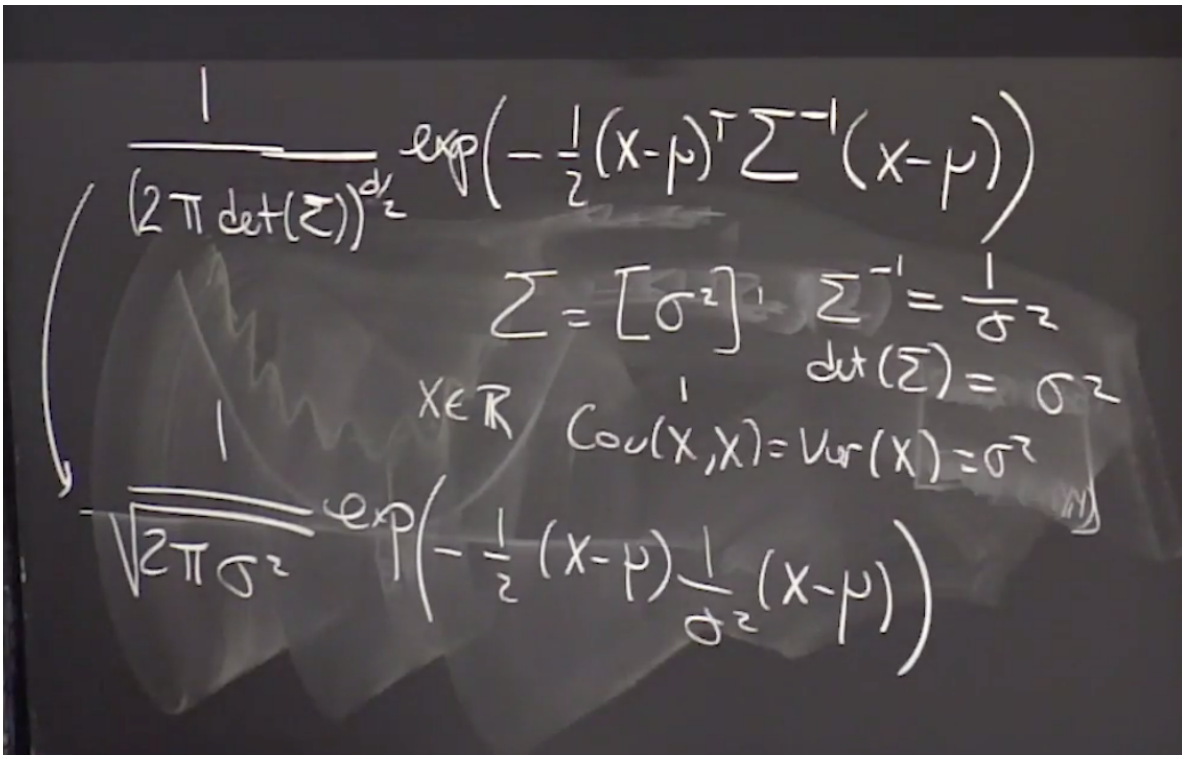
Recall from an earlier problem that for any pair of random variables A, B with the same variance $\mathrm{Var}(A) = \mathrm{Var}(B) = \sigma^2$, $\mathrm{Cov}(A - B, A + B) = \mathrm{Var}(A) - \mathrm{Var}(B) = 0$.

Therefore, given the matrix M, $\Sigma_{\mathbf{Y}}$ must be a diagonal matrix.

We have
$\mathrm{Cov}(Y^{(1)}, Y^{(1)}) = \mathrm{Cov}(X^{(1)} - X^{(2)}, X^{(1)} - X^{(2)}) = \mathrm{Cov}(X^{(1)}, X^{(1)}) - 2\mathrm{Cov}(X^{(1)}, X^{(2)}) + \mathrm{Cov}(X^{(2)}, X^{(2)}) = 1 - 1 + 1 = 1.$

Similarly,
$\mathrm{Cov}(Y^{(2)}, Y^{(2)}) = \mathrm{Cov}(X^{(1)} + X^{(2)}, X^{(1)} + X^{(2)}) = \mathrm{Cov}(X^{(1)}, X^{(1)}) + 2\mathrm{Cov}(X^{(1)}, X^{(2)}) + \mathrm{Cov}(X^{(2)}, X^{(2)}) = 1 + 1 + 1 = 3.$

slide 33
Gaussian pdf



Multivariate Gaussian Random Variable

A random vector $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)})^T$ is a **Gaussian vector**, or **multivariate Gaussian or normal variable**, if any linear combination of its components is a (univariate) Gaussian variable or a constant (a "Gaussian" variable with zero variance), i.e., if $\alpha^T\mathbf{X}$ is (univariate) Gaussian or constant for any constant non-zero vector $\alpha \in \mathbb{R}^d$.

The distribution of $\mathbf{X}$, the ***d*-dimensional Gaussian or normal distribution**, is completely specified by the vector mean $\mu = \mathbb{E}[\mathbf{X}] = (\mathbb{E}[X^{(1)}], \ldots, \mathbb{E}[X^{(d)}])^T$ and the $d \times d$ covariance matrix Σ. If Σ is invertible, then the pdf of $\mathbf{X}$ is

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}, \quad \mathbf{x} \in \mathbb{R}^d$$

where $\det(\Sigma)$ is the determinant of the Σ, which is positive when Σ is invertible.

If $\mu = \mathbf{0}$ and Σ is the identity matrix, then $\mathbf{X}$ is called a **standard normal random vector**.

Note that when the covariant matrix Σ is diagonal, the pdf factors into pdfs of univariate Gaussians, and hence the components are independent.

Linear Transformation of a Multivariate Gaussian Random Vector

0/1 point (graded)

Consider the 2-dimensional Gaussian $\mathbf{X} = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}$ with covariance matrix $\Sigma_X = \begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}$ and mean $\mu_{\mathbf{X}} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$.

Consider the vector $\alpha = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$, so that $Y = \alpha^T\mathbf{X}$ is a 1-dimensional Gaussian.

What is the variance $\mathrm{Var}(Y)$ of Y?

$\mathrm{Var}(Y) =$ 1 ✖ **Answer:** 2

Solution:

One way to answer this is to notice that $Y = X^{(1)} - X^{(2)}$, so

$$\mathrm{Var}(Y) = \mathrm{Cov}(Y, Y) = \mathrm{Var}(X^{(1)}) + \mathrm{Var}(X^{(2)}) - 2\mathrm{Cov}(X^{(1)}, X^{(2)}) = 1 + 5 - 4 = 2.$$

Another way is to define the matrix $M \triangleq \alpha^T = \begin{pmatrix} 1 & -1 \end{pmatrix}$, and apply the formula $\Sigma_Y = M\Sigma_{\mathbf{X}}M^T = 2$.

Singular Covariance Matrices

1/1 point (graded)

Consider again a 2-dimensional Gaussian $\mathbf{X} = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}$. But instead, Σ_X is $\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$ and $\alpha = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$, what is the variance $\mathrm{Var}(Y)$ of $Y = \alpha^T\mathbf{X}$?

$\mathrm{Var}(Y) =$ 0 ✔ **Answer:** 0

This result tells us that the Gaussian $(X^{(1)}, X^{(2)})^T$ is actually a one-dimensional Gaussian, orthogonal to the direction of $\begin{pmatrix} 2 \\ -1 \end{pmatrix}$.

Solution:

Define a matrix $M = \alpha^T$. We have $\Sigma_Y = M\Sigma_X M^T = 0$, since M^T is a column vector in the nullspace of Σ_X.

Such a Gaussian (with a singular covariance matrix) is sometimes referred to as a **degenerate** Gaussian.

(Optional) Gaussian Random Vectors I

0 points possible (ungraded)

Recall from an earlier part of this lecture that the covariance between two random variables being 0 does not necessarily imply that the random variables are independent. However, this is true if the random variables are multivariate Gaussian.

Let $\mathbf{X}$ be a Gaussian random vector with mean μ and covariance Σ. Assume that Σ is positive definite. Determine if the following statement is true or false.

"There exists a vector B and a matrix A such that $A(\mathbf{X}+B)$ is a Gaussian random vector whose components are independent and each of mean 0".

- (●) True
- ( ) False

✔

Hint: Refer to the note above on diagonalization of the covariance matrix.

Solution:

True. First, in order to remove the effect of μ we can set $B=-\mu$ to make the individual Gaussian random variables be of zero mean. Let $\widehat{\mathbf{X}}=\mathbf{X}-\mu$. From an earlier problem we know that the covariance matrix of $\widehat{\mathbf{X}}$ is the same as Σ.

From the above note on covariance matrices we can see that there exists an orthogonal matrix U such that $D=U\Sigma U^T$.

Consider the following transformation: $\mathbf{Y}=U\widehat{\mathbf{X}}$.

The covariance matrix of $\mathbf{Y}$ is (from an earlier problem)

$$U\Sigma U^T,$$

which is precisely equal to the diagonal matrix D. Therefore, $\mathbf{Y}$ has component Gaussian random variables that are uncorrelated and hence independent.

Vector Version of the Central Limit Theorem

0/1 point (graded)

Let $\mathbf{X}$ be a random vector of dimension $d\times 1$ and let μ and Σ be its mean and covariance. Let $\mathbf{X}_1,\ldots,\mathbf{X}_n$ be i.i.d. copies of $\mathbf{X}$. Let $\overline{\mathbf{X}}_n \triangleq \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i$.

Based on your knowledge of the central limit theorem for a single random variable, select from the following the correct shift and scale factor for $\overline{\mathbf{X}}_n$ so that $\overline{\mathbf{X}}_n$ could potentially converge to the Gaussian random vector $\mathcal{N}(0, I_{d\times d})$.

- ( ) $\sqrt{d}\cdot\Sigma^{-\frac{1}{2}}\left(\overline{\mathbf{X}}_n-\mu\right)$
- ( ) $\sqrt{d}\cdot\Sigma^{-1}\left(\overline{\mathbf{X}}_n-\mu\right)$
- (●) $\sqrt{n}\cdot\Sigma^{-1}\left(\overline{\mathbf{X}}_n-\mu\right)$
- ( ) $\sqrt{n}\cdot\Sigma^{-\frac{1}{2}}\left(\overline{\mathbf{X}}_n-\mu\right)$ ✔
- ( ) None of the above

✖

Solution:

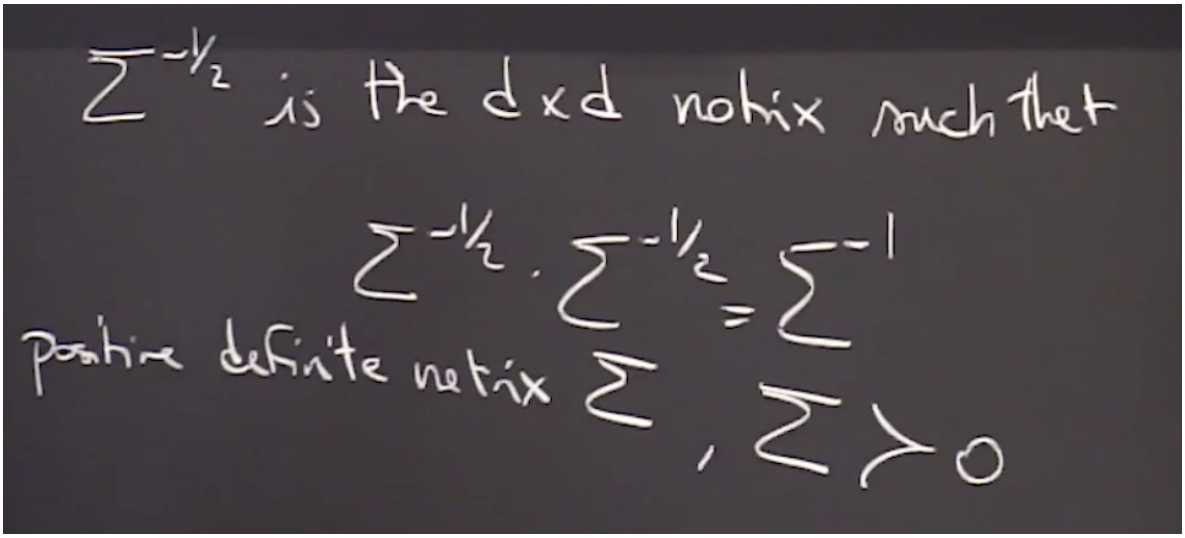
Solution:

The shift of $\mathbf{X}$ by μ is the correct shift that needs to be applied in order to center the random vector.

The scaling factor should be $\sqrt{n}\Sigma^{-\frac{1}{2}}$ because it mimics the single variable CLT case most closely. In particular, the division by $\sqrt{\sigma^2}$ in the single variable CLT case is being taken care of by the inverse of the square root of Σ.

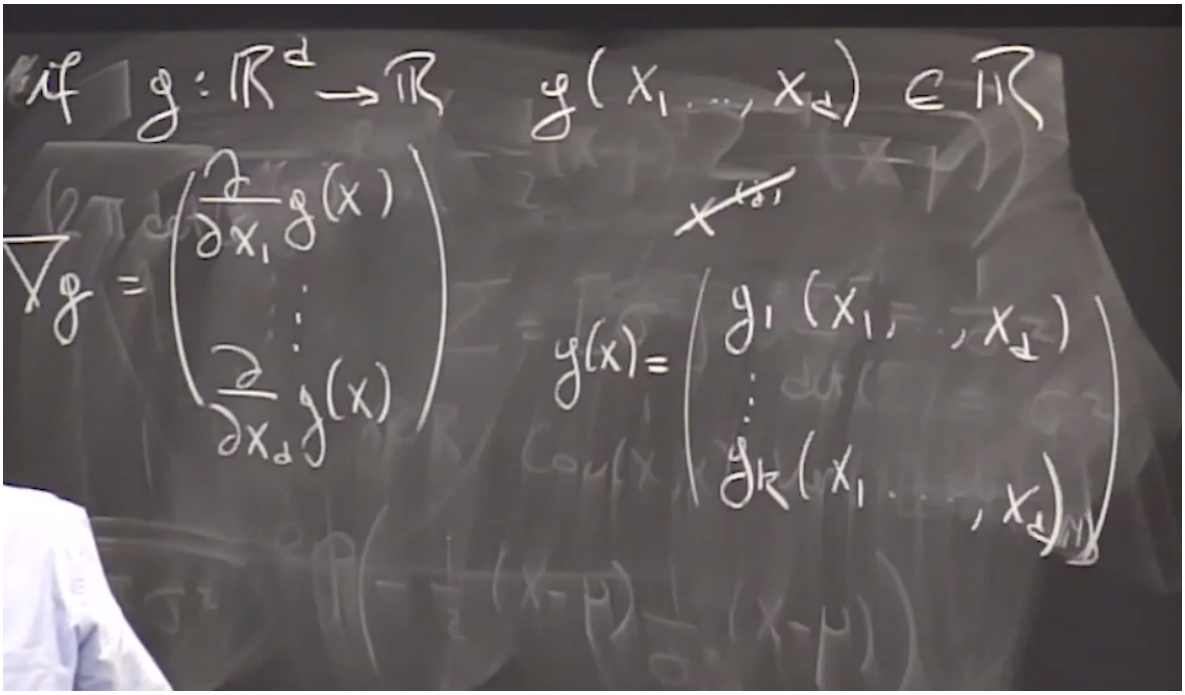
Note: Of course, this is only a heuristic discussion that is meant to test how you can potentially generalize the single variable CLT. This is not a proof and the solution is also written as guesswork.

slide 34

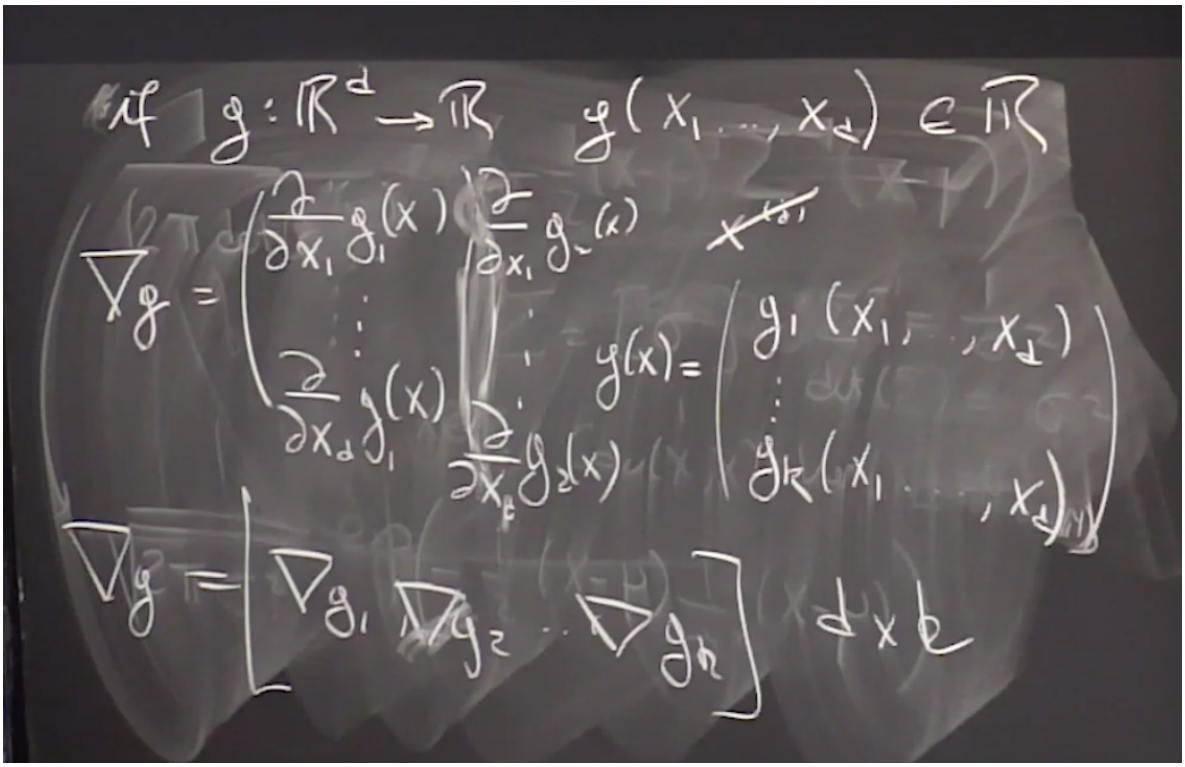


any positive definite matrix is a covariance matrix

slide 35
Delta method, take a function that converges to Xn_bar and put it into another function that converges to the mean



can stack the functions into a gradient matrix



each g vector is of size dx1

Gradient Matrix of a Vector Function

3/4 points (graded)

Given a vector-valued function $f : \mathbb{R}^d \to \mathbb{R}^k$, the **gradient** or the **gradient matrix** of f, denoted by ∇f, is the $d \times k$ matrix

$$\nabla f = \begin{pmatrix} | & | & | & | \\ \nabla f_1 & \nabla f_2 & \ldots & \nabla f_k \\ | & | & | & | \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \ldots & \frac{\partial f_k}{\partial x_1} \\ \vdots & \ldots & \vdots \\ \frac{\partial f_1}{\partial x_d} & \ldots & \frac{\partial f_k}{\partial x_d} \end{pmatrix}.$$

This is also the transpose of what is known as the **Jacobian matrix** $\mathbf{J}_f$ of f.

Let $f(x, y, z) = \begin{pmatrix} x^2 + y^2 + z^2 \\ 2xy \\ y^3 + z^3 \\ z^4 \end{pmatrix}$.

How many rows does $\nabla f(x, y, z)$ have?

3 ✔ **Answer:** 3

How many columns does $\nabla f(x, y, z)$ have?

4 ✔ **Answer:** 4

What does column 2 represent in the gradient matrix?

- (●) Derivative of $2xy$ with respect to x, y, z stacked as a column
- ( ) Derivative of the individual functions with respect to y stacked as a column

✔

What is $\nabla f(x, y, z)_{3,2}$?

2 ✖ **Answer:** 0*x

2

Solution:

According to notation developed in the video, the gradient for f is of size 3×4 because it is a function of 3 variables and it outputs 4 values as a column. Column $j \in \{1, 2, 3, 4\}$ of the gradient matrix represents the derivative of the j^{th} function of $f(x, y, z)$ with respect to x, y, z stacked as a column.

$\nabla f(x, y, z)_{3,2}$ is the derivative of function $2xy$ (2nd function) with respect to z (3rd variable). This derivative is equal to 0.

General Statement of the Multivariate Delta Method

The multivariate delta method states that given

- a sequence of random vectors $(\mathbf{T}_n)_{n\geq 1}$ satisfying $\sqrt{n}\left(\mathbf{T}_n - \vec{\theta}\right) \xrightarrow[n\to\infty]{(d)} \mathbf{T}$,
- a function $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^k$ that is continuously differentiable at $\vec{\theta}$,

then

$$\sqrt{n}\left(\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\vec{\theta})\right) \xrightarrow[n\to\infty]{(d)} \nabla\mathbf{g}(\vec{\theta})^T \mathbf{T} \qquad \text{where } \nabla\mathbf{g} = \begin{pmatrix} | & | & \dots & | \\ \nabla\mathbf{g}_1 & \nabla\mathbf{g}_2 & \dots & \nabla\mathbf{g}_k \\ | & | & \dots & | \end{pmatrix}.$$

Common Application

In the lecture and in most applications, $\mathbf{T}_n = \overline{\mathbf{X}}_n$ where $\overline{\mathbf{X}}_n$ is the sample average of $\mathbf{X}_1, \dots, \mathbf{X}_n \overset{iid}{\sim} \mathbf{X}$, and $\vec{\theta} = \mathbb{E}[\mathbf{X}]$. The (multivariate) CLT then gives $\mathbf{T} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{X}})$ where $\Sigma_{\mathbf{X}}$ is the covariance of $\mathbf{X}$. In this case, we have

$$\sqrt{n}\left(\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\vec{\theta})\right) \xrightarrow[n\to\infty]{(d)} \nabla\mathbf{g}(\vec{\theta})^T \mathbf{T} \sim \mathcal{N}\left(0, \nabla\mathbf{g}(\vec{\theta})^T \Sigma_{\mathbf{X}} \nabla\mathbf{g}(\vec{\theta})\right) \qquad (\mathbf{T} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{X}})).$$