Unit 3 Methods of Estimation

Interpreting Total Variation Distance

1/1 point (graded)

Recall from lecture that the **total variation distance** between two probability measures $\mathbf{P}_\theta$ and $\mathbf{P}_{\theta'}$ with sample space E is defined by

$$\mathrm{TV}\left(\mathbf{P}_\theta, \mathbf{P}_{\theta'}\right) = \max_{A \subset E} \left| \mathbf{P}_\theta\left(A\right) - \mathbf{P}_{\theta'}\left(A\right) \right|$$

Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathbf{P}_{\theta^*}$ where $\theta^* \in \mathbb{R}$ is an unknown parameter. You construct a statistical model $\left(E, \{\mathbf{P}_\theta\}_{\theta \in \mathbb{R}}\right)$ for your data. By analyzing your data, you are able to produce an estimator $\hat{\theta}$ such that the distributions $\mathbf{P}_{\hat{\theta}}$ and $\mathbf{P}_{\theta^*}$ are close in **total variation distance**. More precisely, you know that

$$\mathrm{TV}\left(\mathbf{P}_{\hat{\theta}}, \mathbf{P}_{\theta^*}\right) \leq \epsilon,$$

where ϵ is a very small positive number.

Which of the following can you conclude about the distributions $\mathbf{P}_{\hat{\theta}}$ and $\mathbf{P}_{\theta^*}$? (Choose all that apply.)

- [x] Let A be an event. Then $|\mathbf{P}_{\theta^*}\left(A\right) - \mathbf{P}_{\hat{\theta}}\left(A\right)| \leq \epsilon$.
- [x] Let $X \sim \mathbf{P}_{\theta^*}$, let $Y \sim \mathbf{P}_{\hat{\theta}}$, and suppose $a, b \in \mathbb{R}$ where $a \leq b$. Then $|\mathbf{P}_{\theta^*}\left(a \leq X \leq b\right) - \mathbf{P}_{\hat{\theta}}\left(a \leq Y \leq b\right)| \leq \epsilon$.
- [ ] $|\theta^* - \hat{\theta}| \leq \epsilon$.

✔

Solution:

Recall that by definition,

$$\mathrm{TV}\left(\mathbf{P}_{\hat{\theta}}, \mathbf{P}_{\theta^*}\right) = \max_{A \subset E} |\mathbf{P}_{\hat{\theta}}\left(A\right) - \mathbf{P}_{\theta^*}\left(A\right)|$$

where the maximum is over all events A. Since we are given that $\mathrm{TV}\left(\mathbf{P}_{\hat{\theta}}, \mathbf{P}_{\theta^*}\right) \leq \epsilon$, we conclude that $|\mathbf{P}_{\hat{\theta}}\left(A\right) - \mathbf{P}_{\theta^*}\left(A\right)| \leq \epsilon$ for every event A. Hence, the first choice is correct.

Let A be the event given by the interval (a, b). Then,

$$|\mathbf{P}_{\theta^*}\left(a \leq X \leq b\right) - \mathbf{P}_{\hat{\theta}}\left(a \leq Y \leq b\right)| \leq \epsilon$$

is the same as saying $|\mathbf{P}_{\hat{\theta}}\left(A\right) - \mathbf{P}_{\theta^*}\left(A\right)| \leq \epsilon$. Thus, the second choice is true as well.

The third choice, "$|\theta^* - \hat{\theta}| \leq \epsilon$.", is incorrect. In general, even if distributions $\mathbf{P}_{\hat{\theta}}$ and $\mathbf{P}_{\theta^*}$ are close, there is no reason to expect the parameters θ^* and $\hat{\theta}$ to be close. To conclude that the estimated parameter is close to the true parameter given their distributions are close, we would need some assumptions on the map $\theta \mapsto \mathbf{P}_\theta$. No such assumption is given here.

5. Total Variation Distance for Discrete Random Variables

Bookmark this page

Quiz: Probability Mass Functions

1/1 point (graded)

Let X be a discrete random variable whose sample space is $\mathbb{Z}$, the set of integers. Let $p : \mathbb{Z} \to [0, 1]$ denote the **probability mass function** (**pmf**) of X. What does $p(7) + p(10)$ represent?

- ○ The probability that $X = 10$.
- ○ The probability that $X = 7$.
- ◉ The probability that $X = 7$ **or** $X = 10$.
- ○ The probability that $X = 7$ **and** $X = 10$.

✔

Solution:

By definition, $p(7) + p(10) = P(X = 10) + P(X = 7)$. The events $X = 10$ and $X = 7$ are disjoint, so in fact $p(7) + p(10) = P(X = 10 \text{ or } X = 7)$.

Preparation: Probability of Complements

1/1 point (graded)

What is $\mathbf{P}_\theta(A^c) - \mathbf{P}_{\theta'}(A^c)$ in terms of $\mathbf{P}_\theta(A)$ and $\mathbf{P}_{\theta'}(A)$? (Recall A^c is the complement of A in the sample space.)

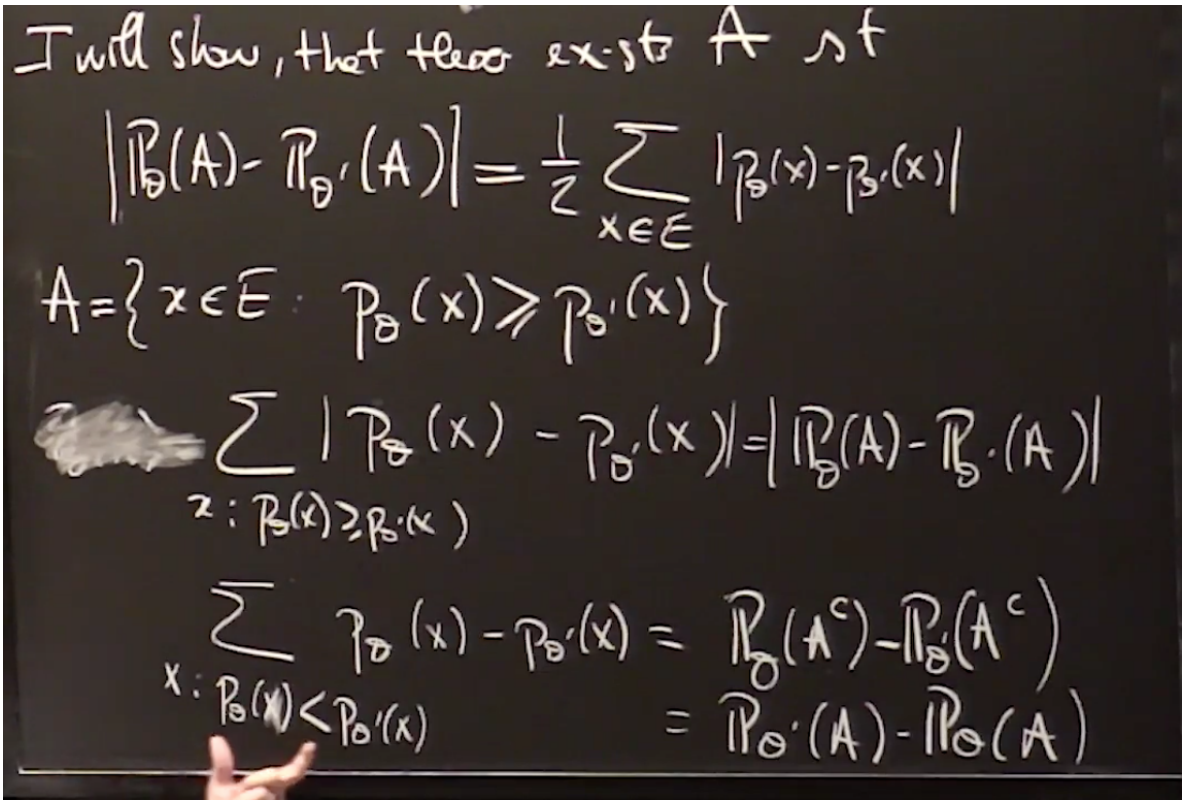
- ◉ $\mathbf{P}_{\theta'}(A) - \mathbf{P}_\theta(A)$
- ○ $\mathbf{P}_\theta(A) - \mathbf{P}_{\theta'}(A)$

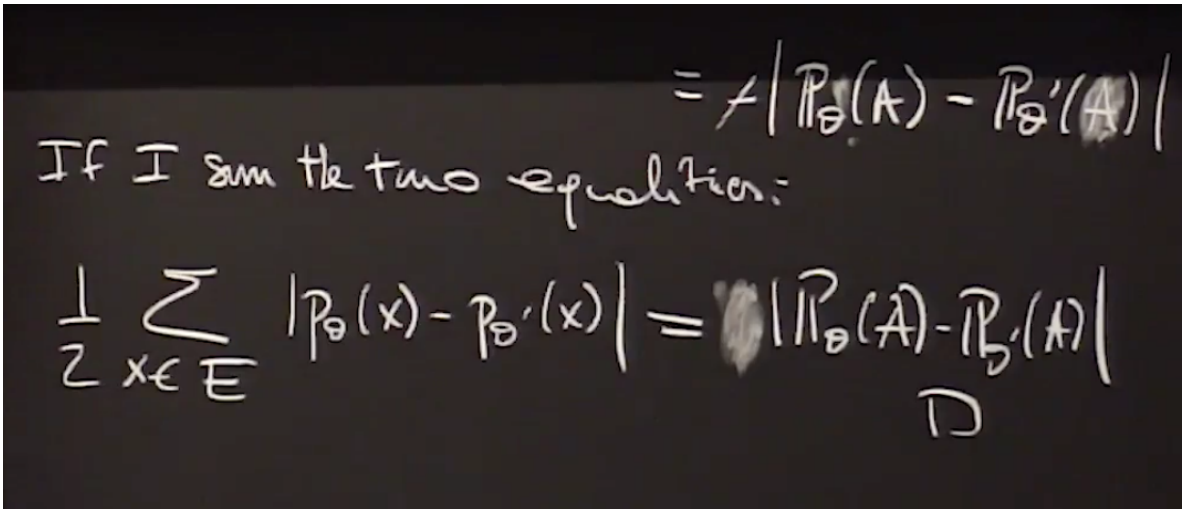
✔

Solution:

$$\mathbf{P}_\theta(A^c) - \mathbf{P}_{\theta'}(A^c) = (1 - \mathbf{P}_\theta(A)) - (1 - \mathbf{P}_{\theta'}(A)) = \mathbf{P}_{\theta'}(A) - \mathbf{P}_\theta(A).$$

Using the fact that when we know something is positive then we can write it as an absolute value and if we know something is negative then we can write it as minus the absolute value
Because the p_theta is always > p_theta’





this shows where the 1/2 comes from (rubbed out bit on right was 2 then divided both sides by 2)

Let $\mathbf{P}$ and $\mathbf{Q}$ be probability measures with a discrete sample space E and probability mass functions f and g. Then, the total variation distance between $\mathbf{P}$ and $\mathbf{Q}$:

$$\mathrm{TV}(\mathbf{P}, \mathbf{Q}) = \max_{A \subset E} |\mathbf{P}(A) - \mathbf{Q}(A)|,$$

can be computed as

$$\mathrm{TV}(\mathbf{P}, \mathbf{Q}) = \frac{1}{2} \sum_{x \in E} |f(x) - g(x)|.$$

Equivalence of Formulas

4/4 points (graded)

Let $E = \{1, 2, 3, 4\}$ be a discrete sample space. Let $\mathbf{P}$ and $\mathbf{Q}$ be probability measures with probability mass functions f and g as follows:

$$f(1) = 1/4,\ f(2) = 1/4,\ f(3) = 1/8,\ f(4) = 3/8$$

$$g(1) = g(2) = g(3) = g(4) = 1/4$$

Find the value of $|\mathbf{P}(A) - \mathbf{Q}(A)|$ for the following choices of A.

For $A = \{3\}$:

$|\mathbf{P}(A) - \mathbf{Q}(A)| =$ 1/8 ✔

For $A = \{4\}$:

$|\mathbf{P}(A) - \mathbf{Q}(A)| =$ 1/8 ✔

For $A = \{3, 4\}$?

$|\mathbf{P}(A) - \mathbf{Q}(A)| =$ 0 ✔

What is the value of $\max_{A \subset E} |\mathbf{P}(A) - \mathbf{Q}(A)|$?

1/8 ✔

Solution:

First, compute $|\mathbf{P}(A) - \mathbf{Q}(A)|$ for the different choices of A:

- When $A = \{3\}$, $\mathbf{P}(A) = f(3) = 1/8$ and $\mathbf{Q}(A) = g(3) = 1/4$. Therefore, $|\mathbf{P}(A) - \mathbf{Q}(A)| = 1/8$.
- When $A = \{4\}$, $\mathbf{P}(A) = f(4) = 3/8$ and $\mathbf{Q}(A) = g(4) = 1/4$. Therefore, $|\mathbf{P}(A) - \mathbf{Q}(A)| = 1/8$.
- When $A = \{3, 4\}$, $\mathbf{P}(A) = f(3) + f(4) = 1/2$ and $\mathbf{Q}(A) = g(3) + g(4) = 1/2$. Therefore, $|\mathbf{P}(A) - \mathbf{Q}(A)| = 0$.

Now, we find $\max_{A \subset E} |\mathbf{P}(A) - \mathbf{Q}(A)|$. We have already considered $A = \{3\}$, $A = \{4\}$, and $A = \{3, 4\}$. For any other non-empty set A, $|\mathbf{P}(A) - \mathbf{Q}(A)|$ takes on one of the values that we have already computed because $f(1) = f(2) = g(1) = g(2) = 1/4$.

In particular, for any set that includes 3 but does not include 4, $|\mathbf{P}(A) - \mathbf{Q}(A)| = |-1/8| = 1/8$. For any set that includes 4 but does not include 3, $|\mathbf{P}(A) - \mathbf{Q}(A)| = |1/8| = 1/8$. And finally, for any set that includes both 3 and 4, $|\mathbf{P}(A) - \mathbf{Q}(A)| = 0$.

Therefore, $\max_{A \subset E} |\mathbf{P}(A) - \mathbf{Q}(A)| = 1/8$, with the maximum achieved with numerous sets as discussed above.

Equivalence of Formulas (cont.)

1/1 point (graded)

Setup as above:
Let $E = \{1, 2, 3, 4\}$ be a discrete sample space. Let $\mathbf{P}$ and $\mathbf{Q}$ be probability measures with probability mass functions f and g as follows:

$$f(1) = 1/4,\ f(2) = 1/4,\ f(3) = 1/8,\ f(4) = 3/8$$

$$g(1) = g(2) = g(3) = g(4) = 1/4$$

Question: What is the value of $\frac{1}{2}\sum_{x \in E} |f(x) - g(x)|$?

1/8 ✔ **Answer:** 1/8

Solution:

$$\frac{1}{2}\sum_{x \in E} |f(x) - g(x)| = \frac{1}{2}\left(0 + 0 + \frac{1}{8} + \frac{1}{8}\right) = \frac{1}{8}.$$

This is the same result as in the previous problem.

Computing Total Variation Distance I

1/1 point (graded)

Let $X \sim \mathbf{P} = \text{Ber}(1/2)$ and $Y \sim \mathbf{Q} = \text{Ber}(1/2)$. What is $\text{TV}(\mathbf{P}, \mathbf{Q})$, the total variation distance between the distributions of the Bernoulli random variables X and Y?

Note that we make no assumptions about X and Y being independent.

0 ✔ **Answer:** 0.0

Solution:

Intuitively, since X and Y have the same distribution, we expect the (total variation) distance between their distributions to be 0. And indeed this is the case. Observe that for any event, $\mathbf{P}(A) = \mathbf{Q}(A)$ since $\mathbf{P}$ and $\mathbf{Q}$ are both $\text{Ber}(1/2)$.

$$\text{TV}(\mathbf{P}, \mathbf{Q}) = \max_{A \subset E} |\mathbf{P}(A) - \mathbf{Q}(A)| = 0.$$

Note that the distance between two distributions only depends on the distributions themselves and *not* their relation to each other (the joint distribution). This is why assuming X and Y are independent (or not) does not affect the total variation distance.

Computing Total Variation II

1/1 point (graded)

Let $X \sim \mathbf{P} = \text{Ber}(1/2)$ and $Y \sim \mathbf{Q} = \text{Ber}(1/3)$. What is $\text{TV}(\mathbf{P}, \mathbf{Q})$, the total variation distance between the distributions of the Bernoulli random variables X and Y?

1/6

✔ **Answer:** 1/6

Solution:

For this problem, the sample space of X and Y is $\{0, 1\}$. Let f be the pmf of X and let g be the pmf of Y. Note that $f(1) = f(0) = 1/2$ and $g(1) = 1/3$, $g(0) = 2/3$. Hence, we can apply the given formula:

$$\begin{aligned} \text{TV}(\mathbf{P}, \mathbf{Q}) &= \frac{1}{2} \sum_{x \in E} |f(x) - g(x)| \\ &= \frac{1}{2}(|f(0) - g(0)| + |f(1) - g(1)|) \\ &= \frac{1}{2}(1/6 + 1/6) = 1/6 \approx 0.16667. \end{aligned}$$

Remark: In general, we have the formula

$$\text{TV}(\text{Ber}(p), \text{Ber}(q)) = |p - q|.$$

Let $\mathbf{P}$ and $\mathbf{Q}$ be probability distributions on a **continuous** sample space E with probability density functions f and g. Then, the total variation distance between $\mathbf{P}$ and $\mathbf{Q}$

$$\text{TV}(\mathbf{P}, \mathbf{Q}) = \max_{A \subset E} |\mathbf{P}(A) - \mathbf{Q}(A)|,$$

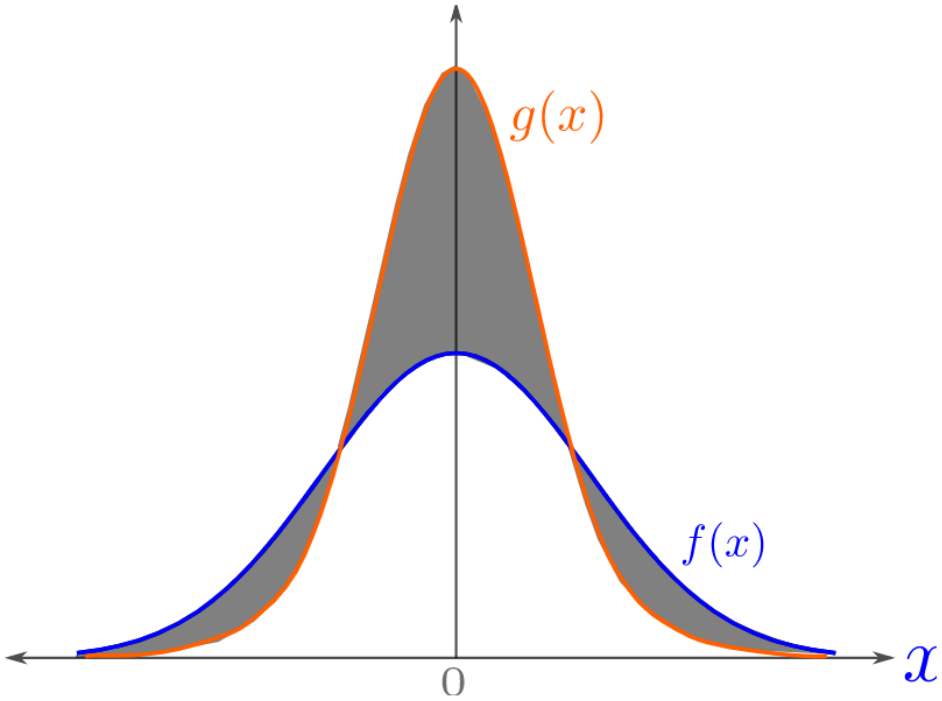
can be computed as

$$\text{TV}(\mathbf{P}, \mathbf{Q}) = \frac{1}{2} \int_{x \in E} |f(x) - g(x)| \, \mathrm{dx}.$$

Graphical Interpretation of Total Variation

1/1 point (graded)

Let $X \sim \mathbf{P}$ and $Y \sim \mathbf{Q}$ be Gaussian random variables with mean 0. Let f denote the probability density function of X and g denote the density of Y. Which answer is a correct graphical interpretation of $2\text{TV}(\mathbf{P}, \mathbf{Q})$, 2 times the total variation distance between $\mathbf{P}$ and $\mathbf{Q}$?



Let d be a function that takes two probability measures $\mathbf{P}$ and $\mathbf{Q}$ and maps them to a real number $d(\mathbf{P}, \mathbf{Q})$. Then d is a **distance** on probability measures if the following four axioms hold. (Here, $\mathbf{P}$, $\mathbf{Q}$, and $\mathbf{V}$ are all probability measures.)

- $d(\mathbf{P}, \mathbf{Q}) = d(\mathbf{Q}, \mathbf{P})$ (symmetric)
- $d(\mathbf{P}, \mathbf{Q}) \geq 0$ (nonnegative)
- $d(\mathbf{P}, \mathbf{Q}) = 0 \iff \mathbf{P} = \mathbf{Q}$ (definite)
- $d(\mathbf{P}, \mathbf{V}) \leq d(\mathbf{P}, \mathbf{Q}) + d(\mathbf{Q}, \mathbf{V})$ (triangle inequality)

In the above, $\mathbf{P} = \mathbf{Q}$ means $\mathbf{P}(A) = \mathbf{Q}(A)$ for $A \subset E$, where E is the common sample space of $\mathbf{P}$ and $\mathbf{Q}$.

The total variation distance (TV) is a distance on probability measures.

Symmetry and Definiteness of Total Variation Distance

1/1 point (graded)

Let $\mathbf{P}$ and $\mathbf{Q}$ be probability measures. Which of the following is (are) true?

- [ ] One can find a measure $\mathbf{Q} \neq \mathbf{P}$ such that $\mathrm{TV}(\mathbf{P}, \mathbf{Q}) = 0$.
- [x] $\mathrm{TV}(\mathbf{P}, \mathbf{Q}) = \mathrm{TV}(\mathbf{Q}, \mathbf{P})$.

✔

Triangle Inequality

1/1 point (graded)

Which of the following quantities is greater than or equal to $\mathrm{TV}\left(\mathrm{Ber}\left(.5\right),\mathrm{Ber}\left(0.3\right)\right)$?
(Choose all that apply.)

- [x] $\mathrm{TV}\left(\mathrm{Ber}\left(0.5\right),\mathrm{Ber}\left(0.1\right)\right)+\mathrm{TV}\left(\mathrm{Ber}\left(0.1\right),\mathrm{Ber}\left(0.3\right)\right)$
- [x] $\mathrm{TV}\left(\mathrm{Ber}\left(0.5\right),\mathrm{Poiss}\left(5\right)\right)+\mathrm{TV}\left(\mathrm{Ber}\left(0.3\right),\mathrm{Poiss}\left(5\right)\right)$
- [x] $\mathrm{TV}\left(\mathrm{Bin}\left(7,0.4\right),\mathrm{Ber}\left(0.5\right)\right)+\mathrm{TV}\left(\mathrm{Ber}\left(0.3\right),\mathrm{Bin}\left(7,0.4\right)\right)$

✔

Solution:

Recall the triangle inequality states that for distributions $\mathbf{P}$, $\mathbf{Q}$, and $\mathbf{V}$:

$$\mathrm{TV}\left(\mathbf{P},\mathbf{V}\right)\leq\mathrm{TV}\left(\mathbf{P},\mathbf{Q}\right)+\mathrm{TV}\left(\mathbf{Q},\mathbf{V}\right).$$

- If we set $\mathbf{P}=\mathrm{Ber}\left(0.5\right)$, $\mathbf{V}=\mathrm{Ber}\left(0.3\right)$, and $\mathbf{Q}=\mathrm{Ber}\left(0.1\right)$, then applying the triangle inequality above gives the first upper bound.
- In the second choice, set $\mathbf{P}=\mathrm{Ber}\left(0.5\right)$, $\mathbf{V}=\mathrm{Ber}\left(0.3\right)$, and $\mathbf{Q}=\mathrm{Poiss}\left(5\right)$ and apply the triangle inequality.
- In the third choice, set $\mathbf{P}=\mathrm{Ber}\left(0.5\right)$, $\mathbf{V}=\mathrm{Ber}\left(0.3\right)$, and $\mathbf{Q}=\mathrm{Bin}\left(7,0.4\right)$ and apply the triangle inequality.

Remark: Implicitly we are also using the symmetry property of total variation: $\mathrm{TV}\left(\mathbf{P},\mathbf{Q}\right)=\mathrm{TV}\left(\mathbf{Q},\mathbf{P}\right)$.

Concept Check: Upper Bound on TV

1 point possible (graded)

Give the smallest number M such that $\mathrm{TV}\left(\mathbf{P},\mathbf{Q}\right)\leq M$ for **any** probability measures $\mathbf{P},\mathbf{Q}$.

$M=$ [ ] **Answer:** 1

(Find a pair of distributions $\mathbf{P},\mathbf{Q}$ such that $\mathrm{TV}\left(\mathbf{P},\mathbf{Q}\right)=M$.)

Solution:

Using the definition of total variation distance,

$$\mathrm{TV}\left(\mathbf{P},\mathbf{Q}\right)=\max_{A\subset E}|\mathbf{P}\left(A\right)-\mathbf{Q}\left(A\right)|,$$

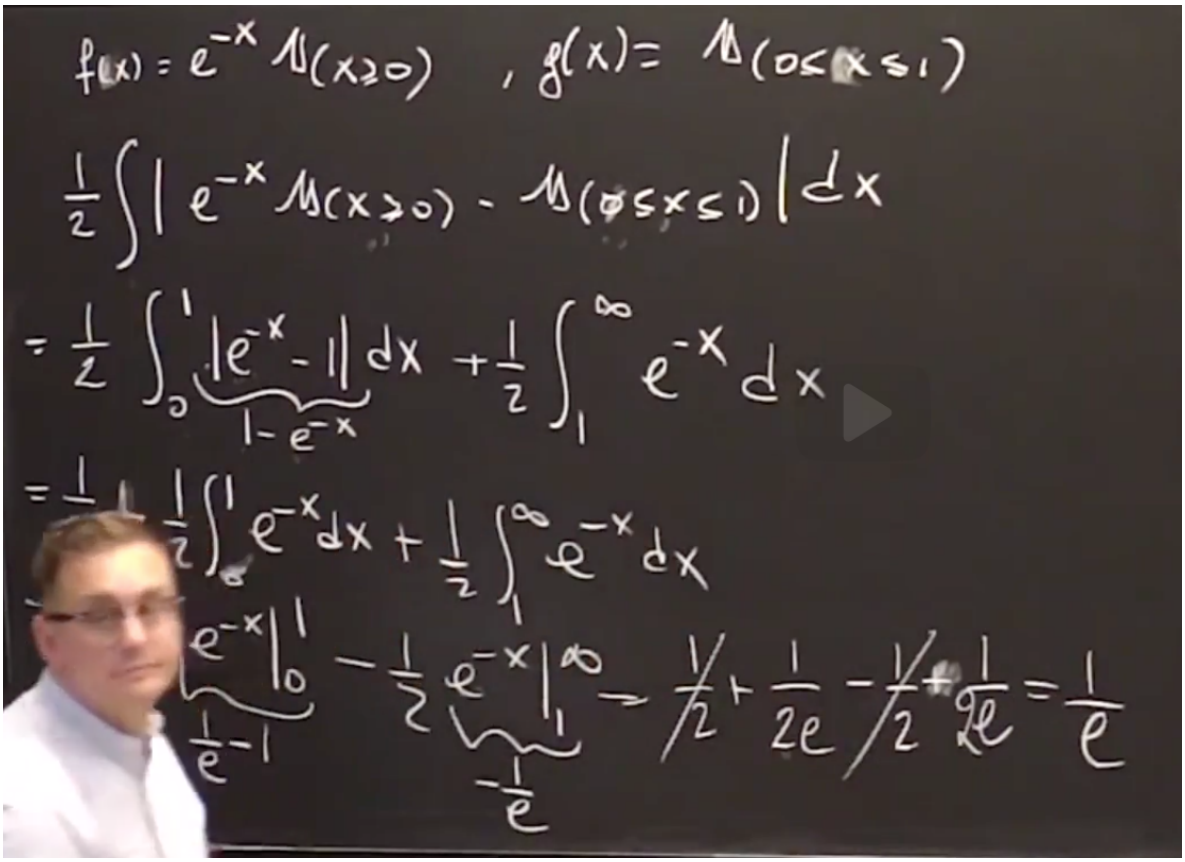
we can say that if the maximum is obtained using a set A_1 such that $\mathbf{P}\left(A_1\right)\geq\mathbf{Q}\left(A_1\right)$, then

$$\mathrm{TV}\left(\mathbf{P},\mathbf{Q}\right)=|\mathbf{P}\left(A_1\right)-\mathbf{Q}\left(A_1\right)|\leq\mathbf{P}\left(A_1\right)\leq 1.$$

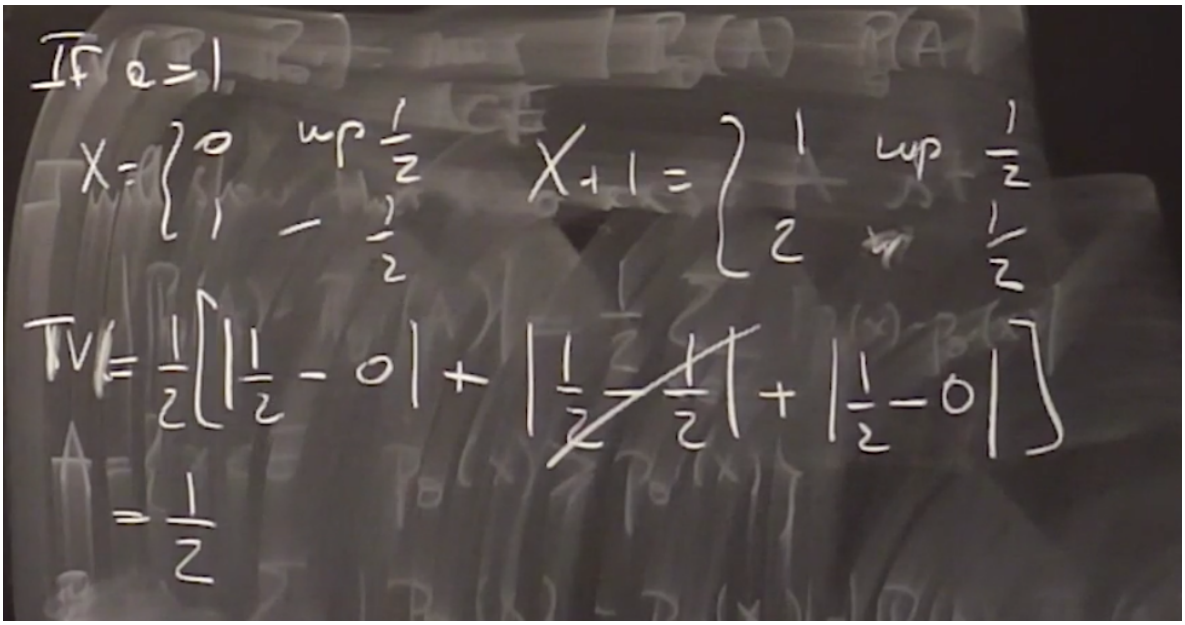
A similar argument can be made for the case when the the maximum is obtained using a set A_2 such that $\mathbf{Q}\left(A_2\right)>\mathbf{P}\left(A_2\right)$.

An example pair $\mathbf{P},\mathbf{Q}$ where the bound is met with equality: $E=\{1,2\},\mathbf{P}\left(1\right)=1,\mathbf{Q}\left(2\right)=1$.
Remark:. In general, when the support of $\mathbf{P}$ does not intersect with the support $\mathbf{Q}$, then $\mathrm{TV}\left(\mathbf{P},\mathbf{Q}\right)=1$.

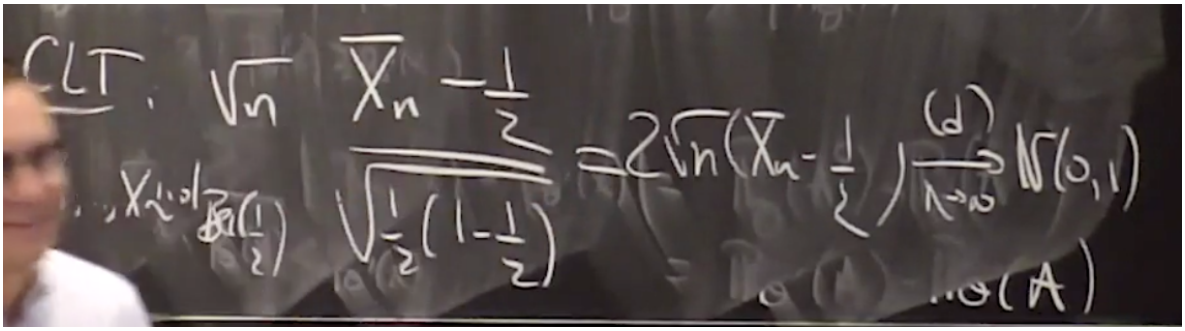
slide 7 (c)



slide 7 (d) where a=1, example



(e)



distance (TV) should become smaller as n-> inf

Computing Total Variation IV

1 point possible (graded)

So far, we have defined the total variation distance to be a distance $\mathrm{TV}(\mathbf{P}, \mathbf{Q})$ between **two probability measures** $\mathbf{P}$ and $\mathbf{Q}$. However, we will also refer to the total variation distance between **two random variables** or between **two pdfs** or **two pmfs**, as in the following.

Compute $\mathrm{TV}(X, X+a)$ for any $a \in (0,1)$, where $X \sim \mathsf{Ber}(0.5)$.

$\mathrm{TV}(X, X+a) =$ [ ] **Answer:** 1

Solution:

Since $a \in (0,1)$, X and $X+a$ have no support points where both pmf's are non-zero. Therefore, the total variation distance is equal to 1.

Computing Total Variation V

1 point possible (graded)

Compute $\mathrm{TV}\left(2\sqrt{n}\left(\bar{X}_n - 1/2\right), Z\right)$ where $X_i \overset{i.i.d}{\sim} \mathsf{Ber}(0.5)$ and $Z \sim \mathcal{N}(0,1)$.

$\mathrm{TV}\left(2\sqrt{n}\left(\bar{X}_n - 1/2\right), Z\right) =$ [ ] **Answer:** 1

Solution:

Let $\mathbf{P}$ and $\mathbf{Q}$ denote the probability measures of $2\sqrt{n}\left(\bar{X}_n - 1/2\right)$ and Z, respectively. Recall the total variation distance is defined as

$$\max_{A \subset E} |\mathbf{P}(A) - \mathbf{Q}(A)|$$

Let $B \triangleq \left\{a_i = 2\sqrt{n}\left(\frac{i}{n} - \frac{1}{2}\right) \mid i = 0, 1, \ldots, n\right\}$ be set of $n+1$ points where the pmf of $2\sqrt{n}\left(\bar{X}_n - 1/2\right)$ is non-zero.

Consider the set $A = \mathbb{R} \setminus B \left(= R \cap B^c\right)$. For this set, $\mathbf{P}(A) = 0$ and $\mathbf{Q}(A) = 1$. Therefore, $|\mathbf{P}(A) - \mathbf{Q}(A)| = 1$. We know from a previous problem that the total variation distance is upper bounded by 1 for any two distributions. Since we have produced a set where this bound is met with equality, $\mathrm{TV}\left(2\sqrt{n}\left(\bar{X}_n - 1/2\right), Z\right) = 1$.

Definition of Kullback-Leibler (KL) Divergence

Let $\mathbf{P}$ and $\mathbf{Q}$ be **discrete** probability distributions with pmfs p and q respectively. Let's also assume $\mathbf{P}$ and $\mathbf{Q}$ have a common sample space E. Then the **KL divergence** (also known as **relative entropy**) between $\mathbf{P}$ and $\mathbf{Q}$ is defined by

$$\mathrm{KL}(\mathbf{P}, \mathbf{Q}) = \sum_{x \in E} p(x) \ln\left(\frac{p(x)}{q(x)}\right),$$

where the sum is only over the support of $\mathbf{P}$.

Why do we sum only over the support of P?

We use the following limit to justify the definition above. At any point $x \in E$ outside the support of $\mathbf{P}$ but where $q(x) \neq 0$:

$$\begin{aligned} \lim_{p/q \to 0^+} q\left(\frac{p}{q}\right) \ln\left(\frac{p}{q}\right) &= q \lim_{p/q \to 0^+} \left(\frac{p}{q}\right) \ln\left(\frac{p}{q}\right) \\ &= q \cdot (0) = 0 \qquad \text{(by L'hopital's rule)}. \end{aligned}$$

Hide

Analogously, if $\mathbf{P}$ and $\mathbf{Q}$ are **continuous** probability distributions with pdfs p and q on a common sample space E, then

$$\mathrm{KL}(\mathbf{P}, \mathbf{Q}) = \int_{x \in E} p(x) \ln\left(\frac{p(x)}{q(x)}\right) dx,$$

where the integral is again only over the support of $\mathbf{P}$.

Computing KL Divergence I

1 point possible (graded)

Let $X \sim \mathbf{P}_X = \mathrm{Ber}(1/2)$ and let $Y \sim \mathbf{P}_Y = \mathrm{Ber}(1/2)$. What is $\mathrm{KL}(\mathbf{P}_X, \mathbf{P}_Y)$?

$\mathrm{KL}(\mathbf{P}_X, \mathbf{P}_Y) =$ **Answer:** 0.0

Solution:

Let p be the pmf of the distribution $\mathrm{Ber}(1/2)$. Note that the sample space is the discrete set $E = \{0, 1\}$. Then

$$\begin{aligned} \mathrm{KL}(\mathbf{P}_X, \mathbf{P}_Y) &= p(1) \ln(p(1)/p(1)) + p(0) \ln(p(0)/p(0)) \\ &= (1/2) \ln(1) + (1/2) \ln(1) = 0. \end{aligned}$$

Remark: Although KL divergence is not a distance on probability distributions (as we defined above), it does satisfy some of the axioms. For example,

- $\mathrm{KL}(\mathbf{P}, \mathbf{Q}) \geq 0$ (nonnegative), and
- $\mathrm{KL}(\mathbf{P}, \mathbf{Q}) = 0$ only if $\mathbf{P}$ and $\mathbf{Q}$ are the same distribution (definite).

Note that the result of this problem is consistent with the second property.

Computing KL Divergence II

3 points possible (graded)

Let $X \sim \mathbf{P}_X = \text{Ber}(1/2)$ and let $Y \sim \mathbf{P}_Y = \text{Ber}(1/3)$. What is $\text{KL}(\mathbf{P}_X, \mathbf{P}_Y)$?

(If applicable, enter **ln(x)** for $\ln(x)$.)

$\text{KL}(\mathbf{P}_X, \mathbf{P}_Y) =$ **Answer:** 0.0588915

What is $\text{KL}(\mathbf{P}_Y, \mathbf{P}_X)$?

$\text{KL}(\mathbf{P}_Y, \mathbf{P}_X) =$ **Answer:** 0.05663301

Is $\text{KL}(\mathbf{P}_X, \mathbf{P}_Y) = \text{KL}(\mathbf{P}_Y, \mathbf{P}_X)$?

- ◯ Yes
- ◯ No ✔

Solution:

Let f and g denote the pmfs of $\text{Ber}(1/2)$ and $\text{Ber}(1/3)$, respectively. Note that the sample space is $E = \{0, 1\}$. Then

$$\begin{aligned}\text{KL}(\mathbf{P}_X, \mathbf{P}_Y) &= \sum_{x \in \{0,1\}} f(x) \ln(f(x)/g(x)) \\ &= (1/2)\ln(3/2) + (1/2)\ln(3/4) \approx 0.0588915\end{aligned}$$

Next,

$$\begin{aligned}\text{KL}(\mathbf{P}_Y, \mathbf{P}_X) &= \sum_{x \in \{0,1\}} g(x) \ln(g(x)/f(x)) \\ &= (1/3)\ln(2/3) + (2/3)\ln(4/3) \approx 0.05663301\end{aligned}$$

Remark: In general, we have the formula

$$\text{KL}(\text{Ber}(p), \text{Ber}(q)) = p \ln\left(\frac{p}{q}\right) + (1-p)\ln\left(\frac{1-p}{1-q}\right).$$

Properties of KL Divergence I

2 points possible (graded)

Let $\mathbf{P}$ be a distribution such that $\mathrm{KL}\left(\mathrm{Ber}\left(1/2\right), \mathbf{P}\right) = 0$. What can we conclude about $\mathbf{P}$?

- ◯ $\mathbf{P} = \mathrm{Ber}\left(1/2\right)$. ✔
- ◯ It is possible that $\mathbf{P} = \mathrm{Ber}\left(p\right)$ for any $0 \leq p \leq 1$.
- ◯ $\mathbf{P}$ could be any Gaussian distribution with mean 0 and variance $1/4$.
- ◯ None of the above.

What property of the KL divergence did you use to make your conclusion?

- ◯ Symmetric
- ◯ Nonnegative
- ◯ Definite ✔
- ◯ Triangle Inequality

Solution:

The definite property of the KL divergence implies that if $\mathrm{KL}\left(\mathbf{P}, \mathbf{Q}\right) = 0$, then $\mathbf{P}$ and $\mathbf{Q}$ are the same distribution. Hence, we use this property to conclude that $\mathbf{P} = \mathrm{Ber}\left(1/2\right)$.

Note that while the KL divergence is nonnegative and definite, it is not a distance because it does not satisfy the triangle inequality nor is it symmetric.

Concept check: Properties of KL Divergence

1 point possible (graded)

Which of the following are properties of the **Kullback-Leibler KL divergence**?
(Choose all that apply.)

- ☐ The KL divergence is symmetric, *i.e.*, $\mathrm{KL}\left(\mathbf{P}, \mathbf{Q}\right) = \mathrm{KL}\left(Q, P\right)$ for all distributions P, Q.
- ☐ The KL divergence is, strictly speaking, a distance function between probability distributions.
- ☐ The KL divergence $\mathrm{KL}\left(P_{\theta^*}, P_{\theta}\right)$ can be written as an expectation with respect to the distribution P_{θ^*}. ✔
- ☐ In general, it is easier to build an estimator for the KL divergence than it is to build an estimator for the total variation distance. ✔

Solution:

- The first choice is incorrect. The second problem in this section shows that the KL divergence is not symmetric.
- The second choice is also incorrect. A distance function, strictly, speaking must be symmetric and satisfy the triangle inequality. The KL divergence is not symmetric and does not satisfy the triangle inequality in general, so it is not a proper distance.
- The third choice is correct. Suppose that the distributions P_θ and P_{θ^*} are discrete and have pmfs p_θ and p_{θ^*}, respectively. Then

$$\mathrm{KL}\left(P_{\theta^*}, P_\theta\right) = \sum_{x \in E} p_{\theta^*}(x) \ln\left(\frac{p_{\theta^*}(x)}{p_\theta(x)}\right) = \mathbb{E}_{\theta^*}\left[\ln\left(\frac{p_{\theta^*}(X)}{p_\theta(X)}\right)\right].$$

  Notation: Here we use the notation $\mathbb{E}_{\theta^*}$ to denote the expectation with respect to the distribution P_{θ^*}.

- The fourth choice is also correct. The total variation distance is not an expectation with respect to either of the probability measures. Therefore there is no natural way to estimate it without requiring an estimate of the true parameter θ^*. The KL divergence, by contrast, is an expectation of some function with respect to one of the probability measures. This means that it can be estimated naturally by replacing the expectation with a sample average. Note that this application of the law of large numbers does not require knowledge of the true parameter value, only a random sample generated from the true distribution.

The next four problems concern the following statistical set-up.

You observe discrete random variables

$$X_1, \ldots, X_n \overset{iid}{\sim} P_{\theta^*}$$

where θ^* is the true parameter. You construct an associated statistical model $(E, \{P_\theta\}_{\theta \in \mathbb{R}})$ with a discrete sample space E.

Your goal is to find an estimator $\hat{\theta}_n$ so that the distributions $P_{\hat{\theta}_n}$ and P_{θ^*} are close. More precisely, you want to find an estimator $\hat{\theta}_n$ so that the quantity

$$\mathrm{KL}\left(P_{\theta^*}, P_{\hat{\theta}_n}\right) = \sum_{x \in E} p_{\theta^*} \ln \frac{p_{\theta^*}(x)}{p_{\hat{\theta}_n}(x)}$$

is as small as possible.

This approach will naturally lead to the construction of the **maximum likelihood estimator** .

Consider the optimization problem in which we minimize the KL divergence between P_{θ^*}, the true distribution, and P_θ. Formally, we want to solve

$$\min_{\theta \in \mathbb{R}} \mathrm{KL}\left(P_{\theta^*}, P_\theta\right).$$

We are not so much interested in the minimum value attained by the objective function $\mathrm{KL}\left(P_{\theta^*}, P_\theta\right)$, but rather the value of θ where the minimum is attained. We refer to such a θ as a **minimizer** .

Let's suppose that there is a unique minimizer for the above optimization problem– *i.e.*, if m is the minimum value of $\mathrm{KL}\left(P_{\theta^*}, P_\theta\right)$, there is only one point $\theta_{\min}$ such that

$$m = \mathrm{KL}\left(P_{\theta^*}, P_{\theta_{\min}}\right).$$

For which θ is the minimum value of $\mathrm{KL}\left(P_{\theta^*}, P_\theta\right)$ attained? (Equivalently, what is $\theta_{\min}$?)

- ◯ θ^* ✔
- ◯ θ
- ◯ 0
- ◯ None of the above.

Solution:

The KL divergence is nonnegative, so $\mathrm{KL}\left(P_{\theta^*}, P_\theta\right) \geq 0$. The right-hand side is achieved if we set $\theta = \theta^*$: $\mathrm{KL}\left(P_{\theta^*}, P_{\theta^*}\right) = 0$. Since the minimizer is unique by assumption, we conclude that the minimum value is attained at $\theta = \theta^*$.

Can we Minimize KL Divergence Directly?

2 points possible (graded)

Let's use the same statistical set-up as above. Recall that you have access to the iid samples $X_1, \ldots, X_n$. You use these samples to build an estimator $\hat{\theta}_n$. Can you compute

$$\mathrm{KL}\left(P_{\hat{\theta}_n}, P_{1/2}\right)$$

without knowing θ^*, the true parameter?

- ◯ Yes ✔
- ◯ No

Can you compute

$$\mathrm{KL}\left(P_{\theta^*}, P_{1/2}\right)$$

without knowing θ^*?

- ◯ Yes
- ◯ No ✔

Solution:

In general, we can compute $\mathrm{KL}(\mathbf{P}, \mathbf{Q})$ if and only if we know both distributions $\mathbf{P}$ and Q. Moreover, by our statistical model, we can compute P_θ if and only if we know the real number θ. Putting these last two facts together, we can compute

$$\mathrm{KL}\left(P_{\hat{\theta}_n}, P_{1/2}\right)$$

because $\hat{\theta}_n$ is known–it is an estimator so its expression does not depend on θ^*, the true parameter. However, regardless of how many samples we take, we cannot compute $\mathrm{KL}\left(P_{\theta^*}, P_{1/2}\right)$ exactly because the distribution P_{θ^*} is unknown.

Remark: Since we cannot even compute the function $\mathrm{KL}\left(P_{\theta^*}, P_\theta\right)$ for general θ, this implies that the optimization problem

$$\min_{\theta \in \mathbb{R}} \mathrm{KL}\left(P_{\theta^*}, P_\theta\right)$$

cannot be solved exactly, regardless of the number of samples we have. So to estimate the minimizer of this optimization problem (which is the true parameter θ^*) we will have to consider an approximation for $\mathrm{KL}\left(P_{\theta^*}, P_\theta\right)$.

Finding the Minimizer for an Approximation of KL Divergence

1 point possible (graded)

We use the same statistical set-up as above. Recall that $X_1, \ldots, X_n \overset{iid}{\sim} P_{\theta^*}$. Let p_θ be the pmf of P_θ.

Which of the following is a (weakly) **consistent** estimator for

$$\mathbb{E}_{\theta^*}\left[\ln p_\theta(X)\right] = \sum_{x \in E} p_{\theta^*} \ln p_\theta(x) \ ?$$

- ○ $\frac{1}{n}\sum_{i=1}^{n} X_i$
- ○ $\frac{1}{n}\sum_{i=1}^{n} \ln\left(p_\theta(X_i)\right)$ ✔
- ○ $\frac{1}{n}\sum_{i=1}^{n} \ln\left(p_{\theta^*}(X_i)\right) - \frac{1}{n}\sum_{i=1}^{n} \ln\left(p_\theta(X_i)\right)$
- ○ $\theta^* - \mathbb{E}_{\theta^*}\left[\ln p_{\theta^*}\right]$

Solution:

By the law of large numbers, $\frac{1}{n}\sum_{i=1}^{n} \ln\left(p_\theta\left(X_i\right)\right) \to \mathbb{E}_{\theta^*}\left[\ln p_\theta\right]$ in probability. Hence, the second choice is correct.

Remark 1: The KL divergence between P_{θ^*} and P_θ can be written

$$\mathrm{KL}\left(P_{\theta^*}, P_\theta\right) = \sum_{x\in E} p_{\theta^*} \ln p_{\theta^*}(x) - \sum_{x\in E} p_{\theta^*} \ln p_\theta(x) = \mathbb{E}_{\theta^*}\left[\ln p_{\theta^*}(X)\right] - \mathbb{E}_{\theta^*}\left[\ln p_\theta(X)\right]$$

where $X \sim P_{\theta^*}$.

Remark 2: While we can't find θ that minimizes $\mathrm{KL}\left(P_{\theta^*}, P_\theta\right)$, we can find θ that minimizes

$$\hat{\mathrm{KL}}\left(P_{\theta^*}, P_\theta\right) := \mathbb{E}_{\theta^*}\left[\ln p_{\theta^*}\right] - \frac{1}{n}\sum_{i=1}^{n} \ln\left(p_\theta\left(X_i\right)\right).$$

Here's why: the first term on the RHS, $\mathbb{E}_{\theta^*}\left[\ln p_{\theta^*}\right]$, does not depend on θ. Hence, the θ that minimizes $\hat{\mathrm{KL}}\left(P_{\theta^*}, P_\theta\right)$ is the same as the θ that minimizes $-\frac{1}{n}\sum_{i=1}^{n} \ln\left(p_\theta\left(X_i\right)\right)$.

Deriving the Maximum Likelihood Estimator

1 point possible (graded)

We use the same statistical set-up as above. Recall that p_θ is the pmf of P_θ and $X_1, \ldots, X_n \overset{iid}{\sim} P_{\theta^*}$.

Suppose that $\theta_{\min}$ is a minimizer for the function

$$f(\theta) := -\frac{1}{n}\sum_{i=1}^{n} \ln\left(p_\theta\left(X_i\right)\right)$$

Which of the following functions is also minimized at $\theta_{\min}$?

- $g_1(\theta) = -\prod_{i=1}^{n} p_\theta\left(X_i\right)$
- $g_2(\theta) = 25 - \prod_{i=1}^{n} p_\theta\left(X_i\right)$
- $g_3(\theta) = h(\theta^*) - \prod_{i=1}^{n} p_\theta\left(X_i\right)$ where h is a function of θ^* that does **not** depend on θ.
- $g_4(\theta) = \theta^* - \prod_{i=1}^{n} p_\theta\left(X_i\right)$
- All of the above ✔

Solution:

Observe that rescaling by n does not change where the minimum of a function is attained. Hence, $f(\theta)$ and $nf(\theta)$ have the same minimizer. Next, by the addition property of logarithms,

$$nf(\theta) = \sum_{i=1}^{n} \ln\left(p_\theta(X_i)\right) = \ln\left(\prod_{i=1}^{n} p_\theta(X_i)\right).$$

Since $\ln$ is an increasing function, the function

$$\theta \mapsto \prod_{i=1}^{n} p_\theta(X_i)$$

has the same minimizer as $\ln\left(\prod_{i=1}^{n} p_\theta(X_i)\right)$. Thus the first choice is correct.

Moreover, the second and third choices are also correct. Whenever we have an optimization problem

$$\min_{\theta \in \mathbb{R}} C + g(\theta)$$

where C does not depend on θ, then the above will have the same minimizer as the optimization problem

$$\min_{\theta \in \mathbb{R}} g(\theta).$$

In the second choice, $C = 25$ (which is independent of θ), and in the third choice, $C = h(\theta^*)$ (which by assumption is independent of θ).

Remark 1: The quantity

$$\hat{\theta}_n := \text{maximizer of } \prod_{i=1}^{n} p_\theta(X_i)$$

is referred to as the **maximum likelihood estimator** . Note that this is the same as the estimator

$$\hat{\theta}_n := \text{minimizer of } -\frac{1}{n}\sum_{i=1}^{n} \ln\left(p_\theta(X_i)\right)$$

considered in Remark 2 in the solution of the previous problem.

Remark 2: Under certain technical conditions, the maximum likelihood estimator is guaranteed to (weakly) converge to the true parameter θ^*.

14. Likelihood of a Discrete Distribution

Exercises due Jun 24, 2020 08:59 JST *Past Due*

Bookmark this page

Preparation: Equivalent Expressions for the pmf of a Bernoulli Distribution

1 point possible (graded)

Which of the following function $f(x)$, when restricted to the domain $x \in \{0, 1\}$, is equal to the pmf f of the probability distribution $\text{Ber}(p)$? Assume that $p \in (0, 1)$. (Choose all that apply.) (Recall that if $X \sim \text{Ber}(p)$, then $p = P(X = 1)$.)

- [x] $f(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$ ✔
- [x] $f(x) = p^x(1 - p)^{1-x}$ ✔
- [x] $f(x) = xp + (1 - x)(1 - p)$ ✔
- [ ] $f(x) = \begin{cases} 1 & \text{if } x = 1 \\ 0 & \text{if } x = 0 \end{cases}$

Solution:

We will explain in the order of the choices.

- $f(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$ is correct. A random variable $X \sim \text{Ber}(p)$, by definition, has sample space $\{0, 1\}$ and satisfies $P(X = 1) = p$ and $P(X = 0) = 1 - p$. The given function is just a restatement of that definition.
- $f(x) = p^x(1 - p)^{1-x}$ is correct. Note that $f(1) = p$ and $f(0) = 1 - p$, so this is the same as the function considered in the first choice.
- $f(x) = xp + (1 - x)(1 - p)$ is correct. It also satisfies $f(1) = p$ and $f(0) = 1 - p$, so f is the same as the function considered in the first choice.
- $f(x) = \begin{cases} 1 & \text{if } x = 1 \\ 0 & \text{if } x = 0 \end{cases}$ is incorrect. This is actually the probability mass function of $\text{Ber}(1)$, but we have assumed $p \in (0, 1)$.

Review: Statistical Model for a Bernoulli Distribution

3 points possible (graded)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Ber}(p^*)$ for some unknown $p^* \in (0, 1)$. Let $(E, \{\text{Ber}(p)\}_{p\in\Theta})$ denote the corresponding statistical model. What is the smallest possible set that could be E?

- ◯ $\{0\}$
- ◯ $\{-1, 1\}$
- ◯ $\{0, 1\}$ ✔
- ◯ $\mathbb{R}$

The parameter space Θ can be written as an interval $[a, b]$. What is the smallest possible interval so that $\{\text{Ber}(p)\}_{p\in\Theta}$ represents all possible Bernoulli distributions?

$a =$ [ ] **Answer:** 0.0

$b =$ [ ] **Answer:** 1.0

Solution:

Since a Bernoulli random variable is either 0 or 1, the smallest possible sample space is $\{0, 1\}$.

If $\Theta = [0, 1]$, then $\{\text{Ber}(p)\}_{p\in[0,1]}$ is the set of all possible Bernoulli distributions, as desired.

Concept Check: Interpreting the Likelihood

1 point possible (graded)

Let $(E, \{P_\theta\}_{\theta\in\Theta})$ denote a discrete statistical model. Let p_θ denote the pmf of P_θ. Let $X_1, \ldots, X_n \overset{iid}{\sim} P_{\theta^*}$ where the parameter θ^* is unknown. Then the **likelihood** is the function

$$\begin{aligned} L_n : E^n \times \Theta &\to \mathbb{R} \\ (x_1, \ldots, x_n, \theta) &\mapsto \prod_{i=1}^{n} p_\theta(x_i). \end{aligned}$$

For our purposes, we think of $x_1, \ldots, x_n$ as observations of the random variables $X_1, \ldots, X_n$.

Which of the following are true about the likelihood L_n? (Choose all that apply.)

- ☐ It is the joint pmf of n i.i.d. samples from the distribution P_θ. ✔
- ☐ For a fixed θ, it is a function of the sample $X_1 = x_1, \ldots, X_n = x_n$. ✔
- ☐ For a fixed sample $X_1 = x_1, \ldots, X_n = x_n$, it is a function of the parameter θ, where θ ranges over all possible values in the parameter space Θ. ✔
- ☐ It is the joint pmf of n iid samples from the true distribution P_{θ^*}.

Solution:

We examine the choices in order.

- "It is the joint pmf of n iid samples from the distribution P_θ." is correct. If $Y_1, \ldots, Y_n \overset{iid}{\sim} P_\theta$, then by independence, the joint pmf of these variables is given by a product:

$$P(Y_1 = x_1, \ldots, Y_n = x_n) = \prod_{i=1}^{n} p_\theta(x_i).$$

  Remark 1: We use Y_i to denote these variables to differentiate from the samples X_i that come from the true distribution P_{θ^*}.

- "It is a function of the sample $X_1 = x_1, \ldots, X_n = x_n$." is correct. To construct the likelihood, we observe samples $X_1 = x_1, \ldots, X_n = x_n$ and then compute $L_n(x_1, \ldots, x_n, \theta)$.

- "It is a function of the parameter θ, where θ ranges over all possible values of in the parameter space Θ" is correct. As θ varies over Θ, the likelihood $L_n(x_1, \ldots, x_n, \theta)$ takes on different values. This is evident from the dependence on θ in the definition of the likelihood.

  Remark 2: Later on we will maximize L_n (as a function of θ) to define the **maximum likelihood estimator** . Hence, it is a crucial property that the likelihood is a function of the parameter.

- "It is the joint pmf of n iid samples from the distribution P_{θ^*}." is incorrect. The likelihood takes as input all possible θ, not just the true parameter θ^*. Note how the likelihood is defined for general θ, not just the true parameter θ^*.

Likelihood of a Bernoulli Statistical Model

1 point possible (graded)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Ber}(p^*)$ for some unknown $p^* \in (0, 1)$. Let $(E, \{\text{Ber}(p)\}_{p \in \Theta})$ denote the corresponding statistical model constructed in the previous question.

What is the likelihood L_n of this statistical model? (Choose all that apply.)

Hint: Use the pmf's in the second and third choices from the first problem on this page: "Preparation Equivalent Expressions for the pmf of a Bernoulli Distribution".

- [x] $L_n(x_1, \ldots, x_n, p) = p^{\sum_{i=1}^n x_i}(1-p)^{n - \sum_{i=1}^n x_i}$. ✔
- [ ] $L_n(x_1, \ldots, x_n, p) = p^{n - \sum_{i=1}^n x_i}(1-p)^{\sum_{i=1}^n x_i}$.
- [ ] $L_n(x_1, \ldots, x_n, p) = p^{\sum_{i=1}^n x_i}(1-p)^{\sum_{i=1}^n x_i}$.
- [x] $L_n(x_1, \ldots, x_n, p) = \prod_{i=1}^n (x_i p + (1 - x_i)(1-p))$ ✔

Solution:

We examine the choices in order.

- As shown in the previous problem, we can write the pmf of a Bernoulli as $x \mapsto p^x(1-p)^{1-x}$. Hence,

$$L_n(x_1, \ldots, x_n, p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

$$\begin{aligned} L(x_1, \ldots, x_n, p) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\ &= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}. \end{aligned}$$

  Hence the first answer choice is correct.

- The second and third choices $L_n(x_1, \ldots, x_n, p) = p^{n-\sum_{i=1}^n x_i}(1-p)^{\sum_{i=1}^n x_i}$ and $L_n(x_1, \ldots, x_n, p) = p^{\sum_{i=1}^n x_i}(1-p)^{\sum_{i=1}^n x_i}$ are incorrect. Note that they are slight algebraic modifications of the first choice, so these formulas cannot be correct.

- If we use the expression $f(x) = xp + (1-x)(1-p)$ for the pmf of $\mathrm{Ber}(p)$, then

$$L(x_1, \ldots, x_n, p) = \prod_{i=1}^{n} (x_i p + (1-x_i)(1-p))$$

  is, by definition, the likelihood. Hence, the last answer choice is also correct.

  Remark: Although the last answer choice is formally correct, the formula is much more difficult to work with. It is often more convenient to use $p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}$ for the likelihood of a Bernoulli statistical model.

Review: Statistical Model for a Poisson Distribution

2 points possible (graded)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathrm{Poiss}(\lambda^*)$ for some unknown $\lambda^* \in (0, \infty)$. Let $(E, \{\mathrm{Poiss}(\lambda)\}_{\lambda \in \Theta})$ denote the corresponding statistical model. What is the smallest possible set that could be E?

- ◯ $\mathbb{N} = \{1, 2, 3, \ldots\}$
- ◯ $\mathbb{N} \cup \{0\}$ ✔
- ◯ $\mathbb{Z}$
- ◯ $\mathbb{R}$

The parameter space Θ can be written as an interval (a, ∞). What is the smallest value of a so that $\{\mathrm{Poiss}(\lambda)\}_{\lambda \in (a, \infty)}$ represents all possible Poisson distributions?

$a =$ [ ] **Answer:** 0.0

Solution:

A Poisson random variable takes values on all non-negative integers $\{0, 1, 2, \ldots\}$. Hence, the smallest possible sample space is $\mathbb{N} \cup \{0\}$.

A Poisson random variable is specified by its mean λ, which is allowed to be any positive real number. Hence, $a = 0$ is the correct choice.

Practice: Compute Likelihood of a Poisson Statistical Model

3 points possible (graded)

Let $X_1, \dots, X_n \overset{iid}{\sim} \text{Poiss}(\lambda^*)$ for some unknown $\lambda^* \in (0, \infty)$. You construct the associated statistical model $(E, \{\text{Poiss}(\lambda)\}_{\lambda \in \Theta})$ where E and Θ are defined as in the answers to the previous question.

Suppose you observe two samples $X_1 = 1, X_2 = 2$. What is $L_2(1, 2, \lambda)$? Express your answer in terms of λ.

$L_2(1, 2, \lambda) =$ **Answer:** e^(-2*lambda)*lambda^3/2

Next, you observe a third sample $X_3 = 3$ that follows $X_1 = 1$ and $X_2 = 2$. What is $L_3(1, 2, 3, \lambda)$?

$L_3(1, 2, 3, \lambda) =$ **Answer:** e^(-3*lambda)*lambda^6/12

Suppose your data arrives in a different order: $X_1 = 2, X_2 = 3, X_3 = 1$. What is $L_3(2, 3, 1, \lambda)$?

$L_3(2, 3, 1, \lambda) =$ **Answer:** e^(-3*lambda)*lambda^6/12

Solution:

The probability mass function of $\text{Poiss}(\lambda)$ is $x \mapsto e^{-\lambda}\frac{\lambda^x}{x!}$ where $x \in \mathbb{N} \cup \{0\}$. Hence by definition

$$L_n(x_1, \dots, x_n, \lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^n x_i}}{x_1! \cdots x_n!}.$$

Hence, first we plug in $n = 2$, $x_1 = 1$, and $x_2 = 2$:

$$L_2(1, 2, \lambda) = e^{-2\lambda}\frac{\lambda^{1+2}}{2!1!} = e^{-2\lambda}\frac{\lambda^3}{2}.$$

When the next sample arrives, we can simply evaluate the density of a Poisson at the observation:

$$P(X_3 = 3) = e^{-\lambda}\frac{\lambda^3}{3!}, \quad X \sim \text{Poiss}(\lambda)$$

and multiply this by the previous response:

$$L_3(1, 2, 3, \lambda) = e^{-\lambda}\frac{\lambda^3}{3!}L_2(1, 2, \lambda) = e^{-3\lambda}\frac{\lambda^6}{12}.$$

Properties of the Likelihood

1 point possible (graded)

Let $(E, \{P_\theta\}_{\theta \in \Theta})$ denote a discrete statistical model. Let p_θ denote the pmf of P_θ. Let $X_1, \ldots, X_n \overset{iid}{\sim} P_{\theta^*}$ where the parameter θ^* is unknown. Then the **likelihood** is the function

$$\begin{aligned} L_n : E^n \times \Theta &\to \mathbb{R} \\ (x_1, \ldots, x_n, \theta) &\mapsto \prod_{i=1}^{n} p_\theta(x_i). \end{aligned}$$

For our purposes, we think of $x_1, \ldots, x_n$ as observations of the random variables $X_1, \ldots, X_n$.

Which of the following are properties of the likelihood L_n? (Choose all that apply.)

Hint: It may be useful to consider your responses from the previous question.

- [ ] The likelihood does not change with the parameter θ.
- [ ] The likelihood can be updated sequentially as new samples are observed. For example, $L_3(x_1, x_2, x_3, \theta) = L_1(x_3, \theta) L_2(x_1, x_2, \theta)$. ✔
- [ ] The likelihood is symmetric: it doesn't matter the order in which we plug in the observations. For example, $L_4(x_1, x_2, x_3, x_4, \theta) = L_4(x_2, x_3, x_1, x_4, \theta)$, and this is true for any rearrangement of x_1, x_2, x_3, x_4. ✔
- [ ] If we eliminate a single observation, then the likelihood remains unchanged. For example, $L_3(x_1, x_2, x_3, \theta) = L_2(x_1, x_2, \theta)$.

Solution:

We examine the choices in order.

- "The likelihood does not change with the parameter θ." is incorrect. Rather, it is crucial that we interpret the likelihood L_n as a function of θ. That is, L_n varies as θ ranges over the parameter space Θ. This is evident in the likelihoods for the Bernoulli and Poisson models in the previous problems.
- "The likelihood can be updated sequentially as new samples are observed. For example, $L_3(x_1, x_2, x_3, \theta) = L_1(x_3, \theta) L_2(x_1, x_2, \theta)$." is also correct. In the previous problem, we saw that to compute the likelihood after observing $X_3 = 3$, we simply took the old likelihood $L_2(1, 2, \lambda)$ and multiplied it by $L_1(3, \lambda)$. Note that $L_1(x_3, \theta) = p_\theta(x_3)$, the density of P_θ evaluated at the new observation. Inspection of the defining formula

  $$L_n(x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} p_\theta(x_i)$$

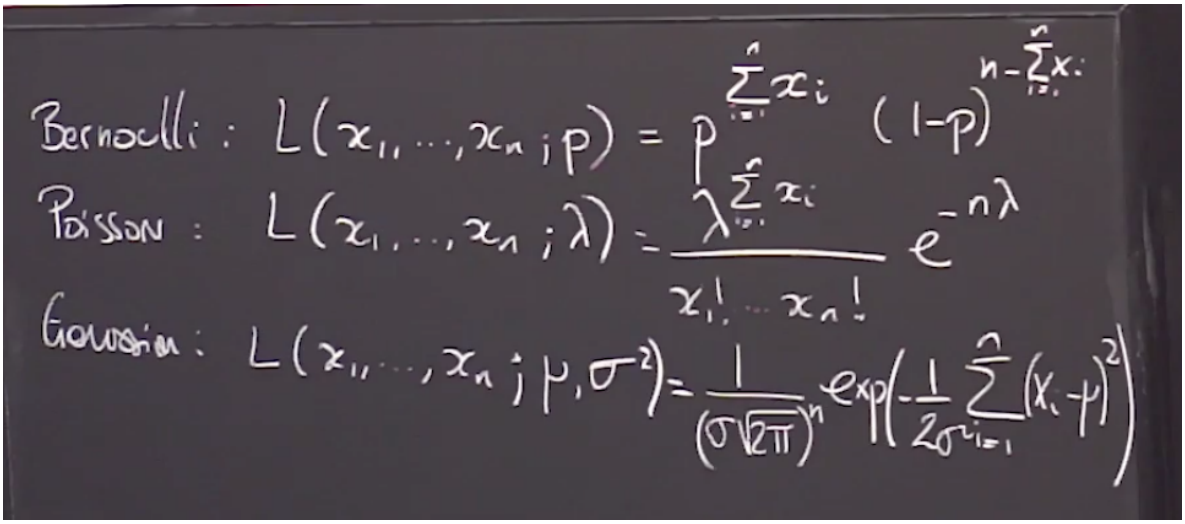
  implies that the likelihood can be updated sequentially in this fashion.
- "The likelihood is symmetric..." is correct. We observed in the previous problem that observing the samples in a different order does not affect the likelihood. This is also evident from the definition of the likelihood: we can take the product

  $$\prod_{i=1}^{n} p_\theta(x_i)$$

  in any order, and the result will still be the same.
- "If we eliminate a single observation, then the likelihood remains unchanged..." is incorrect. In the previous question, we saw that for a Poisson statistical model, $L_2(1, 2, \lambda)$ and $L_3(1, 2, 3, \lambda)$ do not have the same formula. Hence, deleting an observation from the sample will change the likelihood.

Maximum Likelihood Estimation

slides 14-18



Is the Likelihood Discrete or Continuous?

2 points possible (graded)

Setup:
Consider a **discrete** statistical model $M_1 = (\mathbb{Z}, \{\mathbf{P}_\theta\}_{\theta\in\mathbb{R}})$ and a **continuous** statistical model $M_2 = (\mathbb{R}, \{Q_\theta\}_{\theta\in\mathbb{R}})$. Let p_θ denote the pmf of $\mathbf{P}_\theta$, and let q_θ denote the pdf of Q_θ. Assume that p_θ and q_θ both vary continuously with the parameter θ for each fixed $x \in E$.

Let $x_1, \dots, x_n$ be fixed natural numbers and $y_1, \dots, y_n$ be fixed real numbers. Let $(L_1)_n$ denote the likelihood of the discrete model M_1, and let $(L_2)_n$ denote the likelihood of the continuous model M_2. Keeping $x_1, \dots, x_n$ and $y_1, \dots, y_n$ fixed, let's think of $(L_1)_n\,(x_1, \dots, x_n, \theta)$ and $(L_2)_n\,(y_1, \dots, y_n, \theta)$ as functions of θ.

Question
Decide whether the following claims about $(L_1)_n$ and $(L_2)_n$ are true or false.

The map $\theta \mapsto (L_1)_n\,(x_1, \dots, x_n, \theta)$ is a continuous function of θ.

- (●) True ✔
- ( ) False

The map $\theta \mapsto (L_2)_n\,(y_1, \dots, y_n, \theta)$ is a continuous function of θ.

- (●) True ✔
- ( ) False

Solution:

Observe that

$$(L_1)_n(x_1, \dots, x_n, \theta) = \prod_{i=1}^{n} p_\theta(x_i), \quad (L_2)_n(y_1, \dots, y_n, \theta) = \prod_{i=1}^{n} q_\theta(y_i).$$

We are given that p_θ and q_θ are both continuous function of the parameter $\theta \in \mathbb{R}$. Since products of continuous functions are continuous, this implies that the maps $\theta \mapsto (L_1)_n(x_1, \dots, x_n, \theta)$ and $\theta \mapsto (L_2)_n(y_1, \dots, y_n, \theta)$ are continuous functions of the parameter $\theta \in \mathbb{R}$.

Remark: It may be confusing that even the likelihood of a discrete statistical model can be continuous. However, considering the likelihood of a Bernoulli (derived in a previous question),

$$L(x_1, \dots, x_n, p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

we can clearly see that the above varies continuously as a function of the *parameter*. This is also true for a host of other discrete models (for example, the Poisson model).

Quiz: Likelihood of a Gaussian Statistical Model

3 points possible (graded)

Let $X_1, \dots, X_n \overset{iid}{\sim} N(\mu^*, (\sigma^*)^2)$ for some unknown $\mu^* \in \mathbb{R}, (\sigma^*)^2 > 0$. You construct the associated statistical model $\left(\mathbb{R}, \{N(\mu, \sigma^2)\}_{(\mu,\sigma^2)\in\mathbb{R}\times(0,\infty)}\right)$.

The likelihood of this model can be written

$$L_n\left(x_1, \dots, x_n, (\mu, \sigma^2)\right) = \frac{1}{(\sigma\sqrt{2\pi})^C} \exp\left(-\frac{1}{A}\sum_{i=1}^{C} B_i\right)$$

where A depends on σ, B_i depends on μ and x_i. Find A,B_i and C.

(Choose a B_i that has coefficient 1 for the highest degree term in x_i.)

(Type **sigma** for σ, **mu** for μ, and **x_i** for x_i.)

$A =$ [ ] **Answer:** 2*sigma^2

$B_i =$ [ ] **Answer:** (x_i - mu)^2

$C =$ [ ] **Answer:** n

Solution:

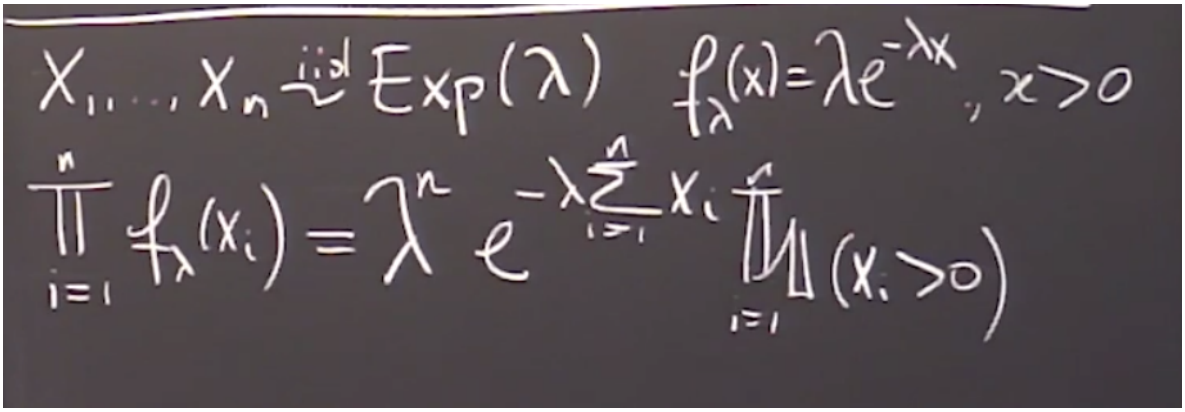
The pdf of a Gaussian distribution $N(\mu, \sigma^2)$ is the function $x \mapsto \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$. Hence, the likelihood is

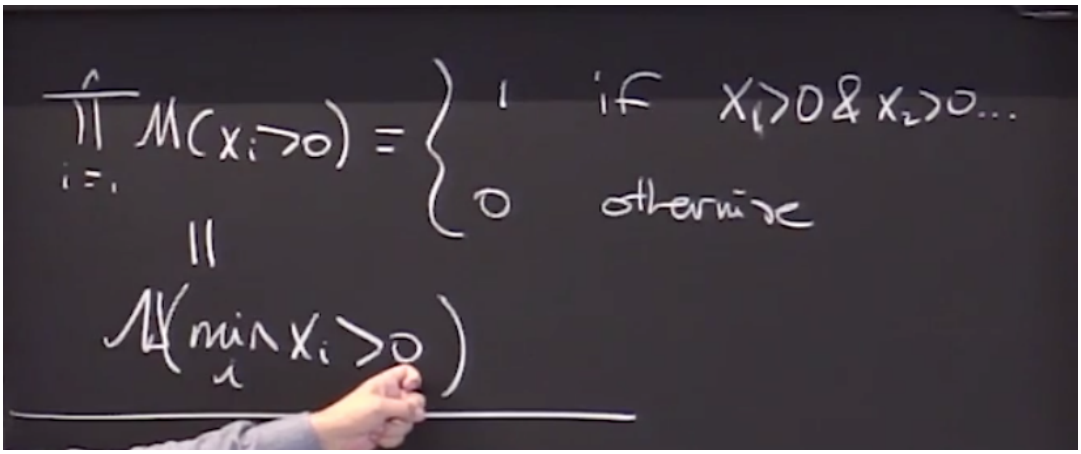
$$L_n(x_1, \dots, x_n, (\mu, \sigma^2)) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right) = \frac{1}{(\sigma\sqrt{2\pi})^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right).$$

Hence,

$$A = 2\sigma^2, \quad B_i = (x_i-\mu)^2, \quad C = n.$$

slide 19





1 only if all the observations are 1, so like saying that the minimum of the observations is positive (since they can only be 0 or 1) - this is a way to reduced many indicators to one indicator which is good practice

Product of Indicators

1 point possible (graded)

Rewrite the product $\mathbf{1}(x_1 \leq 5)\,\mathbf{1}(x_2 \leq 5)$ as a single indicator function. That is, find $f(x_1, x_2)$ in the following equation:

$$\mathbf{1}(x_1 \leq 5)\,\mathbf{1}(x_2 \leq 5) \;=\; \mathbf{1}(f(x_1, x_2) \leq 5)\,.$$

(Choose all that apply.)

- [ ] $f(x_1, x_2) = x_1 x_2$
- [ ] $f(x_1, x_2) = \dfrac{x_1 + x_2}{2}$
- [ ] $f(x_1, x_2) = \operatorname{sign}(x_1)\operatorname{sign}(x_2)$
- [ ] $f(x_1, x_2) = \max(x_1, x_2)$ ✔
- [ ] $f(x_1, x_2) = \min(x_1, x_2)$

Solution:

We need to find $f(x_1, x_2)$ such that

$$f(x_1, x_2) \leq 5 \iff x_1 \leq 5 \text{ and } x_2 \leq 5$$

We go through the choices in order. We leave it to you to find counter examples:

1. $x_1, x_2 \leq 5$ does not imply $x_1 x_2 \leq 5$;
2. $\dfrac{x_1 + x_2}{2} \leq 5$ does not imply $x_1, x_2 \leq 5$;
3. $\operatorname{sign}(x_1)\operatorname{sign}(x_2) \leq 5$ for all $x_1, x_2,$ and in particular does not imply $x_1, x_2 \leq 5$;
4. $\max(x_1, x_2) \leq 5$ if and only if both $x_1, x_2 \leq 5$, so this is a valid choice for $f(x_1, x_2)$;
5. $\min(x_1, x_2) \leq 5$ implies one of x_1, x_2 to be at most 5 but not necessarily both.

slide 20: uniform

$X_1, \dots, X_n \sim \text{Unif}([0, b])$

$f_b(x) = \frac{1}{b} \mathbb{1}(0 \le x \le b)$

$\prod_{i=1}^{n} f_b(x) = \frac{1}{b^n} \prod_{i=1}^{n} \mathbb{1}(0 \le x_i \le b)$

$= \frac{1}{b^n} \mathbb{1}(\min_i x_i \ge 0) \mathbb{1}(\max_i x_i \le b)$

Bernoulli : $L(x_1, \dots, x_n ; p) = p^{\sum_{i=1}^{n} x_i} (1-p)^{n - \sum_{i=1}^{n} x_i}$

Poisson : $L(x_1, \dots, x_n ; \lambda) = \frac{\lambda^{\sum_{i=1}^{n} x_i}}{x_1! \cdots x_n!} e^{-n\lambda}$

Gaussian : $L(x_1, \dots, x_n ; \mu, \sigma^2) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right)$

Exponential : $L(x_1, \dots, x_n, \lambda) = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right)$

Uniform : $L(x_1, \dots, x_n, b) = \frac{1}{b^n} \mathbb{1}(\max_i x_i \le b)$

Review: Maximizing composite functions

1 point possible (graded)

The **arguments of the minima** (*resp.* **arguments of the maxima**) of a function $f(x)$, denoted by $\operatorname{argmin} f(x)$ (*resp.* $\operatorname{argmax} f(x)$), is the value(s) of x at which $f(x)$ is minimum (*resp.* maximum). We can also restrict to a subset S of the domain of f, and denote by $\operatorname{argmin}_{x \in S} f(x)$ (*resp.* $\operatorname{argmax}_{x \in S} f(x)$) the value(s) of $x \in S$ at which $f(x)$ is minimum (*resp.* maximum) over S.

Let $f(x) > 0$ be continuous **positive** function with $\max_x f(x) = 1$. (Note that $\max_x f(x)$ is the maximum value of the function, which is different from $\operatorname{argmax} f(x)$, the value of the argument x at which the function is maximum.)

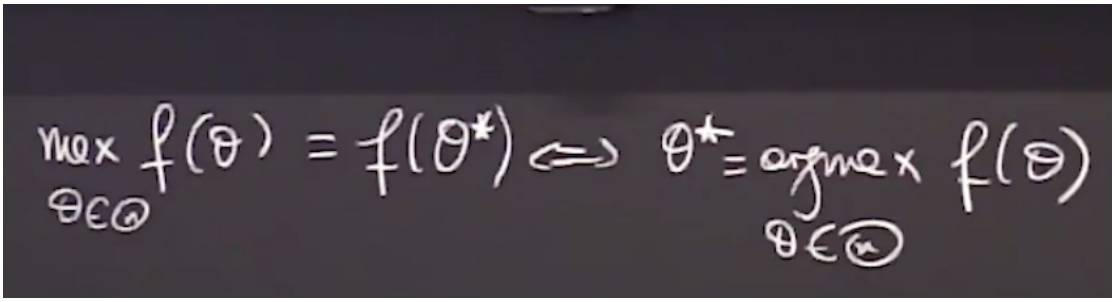
Which of the following functions of $f(x)$ has the same argmax as $f(x)$? In other words, which of the following attain their maxima at the same x-value(s) as $f(x)$?
(Choose all that apply.)

- [x] $f(x)^2$ ✔
- [x] $\sqrt{f(x)}$ ✔
- [ ] $\ln(f(x))$ ✔
- [x] $-\ln\left(\frac{1}{f(x)}\right)$ ✔
- [x] $\cos(f(x))$
- [ ] $-\cos(2f(x))$ ✔

Solution:

We go through the choices in order.

- Since y^2, $\sqrt{y}$, $\ln(y) = -\ln\left(\frac{1}{y}\right)$ are all **strictly increasing** functions, their value increases as y increases. Hence, the functions $f(x)^2$, $\sqrt{f(x)}$, $\ln(f(x))$, $-\ln\left(\frac{1}{f(x)}\right)$ attain their maxima when $f(x)$ attain its maximum, which is at $x = \operatorname{argmax} f(x)$.
- The cosine function is strictly decreasing in $(0, \pi)$. Given $\max_x f(x) = 1 < \pi$, $\cos(f(x))$ is in fact minimum when $f(x)$ is maximum.
- On the other hand, $-\cos(2y)$ is strictly increasing for $0 < 2y < \pi$. Since $\max_x 2f(x) = 2 < \pi$, we conclude that $-\cos(2f(x))$ is maximum again when $f(x)$ is maximum, at $x = \operatorname{argmax} f(x)$.



Concept Check: Interpreting the Maximum Likelihood Estimator

1 point possible (graded)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathbf{P}_{\theta^*}$ be discrete random variables. We construct a statistical model $(E, \{\mathbf{P}_\theta\}_{\theta \in \mathbb{R}})$ where $\mathbf{P}_\theta$ has pmf p_θ. We observe our sample to be $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$. The **maximum likelihood estimator** for θ^* is defined to be

$$\hat{\theta}_n^{MLE} = \text{argmax}_{\theta \in \mathbb{R}} \left(\prod_{i=1}^{n} p_\theta(X_i) \right).$$

Which of the following is a correct interpretation of the maximum likelihood estimator (MLE) when applied to the sample $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$?
(Choose all that apply.)

- [x] The value of θ that maximizes the probability $\mathbf{P}_\theta$ of observing the data set $(x_1, \ldots, x_n)$. ✔
- [x] The value of θ that minimizes an estimator of the KL divergence between $\mathbf{P}_\theta$ and the true distribution $\mathbf{P}_{\theta^*}$. ✔
- [ ] It is the true parameter θ^*

Solution:

- "The value of θ that maximizes the probability that $\mathbf{P}_\theta$ generates the data set $(x_1, \ldots, x_n)$." is correct. Since the likelihood is the joint density of n iid samples from $\mathbf{P}_\theta$,

  $$\mathbf{P}_\theta [X_1 = x_1, \ldots, X_n = x_n] = L_n(x_1, \ldots, x_n, \theta).$$

  Hence, the MLE finds $\hat{\theta}_n$ that maximizes the probability that $x_1, \ldots, x_n$ were sampled from $P_{\hat{\theta}_n}$.

- "The value of θ that minimizes the KL divergence between $\mathbf{P}_\theta$ and the true distribution $\mathbf{P}_{\theta^*}$." is correct. In fact, this is how the MLE was derived from KL divergence. See the third section "Parameter Estimation via KL Divergence" of this lecture to review this fact.

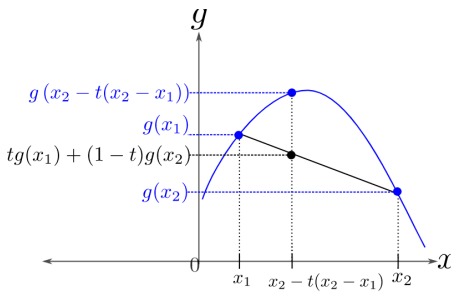
- "It is the true parameter θ^*" is incorrect. The MLE is an estimator– it is constructed from the finite amount of data $x_1, \ldots, x_n$ that we are given– so we can't hope for it to exactly recover the true parameter.

  Remark: Under some technical conditions the MLE is a **weakly consistent estimator** for θ^*, meaning that the MLE will converge to θ^* in probability under these conditions. However, there are examples of statistical models where the maximum likelihood estimator will **not** converge to the true parameter.

A function $g : I \to \mathbb{R}$ is **concave** (or concave down) on an interval I, if for all pairs of real numbers $x_1 < x_2 \in I$

$$g(tx_1 + (1-t)x_2) \geq tg(x_1) + (1-t)g(x_2) \qquad \text{for all } 0 < t < 1.$$

Geometrically, this means that for $x_1 < x < x_2$, the graph of g is **above** the secant line connecting the two points $(x_1, g(x_1))$ and $(x_2, g(x_2))$.



At $x = x_2 - t(x_2 - x_1) = tx_1 + (1-t)x_2$, the y-value of the graph of g is $g(x) = g(tx_1 + (1-t)x_2)$, while the y-value of the secant line is $tg(x_1) + (1-t)g(x_2)$.

If the inequality is strict, i.e. if

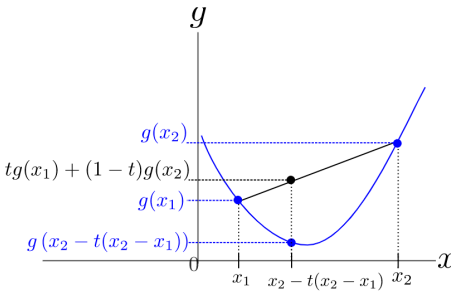
$$g(tx_1 + (1-t)x_2) > tg(x_1) + (1-t)g(x_2) \qquad \text{for all } 0 < t < 1.$$

then g is **strictly concave** .

The definition for **(strictly) convex** is analogous. A function $g : I \to \mathbb{R}$ is **convex** (or concave up), where I is an interval, if for all pairs of real numbers $x_1 < x_2 \in I$

$$g(tx_1 + (1-t)x_2) \leq tg(x_1) + (1-t)g(x_2) \qquad \text{for all } 0 < t < 1.$$

Geometrically, this means that for $x_1 < x < x_2$, the graph of g is **below** the secant line connecting the two points $(x_1, g(x_1))$ and $(x_2, g(x_2))$.



At $x = x_2 - t(x_2 - x_1) = tx_1 + (1-t)x_2$, the y-value of the graph of g is $g(x) = g(tx_1 + (1-t)x_2)$, while the y-value of the secant line is $tg(x_1) + (1-t)g(x_2)$.

If the inequality is strict, i.e. if

$$g(tx_1 + (1-t)x_2) < tg(x_1) + (1-t)g(x_2) \qquad \text{for all } 0 < t < 1.$$

then g is **strictly convex** .

If in addition g is twice differentiable in the interval I, i.e. $g''(x)$ exists for all $x \in I$, then g is

- **concave** if and only if $g''(x) \leq 0$ for all $x \in I$;
- **strictly concave** if $g''(x) < 0$ for all $x \in I$;
- **convex** if and only if $g''(x) \geq 0$ for all $x \in I$;
- **strictly convex** if $g''(x) > 0$ for all $x \in I$;

Note: In the lecture video and slides, we used these inequality conditions on the second derivative to defined concave functions and strictly concave functions *analytically*. The *synthetic* definition above is slightly more general because it does not require differentiability at every point. For example, the function $x \mapsto x^4$ is strictly convex according to the definition above, but has three vanishing derivatives at the origin $x = 0$.

Review: 1D Optimization via Calculus

4 points possible (graded)

(For this problem, you are welcome to use any computational tools that would be helpful.)

Let $f(x) = \frac{1}{3}x^3 - x^2 - 3x + 10$ defined on the interval $[-4, 4]$.

Let x_1 and x_2 be the critical points of f, and let's impose that $x_1 < x_2$. Fill in the next two boxes with the values of x_1 and x_2, respectively: (Recall that the **critical points** of f are those $x \in \mathbb{R}$ such that $f'(x) = 0$.)

$x_1 =$ **Answer:** -1

$x_2 =$ **Answer:** 3

Fill in the next two boxes with the values of $f''(x_1)$ and $f''(x_2)$, respectively:

$f''(x_1) =$ **Answer:** -4

$f''(x_2) =$ **Answer:** 4

Solution:

Observe that

$$f'(x) = x^2 - 2x - 3 = (x - 3)(x + 1).$$

Hence the **critical points** are $x_1 = -1$ and $x_2 = 3$. The **second derivative** is

$$f''(x) = 2x - 2$$

so that

$$f''(x_1) = -4, \quad f''(x_2) = 4.$$

Review: 1D Optimization via Calculus (Continued)

4 points possible (graded)

(For this problem, you are welcome to use any computational tools that would be helpful.)

Recall that x_1 and x_2 are the critical points of the function $f(x) = \frac{1}{3}x^3 - x^2 - 3x + 10$.

According to the second derivative test, x_1 is a ...

- (●) Local Maximum ✔
- ( ) Local Minimum
- ( ) None of the above

and x_2 is a

- ( ) Local Maximum
- (●) Local Minimum ✔
- ( ) None of the above

At what value of x is the (global) minimum value of $f(x)$ attained on the interval $[-4, 4]$?

[ ] **Answer:** -4

At what value of x the (global) maximum value of $f(x)$ attained on the interval $[-4, 4]$?

[ ] **Answer:** -1

Solution:

The previous problem implies that f is concave at x_1 and convex at x_2, so x_1 is a **local maximum** and x_2 is a **local minimum**. To figure out the *global* extrema, we need to test the critical points as well as the endpoints: -4 and 4. We compute that

$$f(x_1) = \frac{35}{3} \approx 11.6666, \quad f(x_2) = 1$$

$$f(-4) = -\frac{46}{3} \approx -15.33333, \quad f(4) = 10/3 \approx 3.3333$$

Hence the **maximum value** of f on $[-4, 4]$ is $\frac{35}{3} \approx 11.6666$ and the **minimum value** is $-\frac{46}{3} \approx -15.33333$.

Remark: It is very important to remember to test the endpoints when doing optimization.

Strict Concavity

1 point possible (graded)

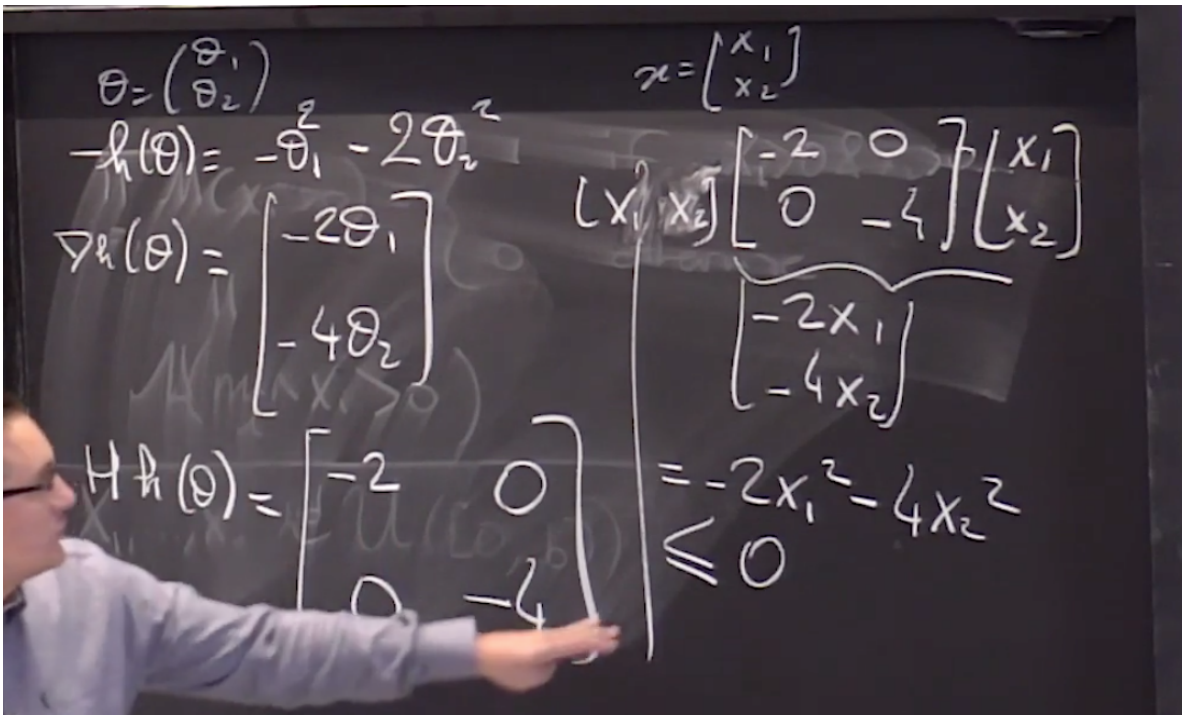
Which of the following functions are strictly concave? (Choose all that apply.) (Recall that a twice-differentiable function $f : I \to \mathbb{R}$, where I is a subset of $\mathbb{R}$, is **strictly concave** if $f''(x) < 0$ for all $x \in I$.)

- [ ] $f_1(x) = x$ on $\mathbb{R}$
- [x] $f_2(x) = -e^{-x}$ on $\mathbb{R}$ ✔
- [x] $f_3(x) = x^{0.99}$ on the interval $(0, \infty)$ ✔
- [ ] $f_4(x) = x^2$ on $\mathbb{R}$

Solution:

- $f_1(x) = x$ is **not** strictly concave because $f_1''(x) = 0$.
- $f_2(x) = -e^{-x}$ is strictly concave because $f_2''(x) = -e^{-x} < 0$ for all $x \in \mathbb{R}$.
- $f_3(x) = x^{0.99}$ is strictly concave because $f_3''(x) = (0.99)(-.01)x^{-1.01} < 0$ for all $x \in (0, \infty)$.
- $f_4(x) = x^2$ is **not** strictly concave because $f_4''(x) = 2 > 0$. In fact, this function is strictly *convex*.

slide 24, showing how to get the Hessian
take derivative with respect to theta1 then derivative with respect to theta 2 then multiply by x and xT



Multivariable Calculus Review: Compute the Gradient

1 point possible (graded)

Let

$$f: \quad \mathbb{R}^d \quad \to \quad \mathbb{R}$$
$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{pmatrix} \quad \mapsto \quad f(\theta).$$

denote a **differentiable** function. The **gradient** of f is the vector-valued function

$$\nabla f: \quad \mathbb{R}^d \quad \to \quad \mathbb{R}^d$$
$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{pmatrix} \quad \mapsto \quad \left. \begin{pmatrix} \frac{\partial f}{\partial \theta_1} \\ \frac{\partial f}{\partial \theta_2} \\ \vdots \\ \frac{\partial f}{\partial \theta_d} \end{pmatrix} \right|_\theta .$$

Consider $f(\theta) = -c_1\theta_1^2 - c_2\theta_2^2 - c_3\theta_3^2$ where $c_1, c_2, c_3 > 0$ are positive real numbers.

Compute the gradient ∇f.

(Enter your answer as a vector, e.g., type **[3,2,x]** for the vector $\begin{pmatrix} 3 \\ 2 \\ x \end{pmatrix}$. Note the square brackets, and commas as separators. Enter **c_i** for c_i, **theta_i** for θ_i.)

$\nabla f =$ [-2*c_1*theta_1, -2*c_2*theta_2, -2*c_3*theta_3] **Answer:** [-2*c_1*theta_1,-2*c_2*theta_2,-2*c_3*theta_3]

$[-2 \cdot c_1 \cdot \theta_1, -2 \cdot c_2 \cdot \theta_2, -2 \cdot c_3 \cdot \theta_3]$

Solution:

$$f(\theta) = -c_1\theta_1^2 - c_2\theta_2^2 - c_3\theta_3^2$$
$$\nabla f(\theta) = \left. \begin{pmatrix} \frac{\partial f}{\partial \theta_1} \\ \frac{\partial f}{\partial \theta_2} \\ \frac{\partial f}{\partial \theta_3} \end{pmatrix} \right|_\theta = \begin{pmatrix} -2c_1\theta_1 \\ -2c_2\theta_2 \\ -2c_3\theta_3 \end{pmatrix}.$$

Multivariable Calulus Review: Compute the Hessian Matrix

1 point possible (graded)

As above, let

$$f:\ \mathbb{R}^d\ \to\ \mathbb{R}\quad \theta=\begin{pmatrix}\theta_1\\ \theta_2\\ \vdots\\ \theta_d\end{pmatrix}\ \mapsto\ f(\theta)\,.$$

denote a **twice-differentiable** function.

The **Hessian** of f is the matrix

$$\mathbf{H}\,f\ :\ \mathbb{R}^d\ \to\ \mathbb{R}^{d\times d}$$

whose entry in the i-th row and j-th column is defined by

$$(\mathbf{H}\,f)_{ij}\ :=\ \frac{\partial^2}{\partial\theta_i\partial\theta_j}f,\qquad 1\le i,j\le d.$$

The Hessian matrix of f in this context is also denoted by $\nabla^2 f$, the **second derivative** of f. This is not to be confused with the "Laplacian" of f, which is also denoted the same way.

Consider the same function $f(\theta)=-c_1\theta_1^2-c_2\theta_2^2-c_3\theta_3^2$ where $c_1,\ c_2,\ c_3>0$ as in the previous problem. Compute the Hessian matrix $\mathbf{H}f$.

(Enter your answer as a matrix, e.g. by typing **[[1,2],[5*x,y-1]]** for the matrix $\begin{pmatrix}1 & 2\\ 5x & y-1\end{pmatrix}$. Note the square brackets, and commas as separaters.)

$\mathbf{H}f=$ [[-2*c_1, 0, 0], [0,-2*c_2,0], [0,0,-2*c_3]]

Answer: [[-2*c_1,0,0],[0,-2*c_2,0],[0,0,-2*c_3]]

Solution:

Recall from the previous problem:

$$\begin{aligned}f(\theta)\ &=\ -c_1\theta_1^2-c_2\theta_2^2-c_3\theta_3^2\\ \nabla f(\theta)\ &=\ \begin{pmatrix}\frac{\partial f}{\partial\theta_1}\\ \frac{\partial f}{\partial\theta_2}\\ \frac{\partial f}{\partial\theta_3}\end{pmatrix}\Bigg|_\theta\ =\ \begin{pmatrix}-2c_1\theta_1\\ -2c_2\theta_2\\ -2c_3\theta_3\end{pmatrix}.\end{aligned}$$

One way to compute the Hessian is to start will in j-th column of the Hessian matrix by the gradient of the j-th component of ∇f. We obtain:

$$\begin{aligned}f(\theta)\ &=\ -c_1\theta_1^2-c_2\theta_2^2-c_3\theta_3^2\\ \mathbf{H}f(\theta)\ &=\ \begin{pmatrix}| & | & |\\ \nabla(-2c_1\theta_1) & \nabla(-2c_2\theta_2) & \nabla(-2c_3\theta_3)\\ | & | & |\end{pmatrix}\\ &=\ \begin{pmatrix}-2c_1 & 0 & 0\\ 0 & -2c_2 & 0\\ 0 & 0 & -2c_3\end{pmatrix}.\end{aligned}$$

Semi-Definiteness

3 points possible (graded)

A symmetric (real-valued) $d \times d$ matrix $\mathbf{A}$ is **positive semi-definite** if

$$\mathbf{x}^T \mathbf{A}\, \mathbf{x} \geq 0 \qquad \text{for all } \mathbf{x} \in \mathbb{R}^d.$$

If the inequality above is strict, i.e. if $\mathbf{x}^T \mathbf{A}\, \mathbf{x} > 0$ for all non-zero vectors $\mathbf{x} \in \mathbb{R}^d$, then $\mathbf{A}$ is **positive definite** .

Analogously, a symmetric (real-valued) $d \times d$ matrix $\mathbf{A}$ is **negative semi-definite** (*resp.* **negative definite**) if $\mathbf{x}^T \mathbf{A}\, \mathbf{x}$ is **non-positive** (*resp.* **negative**) for all $\mathbf{x} \in \mathbb{R}^d - \{\mathbf{0}\}$.

Note that by definition, positive (or negative) definiteness implies positive (or negative) semi-definiteness.

Consider the same function as in the problems above:

$$f(\theta) = -c_1\theta_1^2 - c_2\theta_2^2 - c_3\theta_3^2 \quad \text{where } c_1, c_2, c_3 > 0.$$

Compute $\mathbf{x}^T (\mathbf{H}f)\, \mathbf{x}$ where $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$.

$\mathbf{x}^T (\mathbf{H}f)\, \mathbf{x} =$ [ ]

Answer: -2*c_1*x_1^2-2*c_2*x_2^2-2*c_3*x_3^2

Solution:

Recall from the previous problem that

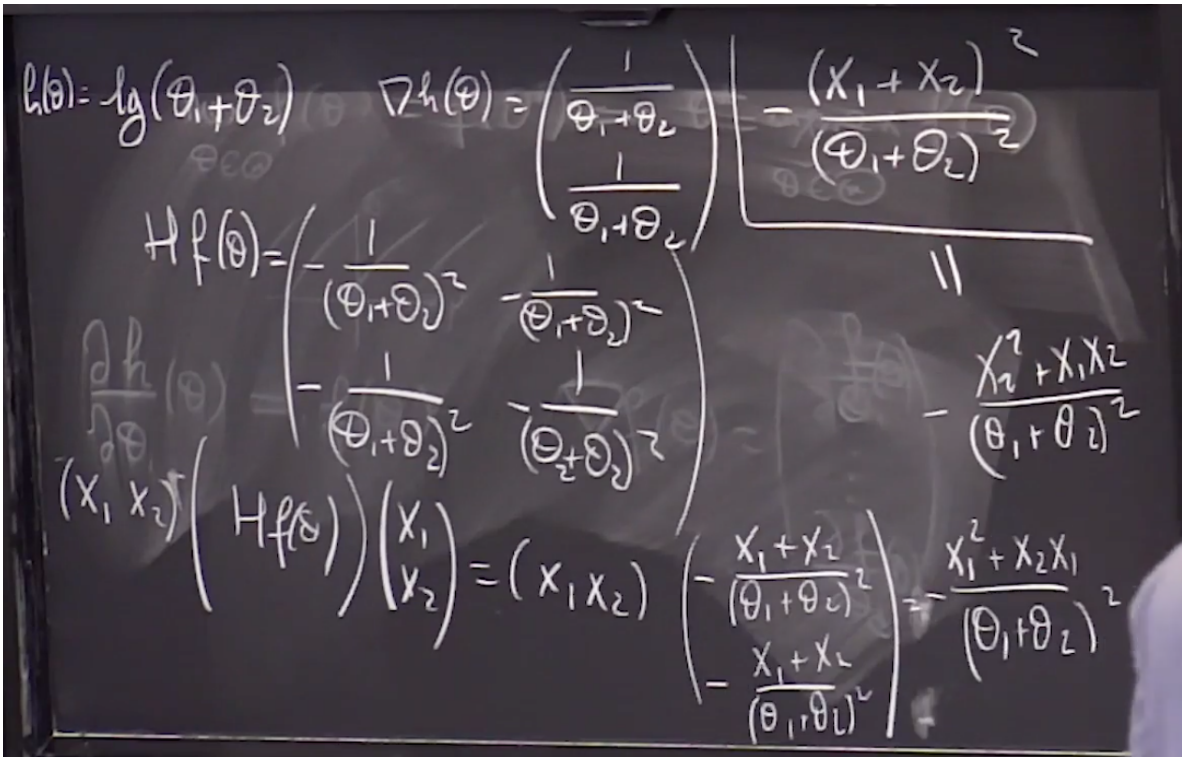
$$\mathbf{H}f(\theta) = \begin{pmatrix} -2c_1 & 0 & 0 \\ 0 & -2c_2 & 0 \\ 0 & 0 & -2c_3 \end{pmatrix}.$$

Then

$$\begin{aligned} \mathbf{x}^T (\mathbf{H}f)\, \mathbf{x} &= \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} -2c_1 & 0 & 0 \\ 0 & -2c_2 & 0 \\ 0 & 0 & -2c_3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \\ &= -2c_1x_1^2 - 2c_2x_2^2 - 2c_3x_3^2 < 0. \end{aligned}$$

Since $c_1, c_2, c_3 > 0$, this means the $\mathbf{H}f$ is negative definite, (also negative semi-definite), and hence f is strictly concave (also concave).

slide 24 last example



Combination of Convex functions

3 points possible (graded)

Let f_1, f_2 be convex functions on $\mathbb{R}$.

Determine if the following functions are necessarily convex or concave.

Hint: Recall that a function $g : I \to \mathbb{R}$ defined on an interval I is convex, if for all pairs of real numbers $x_1 < x_2$ in I we have:

$$g\left(tx_1 + (1-t)\,x_2\right) \leq tg\left(x_1\right) + (1-t)\,g\left(x_2\right) \qquad \text{for all } \; 0 \leq t \leq 1.$$

- $3f_1 + 2f_2$:

  - (●) Convex ✔
  - ( ) Concave
  - ( ) Cannot be determined without more information

- $-10f_1$:

  - ( ) Convex
  - (●) Concave ✔
  - ( ) Cannot be determined without more information

- $f_2 f_1$:

  - (●) Convex
  - ( ) Concave
  - ( ) Cannot be determined without more information ✔

Solution:

Given f_1, f_2 are convex, we have

$$f_1\left(tx_1+(1-t)x_2\right) \leq tf_1\left(x_1\right)+(1-t)f_1\left(x_2\right) \qquad \text{for all } 0 \leq t \leq 1$$

and the same holds for f_2.

- The same inequality holds for $g = 3f_1 + 2f_2$:

$$\begin{aligned} g\left(tx_1+(1-t)x_2\right) &= 3f_1\left(tx_1+(1-t)x_2\right)+2f_2\left(tx_1+(1-t)x_2\right) \\ &\leq 3\left(tf_1\left(x_1\right)+(1-t)f_1\left(x_2\right)\right)+2\left(tf_2\left(x_1\right)+(1-t)f_2\left(x_2\right)\right) \\ &= tg\left(x_1\right)+(1-t)g\left(x_2\right). \end{aligned}$$

  Hence $3f_1 + 2f_2$ is also convex.

  Remark: In general, any function $c_1 f_1 + c_2 f_2$ where $c_1, c_2 > 0$ is convex of f_1, f_2 are.

- $-10f_1$ is concave, because it is negative of a convex function.

- $f_1 f_2$ is not necessary convex For example, is $f_1(x) = x$, and $f_2 = x^2$, then $(f_1 f_2)(x) = x^3$ which is neither convex nor concave. Other examples of f_1 and f_2, e.g. $f_1 = f_2 = x^2$ will lead to $f_1 f_2$ being convex.

10. Concavity in higher dimensions and Eigenvalues

Exercises due Jun 24, 2020 08:59 JST *Past Due*

Bookmark this page

Concavity in 2 dimensions: Compute the Hessian

4 points possible (graded)

What is the Hessian $\mathbf{H}f$ of the function $f(x, y) = -2x^2 + \sqrt{2}xy - \frac{5}{2}y^2$? Fill in the values of the entries of $\mathbf{H}f$.

$(\mathbf{H}f)_{11} =$ -4 **Answer:** -4 $(\mathbf{H}f)_{12} =$ **Answer:** sqrt(2)

$(\mathbf{H}f)_{21} =$ **Answer:** sqrt(2) $(\mathbf{H}f)_{22} =$ **Answer:** -5

Solution:

We compute that

$$(\mathbf{H}f)_{11} = \frac{\partial^2 f}{\partial x^2} = -4, \qquad (\mathbf{H}f)_{12} = \frac{\partial^2 f}{\partial x \partial y} = \sqrt{2}$$
$$(\mathbf{H}f)_{21} = \frac{\partial^2 f}{\partial x \partial y} = \sqrt{2}, \qquad (\mathbf{H}f)_{22} = \frac{\partial^2 f}{\partial y^2} = -5.$$

So this implies that

$$\mathbf{H}f = \begin{pmatrix} -4 & \sqrt{2} \\ \sqrt{2} & -5 \end{pmatrix}.$$

(Optional) Concavity in 2 dimensions: Positive Definiteness and Eigenvalues

0 points possible (ungraded)

A symmetric (real-valued) $d \times d$ matrix $\mathbf{A}$ is **positive semi-definite** (*resp.* **positive definite**) if and only if all of its eigenvalues are **non-negative** (*resp.* **positive**).

Analogously, it is **negative semi-definite** (*resp.* **negative definite**) if and only if all of its eigenvalues are **non-positive** (*resp.* **negative**).

As above, consider $f(x, y) = -2x^2 + \sqrt{2}xy - \frac{5}{2}y^2$.

What are the eigenvalues λ_1, λ_2 of $\mathbf{H}f$? Assume that $\lambda_1 < \lambda_2$.

$\lambda_1 =$ -4/sqrt(2) **Answer:** -6 $\lambda_2 =$ -5/sqrt(2) **Answer:** -3

Based on your answer to the last question, f is ...

- ◯ Convex
- ◉ Concave ✔
- ◯ None of the Above

Solution:

Recall from the previous problem that the Hessian of f is

$$\mathbf{H}f = \begin{pmatrix} -4 & \sqrt{2} \\ \sqrt{2} & -5 \end{pmatrix}.$$

To find the eigenvalues, we need to solve for λ such that

$$\det(\mathbf{H}f - \lambda I) = \det\left(\begin{pmatrix} -4-\lambda & \sqrt{2} \\ \sqrt{2} & -5-\lambda \end{pmatrix}\right) = \lambda^2 + 9\lambda + 18 = 0.$$

Factoring the quadratic: $\lambda^2 + 9\lambda + 18 = (\lambda + 6)(\lambda + 3)$ shows that $\lambda_1 = -6$ and $\lambda_2 = -3$.

The function f is twice-differentiable, so it is concave if $x^T \mathbf{H}f x \leq 0$ for all $x \in \mathbb{R}^2$. By the remark in the problem statement, this is equivalent to all of the eigenvalues of $\mathbf{H}f$ being negative. Hence, f is concave (in fact it is *strictly* concave).

Concavity and Convexity in Higher Dimensions II

2 points possible (graded)

As in the problem on the previous page, $f(x, y) = -2x^2 + \sqrt{2}xy - \frac{5}{2}y^2$. Based on your answer to the question, which of the following is true?

- ◯ f has a unique (global) minimizer.
- ◯ f has a unique (global) maximizer. ✔
- ◯ f has more than one (global) minimizer
- ◉ f has more than one (global) maximizer

Where is the critical point of f? (If there is more than one critical point, just enter one of them.)

(Enter your answer as a vector, e.g., type **[3,2]** for the vector $\begin{pmatrix} 3 \\ 2 \end{pmatrix}$, which corresponds to the point $(3, 2)$ on the (x, y)-plane).

Critical point of f: **Answer:** [0,0]

Solution:

Since f is twice-differentiable and strictly concave, we know there will be a unique global maximum.

The critical points of f are those points where $\nabla f = (\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}) = 0$. We compute that

$$\nabla f = (-4x + \sqrt{2}y, -5y + \sqrt{2}x),$$

which is 0 if and only if $x = y = 0$.

Concavity Concept Check

1 point possible (graded)

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a twice-differentiable function, such that the top-left element of the Hessian matrix $\mathbf{H}f(0, 0)_{1,1} > 0$ is positive. Is f concave?

- ◯ Yes
- ◯ No ✔
- ◉ Not possible to determine from given information

Solution:

Recall that f is concave at $(0,0)$ if for all $(x,y) \in \mathbb{R}^2$,

$$(x,y)\,\mathbf{H}f\,(0,0)\begin{pmatrix} x \\ y \end{pmatrix} < 0.$$

By expanding and using the definition of the Hessian, we see that

$$(x,y)\,\mathbf{H}f\,(0,0)\begin{pmatrix} x \\ y \end{pmatrix} = x^2\frac{\partial^2}{\partial x^2}f\,(0,0) + xy\left(\frac{\partial^2}{\partial x \partial y}f\,(0,0) + \frac{\partial^2}{\partial y \partial x}f\,(0,0)\right) + y^2\frac{\partial^2}{\partial y^2}f\,(0,0)\,.$$

By assumption, we know that $\frac{\partial^2}{\partial x^2}f\,(0,0) > 0$. Hence,

$$(1,0)^T\mathbf{H}f\,(0,0)\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{\partial^2}{\partial x^2}f\,(0,0) > 0.$$

This violates the definition of concavity, so the correct response is "No."

Intuition for Optimizing Concave Functions

1 point possible (graded)

Let $f : \mathbb{R} \to \mathbb{R}$ be a twice-differentiable function that has a critical point and is strictly concave. Recall that the critical point is a **unique maximizer** of f.

You choose an initial guess for the maximizer $x_0 = 0$ (which may be very far from the true maximizer). You compute the derivative and observe that $f'(x_0) < 0$. Where is the maximizer of f?

- (●) To the left of $x_0 = 0$ ✔
- ( ) To the right of $x_0 = 0$
- ( ) Very far from x_0
- ( ) Very close to x_0

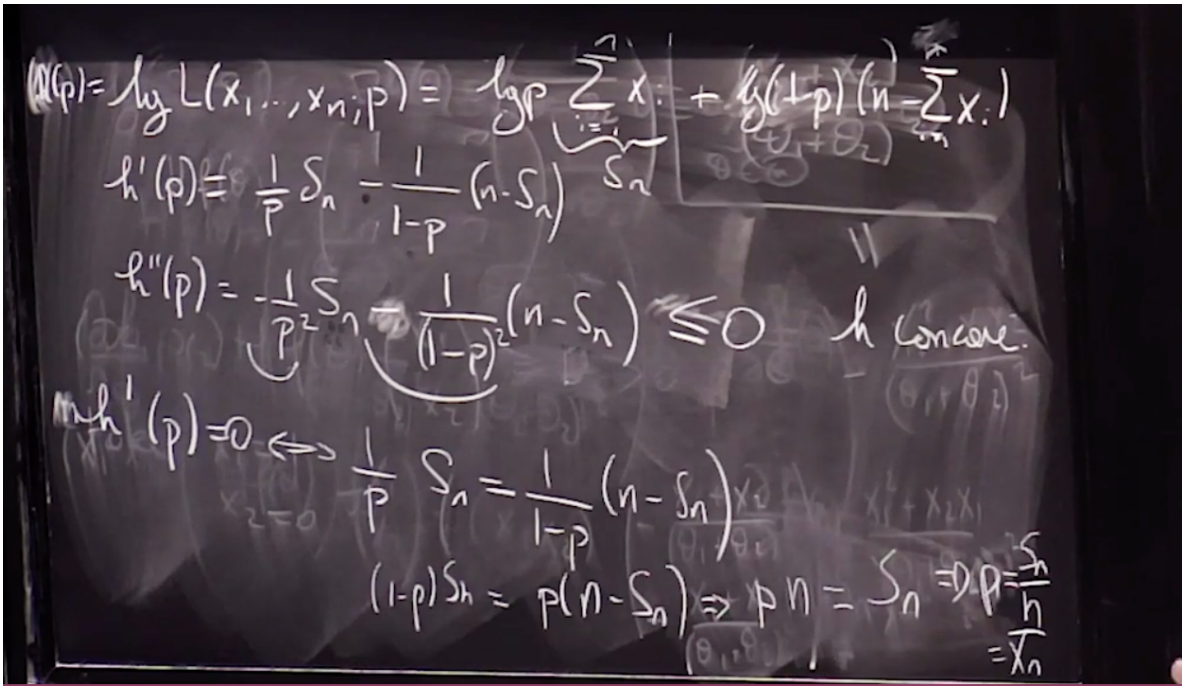
Solution:

Graphically, a strictly concave function that has a critical point looks like a hill. If you are to the right of the peak (*i.e.*, the maximum), then the hill is sloping downward. If you are to the left of the peak, then the hill is sloping upward.

More formally, a differentiable function that is strictly concave has a strictly decreasing derivative. The first derivative is zero at the critical point, so it must be positive to the left of the maximum (which implies the function is increasing) and negative to the right of the maximum (which implies the function is decreasing).

Thus the first choice, "To the left of $x_0 = 0$", is correct.

slide 27 for Bernoulli
gives Sn/n which is just the sample average (Xn_bar)
Sn is just short hand for sum of the x's over n



Maximum Likelihood Estimator of a Bernoulli Statistical Model I

3 points possible (graded)

In the next two problems, you will compute the MLE (maximum likelihood estimator) associated to a Bernoulli statistical model.

Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathrm{Ber}(p^*)$ for some unknown $p^* \in (0, 1)$. You construct the associated statistical model $(\{0, 1\}, \{\mathrm{Ber}(p)\}_{p \in (0,1)})$. Let L_n denote the likelihood of this statistical model. Recall that in the fourth problem "Likelihood of a Bernoulli Statistical Model" from two slides ago that you derived the formula

$$L_n(x_1, \ldots, x_n, p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n - \sum_{i=1}^{n} x_i}.$$

Oftentimes for computating the MLE it is more convenient to work with and optimize the **log-likelihood** $\ell(p) := \ln L_n(x_1, \ldots, x_n, p)$.

The derivative of the log-likelihood can be written

$$\frac{\partial}{\partial p} \ln L_n(x_1, \ldots, x_n, p) = A/p - (n - A)/B$$

where A can be expressed in terms of $\sum_{i=1}^{n} x_i$ and B can be expressed in terms of p. Fill in the blanks with the appropriate values for A and B

(Enter **Sigma_n** for entire sum $\sum_{i=1}^{n} x_i$).

$A =$ [ ] **Answer:** Sigma_n

$B =$ [ ] **Answer:** 1-p

For which p does $\frac{\partial}{\partial p}\ln L_n(x_1, \ldots, x_n, p) = 0$? Denote this critical point by $\hat{p}$.

- ◯ $\hat{p} = 0$
- ◯ $\hat{p} = 1$
- ◯ $\hat{p} = \sum_{i=1}^{n} x_i$
- ◯ $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$ ✔

Solution:

Observe that

$$\begin{aligned}\ln L_n(x_1, \ldots, x_n, p) &= \ln\left(p^{\sum_{i=1} x_i}(1-p)^{n-\sum_{i=1} x_i}\right) \\ &= \left(\sum_{i=1}^{n} x_i\right)\ln p + \left(n - \sum_{i=1}^{n} x_i\right)\ln(1-p).\end{aligned}$$

Taking the derivative with respect to p,

$$\frac{\partial}{\partial p}\ln L_n(x_1, \ldots, x_n, p) = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p}.$$

We set this to be 0 and solve for p:

$$\begin{aligned}\frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} &= 0 \Leftrightarrow \\ \frac{(1-p)\sum_{i=1}^{n} x_i - p\left(n - \sum_{i=1}^{n} x_i\right)}{p(1-p)} &= 0 \Leftrightarrow \\ \frac{\sum_{i=1}^{n} x_i - np}{p(1-p)} &= 0.\end{aligned}$$

Since the derivative blows up at $p = 0, 1$, we can assume $0 < p < 1$ and ignore the denominator for the purpose of solving for p. Hence $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the unique critical point of the log-likelihood.

Maximum Likelihood Estimator of a Bernoulli Statistical Model: Second Derivative Test

5 points possible (graded)

Setup:

As above, let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Ber}(p^*)$ for some unknown $p^* \in (0,1)$. You construct the associated statistical model $(\{0,1\}, \{\text{Ber}(p)\}_{p\in(0,1)})$. Let L_n denote the likelihood of this statistical model. Recall from a previous problem that

$$L_n(x_1, \ldots, x_n, p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

As stated, it will be more convenient to work with the **log-likelihood** $\ell(p) = \ln L_n(x_1, \ldots, x_n, p)$.

Question:

Next we will do the second derivative test to see if the critical point $\hat{p}$ obtained from the previous question is a local maximum. The second derivative of the log-likelihood can be written

$$\frac{\partial^2}{\partial p^2} \ln L_n(x_1, \ldots, x_n, p) = -\frac{C}{p^2} - \frac{n-C}{D}$$

where C depends on $\sum_{i=1}^n x_i$ and D depends on p. Fill in the blanks with the correct values of C and D.

(Type **Sigma_n** for the entire sum $\sum_{i=1}^n x_i$)

$C =$ [ ] **Answer:** Sigma_n

$D =$ [ ] **Answer:** (1-p)^2

Next we will test the endpoints of our optimization problem. Fill in the blanks with the correct values:
(Note that here we are working with the **likelihood**, *not* the **log-likelihood**)

$L_n(x_1, \ldots, x_n, 0) =$ [ ] **Answer:** 0.0

$L_n(x_1, \ldots, x_n, 1) =$ [ ] **Answer:** 0.0

What is the maximum likelihood estimator (MLE) $\hat{p}_n^{MLE}$ for the true parameter p^*?

- ○ 0
- ○ 1
- ○ $\sum_{i=1}^n X_i$
- ○ $\frac{1}{n}\sum_{i=1}^n X_i$ ✔

Solution:

The second derivative is

$$\frac{\partial}{\partial\theta}\left(\frac{\sum_{i=1}^{n} X_i}{p} - \frac{n-\sum_{i=1}^{n} X_i}{1-p}\right) = -\frac{\sum_{i=1}^{n} x_i}{p^2} - \frac{n-\sum_{i=1}^{n} x_i}{(1-p)^2}.$$

Since this expression is always negative, this implies that the critical point $\hat{p}$ is a **local maximum**.

Testing the endpoints we see

$$\begin{aligned} L_n(x_1, \ldots, x_n, 0) &= 0^{\sum_{i=1}^{n} x_i}(1)^{n-\sum_{i=1}^{n} x_i} = 0 \\ L_n(x_1, \ldots, x_n, 1) &= 1^{\sum_{i=1}^{n} x_i}(0)^{n-\sum_{i=1}^{n} x_i} = 0 \end{aligned}$$

Since the likelihood is non-negative, the endpoints are actually **global minima**.

Hence, the global maximum is achieved at $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Plugging in the random variables $X_1, \ldots, X_n$, we derive the MLE

$$\hat{p}_n^{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

which is precisely the **sample mean**.

Poisson example slide 27

Maximum Likelihood Estimator of a Poisson Statistical Model

3 points possible (graded)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Poiss}(\lambda^*)$ for some unknown $\lambda^* \in (0, \infty)$. You construct the associated statistical model $(\mathbb{N} \cup \{0\}, \{\text{Poiss}(\lambda)\}_{\lambda\in(0,\infty)})$. Recall that in the sixth question "Likelihood of a Poisson Statistical Model" two sections ago that you derived the formula

$$L_n(x_1, \ldots, x_n, \lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^{n} x_i}}{x_1! \cdots x_n!}$$

As in the previous question, we will work with the log-likelihood $\ell(\lambda) := \ln L_n(x_1, \ldots, x_n, \lambda)$.

The derivative of the log-likelihood can be written

$$\frac{\partial}{\partial\lambda}\ln L_n(x_1, \ldots, x_n, \lambda) = -n + \frac{A}{B}$$

where A depends on $\sum_{i=1}^{n} x_i$ and B depends on λ. Fill in the boxes with the correct expressions for A and B.

(Type **S_n** for $\sum_{i=1}^{n} x_i$ and **lambda** for λ.)

$A =$ [ ] **Answer:** S_n

$B =$ [ ] **Answer:** lambda

For the Poisson model, given fixed $x_1, \dots, x_n$, the function $\lambda \mapsto \ln L_n(x_1, \dots, x_n, \lambda)$ has a unique critical point $\hat{\lambda}$. You are allowed to assume that this critical point gives the expression for the MLE (*i.e.* given observations $x_1, \dots, x_n$, the global maximum of the log-likelihood is attained at $\hat{\lambda}$). Given this information, suppose you observe the data-set $X_1 = 2, X_2 = 3$, and $X_3 = 1$. What is $\hat{\lambda}_3^{\text{MLE}}(2, 3, 1)$?

$\hat{\lambda}_3^{\text{MLE}}(2, 3, 1) =$ **Answer:** 2.0

STANDARD NOTATION

Solution:

Observe that

$$\ln L_n(x_1, \dots, x_n, \lambda) = \ln\left(e^{-n\lambda}\frac{\lambda^{\sum_{i=1} x_i}}{x_1! \cdots x_n!}\right) = -n\lambda + (\sum_{i=1}^{n} x_i)\ln\lambda - \ln(x_1! \cdots x_n!)$$

Hence,

$$\frac{\partial}{\partial\lambda}\ln L_n(x_1, \dots, x_n, \lambda) = -n + \frac{\sum_{i=1}^{n} x_i}{\lambda}.$$

Setting this equal to 0, we recover the critical point

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

You are encouraged to perform the second derivative test and verify that this critical point is indeed a global maximum. (Don't forget to test the endpoints $\lambda = 0$ and $\lambda = \infty$ as well!)

This verifies that the MLE is

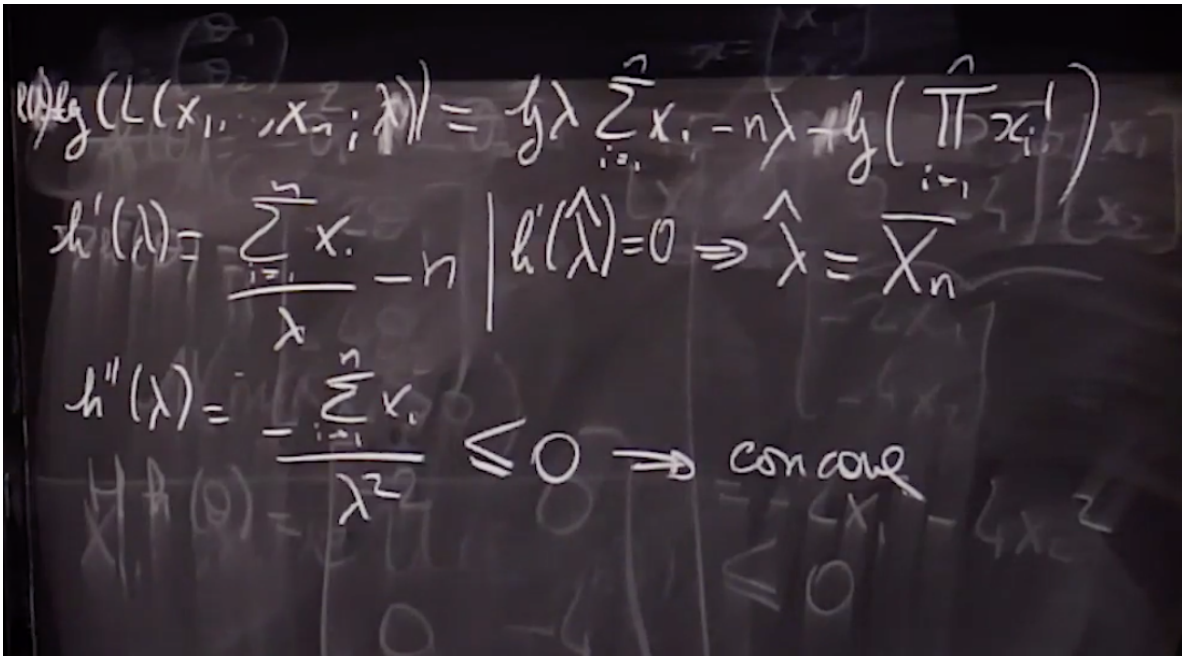
$$\hat{\lambda}_n^{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

which is again the **sample mean**.

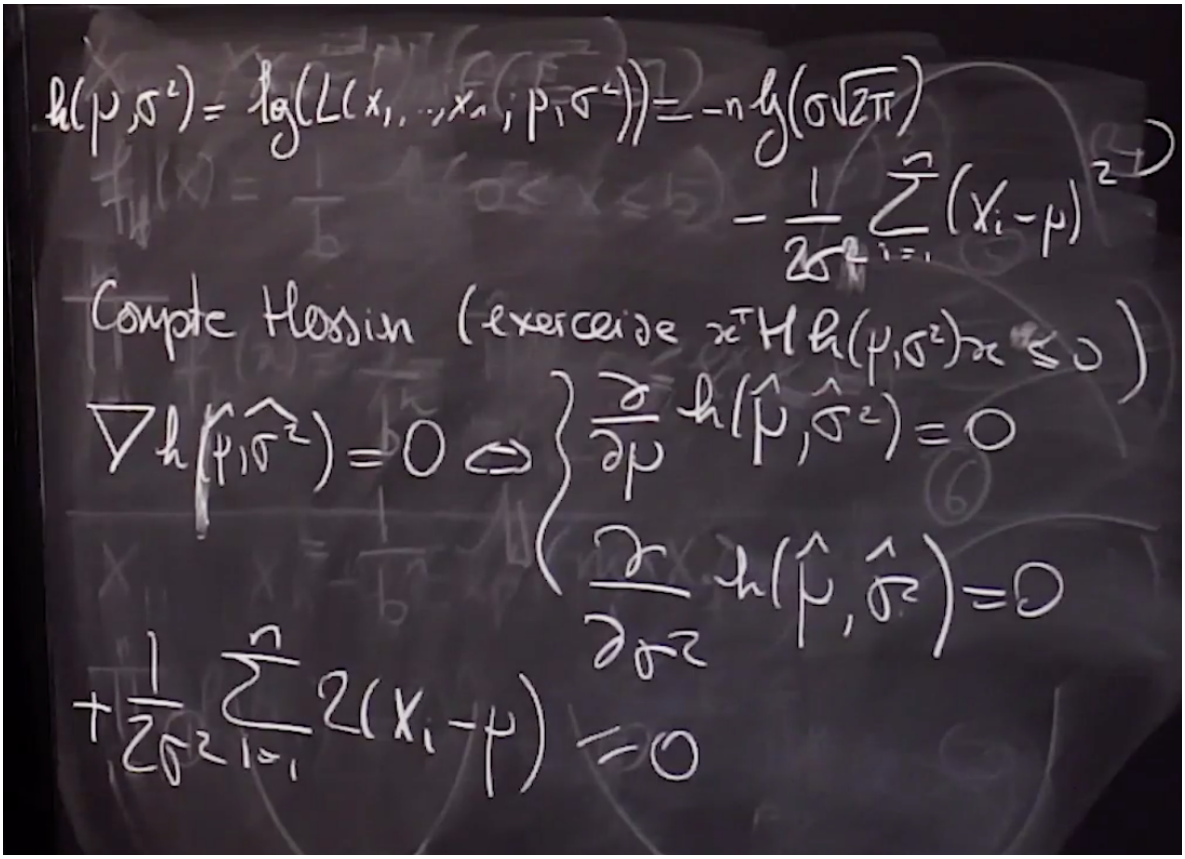
Hence, for the observations $X_1 = 2, X_2 = 3, X_3 = 1$, we get the estimate

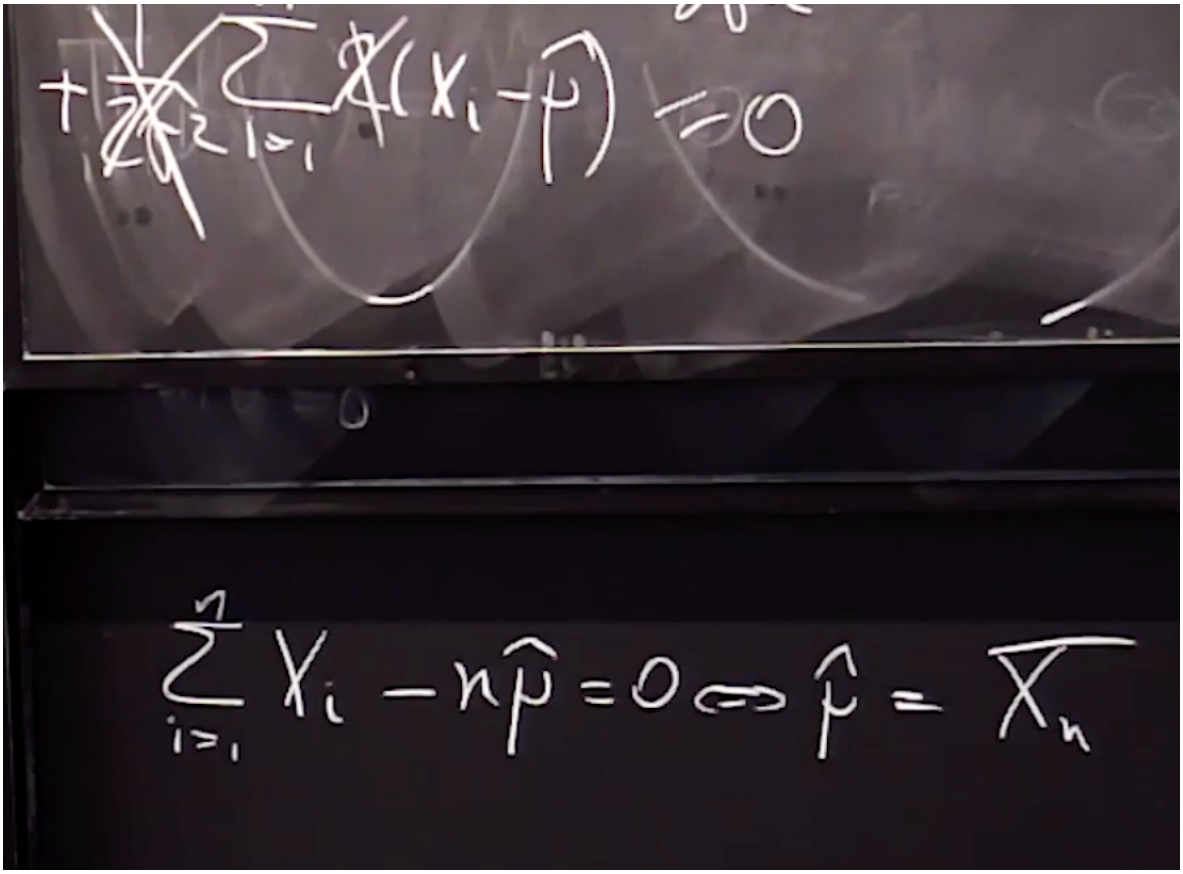
$$\hat{\lambda}_n^{MLE}(2, 3, 1) = \frac{1}{3}(2 + 3 + 1) = 2.$$

Remark: We also see for the Poisson model the conceptually nice fact that the maximum likelihood estimator is the sample mean.


$\log(L(x_1,\dots,x_n;\lambda)) = \log\lambda \sum_{i=1}^{n} x_i - n\lambda + \log\left(\prod_{i=1}^{n} x_i!\right)$
$h'(\lambda) = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n \quad | \quad h'(\hat{\lambda}) = 0 \Rightarrow \hat{\lambda} = \overline{X}_n$
$h''(\lambda) = -\frac{\sum_{i=1}^{n} x_i}{\lambda^2} \leq 0 \Rightarrow$ concave

Gaussian


$h(\mu,\sigma^2) = \log(L(x_1,\dots,x_n;\mu,\sigma^2)) = -n\log(\sigma\sqrt{2\pi})$
$-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$
Compte Hessian (exercice $x^T H h(\mu,\sigma^2) x \leq 0$)
$\nabla h(\mu,\hat{\sigma}^2) = 0 \Leftrightarrow \begin{cases} \frac{\partial}{\partial\mu} h(\hat{\mu},\hat{\sigma}^2) = 0 \\ \frac{\partial}{\partial\sigma^2} h(\hat{\mu},\hat{\sigma}^2) = 0 \end{cases}$
$+\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(x_i-\mu) = 0$



Maximum Likelihood Estimator of the variance a Gaussian Statistical Model with Mean Zero

3 points possible (graded)

Let $X_1, \ldots, X_n \overset{iid}{\sim} N(0, \tau^*)$ for some unknown variance τ^*. You construct the associated statistical model $(\mathbb{R}, \{N(0, \tau)\}_{\tau>0})$. Recall that in the last question from the previous slide, you derived the formula

$$L_n(x_1, \ldots, x_n, (\mu, \sigma^2)) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right).$$

Since we are given $\mu = 0$ and $\tau = \sigma^2$, we may rewrite this

$$L_n(x_1, \ldots, x_n, \tau) = \frac{1}{(\sqrt{2\pi\tau})^n} \exp\left(-\frac{1}{2\tau} \sum_{i=1}^{n} x_i^2\right).$$

As in the previous two questions, it will be more convenient to work with the log-likelihood $\ell(\tau) := \ln L_n$.

The derivative of the log-likelihood can be written

$$\frac{\partial}{\partial\tau}(\ln L_n(x_1, \ldots, x_n, \tau)) = -\frac{n}{A} + \frac{\sum_{i=1}^{n} x_i^2}{B}$$

where A and B both depend on τ. Find A and B.

where A and B both depend on τ. Find A and B.

(Type **tau** for τ.)

$A =$ **Answer:** 2*tau

$B =$ **Answer:** 2*tau^2

What is the maximum likelihood estimator for τ^*? (You are allowed to assume that the critical point found by setting $\frac{\partial}{\partial \tau} \ln L_n(x_1, \ldots, x_n, \tau) = 0$, treating $x_1, \ldots, x_n$ as fixed, gives the global maximum.)

Answer by entering the summand below in terms of the variable X_i.

(Type **X_i** for X_i.)

$\hat{\tau}_n^{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n}$ **Answer:** X_i^2

Solution:

Observe that

$$\begin{aligned} \ln L_n(x_1, \ldots, x_n, \tau) &:= \ln\left(\frac{1}{(\sqrt{2\pi\tau})^n} \exp\left(-\frac{1}{2\tau} \sum_{i=1}^{n} X_i^2\right)\right) \\ &= -\frac{n}{2} \ln(2\pi\tau) - \frac{1}{2\tau} \sum_{i=1}^{n} X_i^2 \end{aligned}$$

Taking derivatives, we get

$$\ln L_n(x_1, \ldots, x_n, \tau) = -\frac{n}{2\tau} + \frac{\sum_{i=1}^{n} X_i^2}{2\tau^2}.$$

Hence $A = 2\tau$ and $B = 2\tau^2$.

Rearranging, this is equal to 0 precisely when

$$\tau = \frac{1}{n} \sum_{i=1}^{n} X_i^2,$$

which is the “empirical second moment".

You are encouraged to perform the second derivative test to verify that $\tau = \frac{1}{n} \sum_{i=1}^{n} X_i^2$ is a local maximum. Moreover, it will be a *global* maximum because $\lim_{\tau \to 0} L_n(X_1, \ldots, X_n, \tau) = 0$ and $\lim_{\tau \to \infty} L_n(X_1, \ldots, X_n, \tau) = 0$.

Hence, we derive the formula for the MLE

$$\hat{\tau}_n^{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Remark: In this example, we want to estimate τ^*, the true variance, and we see the conceptually nice fact that the MLE is the **empirical second moment** $\frac{1}{n}\sum_{i=1}^{n} X_i^2$.

taking derivative wrt sigma^2 instead