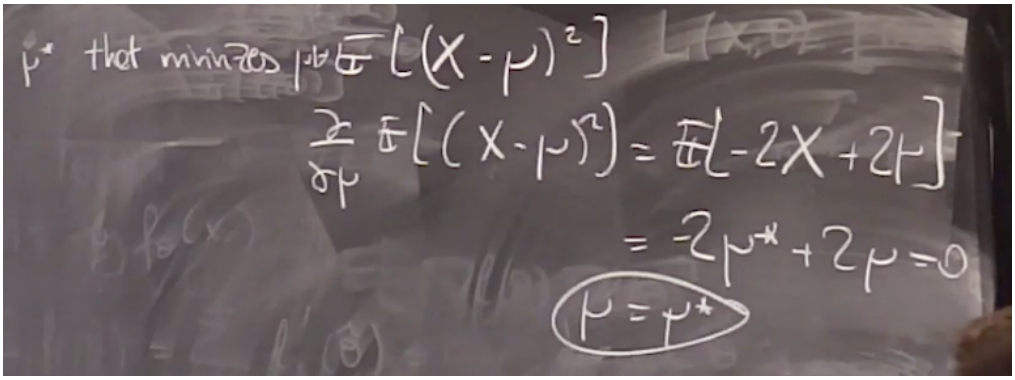
Unit 3 - part 4 M estimation

slide 48 first line
showing that minimising the expectation gives a mu that is equal to mu* (actual value)



M-estimation

Let $X_1, \ldots, X_n$ be i.i.d. with some unknown distribution $\mathbf{P}$ and an associated parameter μ^* on a sample space E. We make no modeling assumption that $\mathbf{P}$ is from any particular family of distributions.

An **M-estimator** $\widehat{\mu}$ of the parameter μ^* is the **argmin of an estimator of a function** $Q(\mu)$ **of the parameter** which satisfies the following:

- $Q(\mu) = \mathbb{E}[\rho(X, \mu)]$ for some function $\rho : E \times \mathcal{M} \to \mathbb{R}$, where $\mathcal{M}$ is the set of all possible values of the unknown true parameter $\mu*$;
- $Q(\mu)$ attains a **unique** minimum at $\mu = \mu^*$, in $\mathcal{M}$. That is, $\operatorname{argmin}_{\mu \in \mathcal{M}} Q(\mu) = \mu^*$.

In general, the goal is to find the **loss function** ρ such $Q(\mu) = \mathbb{E}[\rho(X, \mu)]$ has the properties stated above.

Note that the function $\rho(X, \mu)$ is in particular a function of the random variable X, and the expectation in $\mathbb{E}[\rho(X, \mu)]$ is to be taken against the **true distribution** $\mathbf{P}$ of X, with associated parameter value μ^*.

Because $Q(\mu)$ is an expectation, we can construct a (consistent) estimator of $Q(\mu)$ by replacing the expectation in its definition by the sample mean.

Example: multivariate mean as minimizer

Let $\mathbf{X} = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}$ be a continuous random vector with density $f : \mathbb{R}^2 \to \mathbb{R}$. Recall the mean of $\mathbf{X}$ is

$$\mathbb{E}[\mathbf{X}] = \begin{pmatrix} \mathbb{E}[X^{(1)}] \\ \mathbb{E}[X^{(2)}] \end{pmatrix}$$

Recall the square of the Euclidean norm function on $\mathbb{R}^2$:

$$\begin{aligned} \|\cdot\|^2 : \mathbb{R}^2 &\to \mathbb{R} \\ \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} &\mapsto (y_1)^2 + (y_2)^2. \end{aligned}$$

We now show that the (multivariate) mean of $\mathbf{X}$ satisfies:

$$\mathbb{E}[\mathbf{X}] = \operatorname{argmin}_{\vec{\mu} \in \mathbb{R}^2} \mathbb{E}\left[\|\mathbf{X} - \vec{\mu}\|^2\right].$$

(We will use subscripts to label the components of vectors below.)

First, expand $Q(\vec{\mu}) = \mathbb{E}\left[\|\mathbf{X} - \vec{\mu}\|^2\right]$ as an integral expression, and write down both partial derivatives $\frac{\partial Q}{\partial \mu_1}(\vec{\mu})$ and $\frac{\partial Q}{\partial \mu_2}(\vec{\mu})$:

$$\begin{aligned} \mathbb{E}\left[\|\mathbf{X} - \vec{\mu}\|^2\right] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left((x_1 - \mu_1)^2 + (x_2 - \mu_2)^2\right) f(x_1, x_2)\, dx_1 dx_2 \\ \implies \frac{\partial}{\partial \mu_1} \mathbb{E}\left[\|\mathbf{X} - \vec{\mu}\|^2\right] &= -2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_1 - \mu_1) f(x_1, x_2)\, dx_1 dx_2 \\ \frac{\partial}{\partial \mu_2} \mathbb{E}\left[\|\mathbf{X} - \vec{\mu}\|^2\right] &= -2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_2 - \mu_2) f(x_1, x_2)\, dx_1 dx_2. \end{aligned}$$

To find the argmin of $\mathbb{E}\left[\|\mathbf{X} - \vec{\mu}\|^2\right]$, we set both partial derivatives to 0, and obtain:

$$\operatorname{argmin}_{\vec{\mu} \in \mathbb{R}^2} \mathbb{E}\left[\|\mathbf{X} - \vec{\mu}\|^2\right] = \begin{pmatrix} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 f(x_1, x_2)\, dx_1 dx_2 \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_2 f(x_1, x_2)\, dx_1 dx_2 \end{pmatrix} = \begin{pmatrix} \mathbb{E}[X^{(1)}] \\ \mathbb{E}[X^{(2)}] \end{pmatrix}.$$

Concept check: M-estimators

1/1 point (graded)

Which of the following is true about M-estimation?
(Choose all that apply. Refer to the slides.)

- [x] M-estimation involves estimating some parameter of interest related to the underlying, unknown distribution (e.g. its mean, variance, or quantiles)
- [x] Maximum likelihood estimation is a special case of M-estimation.
- [x] Unlike maximum likelihood estimation and the method of moments, no statistical model needs to be assumed to perform M-estimation.
- [ ] M-estimation cannot be used for parametric statistical models.

✔

Solution:

We examine the choices in order.

- "M-estimation involves estimating some parameter of interested related to the underlying, unknown distribution (e.g. its mean, variance, or quantiles)" is correct. This is precisely the goal of M-estimation, as stated in the slides. It is a flexible approach that applies even outside of parametric statistical models.
- "Maximum likelihood estimation is a special case of M-estimation." is correct. If we set the loss function to be the negative log-likelihood, then the same optimization problem defining the MLE is the one considered for the M-estimator associated to this loss function.
- "Unlike maximum likelihood estimation and the method of moments, no statistical model needs to be assumed to perform M-estimation." is correct. As stated above, M-estimation is a flexible approach that can used to approximate relevant quantities of interest to a distribution, such as its moments.
- "M-estimation cannot be used for parametric statistical models." is incorrect. M-estimation can be used in both a parametric and non-parametric context, though in this lecture, we will only see it applied in parametric examples.

Relating M-estimation and Maximum Likelihood Estimation

0/1 point (graded)

Let $(E, \{\mathbf{P}_\theta\}_{\theta\in\Theta})$ denote a discrete statistical model and let $X_1, \ldots, X_n \overset{iid}{\sim} \mathbf{P}_{\theta^*}$ denote the associated statistical experiment, where θ^* is the true, unknown parameter. Suppose that $\mathbf{P}_\theta$ has a probability mass function given by p_θ. Let $\widehat{\theta}_n^{\text{MLE}}$ denote the maximum likelihood estimator for θ^*.

The maximum likelihood estimator can be expressed as an M-estimator– that is,

$$\widehat{\theta}_n^{\text{MLE}} = \text{argmin}_{\theta\in\Theta} \frac{1}{n} \sum_{i=1}^{n} \rho\left(X_i, \theta\right)$$

for some function ρ.

Which of the following represents the correct choice of the function ρ so that the equation above is satisfied?

- ◯ $-\ln p_\theta\left(X_i\right)$ ✔
- ◉ $\ln p_\theta\left(X_i\right)$
- ◯ $p_\theta\left(X_i\right)$
- ◯ None of the above.

✖

Solution:

The correct response is "$-\ln p_\theta\left(X_i\right)$". Recall that the MLE is defined by

$$\widehat{\theta}_n^{\text{MLE}} = \text{argmax}_{\theta\in\Theta} \frac{1}{n} \sum_{i=1}^{n} \ln p_\theta\left(X_i\right).$$

By symmetry, we also have,

$$\widehat{\theta}_n^{\text{MLE}} = \text{argmin}_{\theta\in\Theta} \frac{1}{n} \sum_{i=1}^{n} -\ln p_\theta\left(X_i\right).$$

Indeed, setting $\rho\left(x, \theta\right) = -\ln p_\theta\left(x\right)$, we recover the maximum likelihood estimator.

Median as a Minimizer

2/3 points (graded)

Assume that X is a continuous random variable with density $f : \mathbb{R} \to \mathbb{R}$. Then a **median** of X is defined to be any point $\text{med}(X) \in \mathbb{R}$ such that

$$P(X > \text{med}(X)) = P(X < \text{med}(X)) = \frac{1}{2}.$$

(Recall that for a continuous distribution, $P(X > \text{med}(X)) = P(X \geq \text{med}(X))$.) Note: A median of a distribution is *not necessarily* unique.)

In this problem, you will show that any median satisfies

$$\text{med}(X) = \text{argmin}_{\mu \in \mathbb{R}} \mathbb{E}[|X - \mu|].$$

Which of the following correctly expresses $\mathbb{E}[|X - \mu|]$ in terms of the density $f(x)$?

- $\int_{-\infty}^{\infty} x f(x)\, dx - \mu$
- $\int_{-\infty}^{\infty} x f(x)\, dx - \mu \left(- \int_{\mu}^{\infty} f(x)\, dx + \int_{-\infty}^{\mu} f(x)\, dx \right)$
- $\int_{\mu}^{\infty} x f(x)\, dx - \int_{-\infty}^{\mu} x f(x)\, dx - \mu$
- $\int_{\mu}^{\infty} x f(x)\, dx - \int_{-\infty}^{\mu} x f(x)\, dx - \mu \left(\int_{\mu}^{\infty} f(x)\, dx - \int_{-\infty}^{\mu} f(x)\, dx \right)$ ✔

✖

Let $\mathcal{Q}(\mu) = \mathbb{E}[|X - \mu|]$ denote the expression obtained in the previous question. Then $\mathcal{Q}(\mu)$ consists of a sum of terms, each of which can be differentiated with respect to μ.

What is $\mathcal{Q}'(\mu) = \frac{d}{d\mu}\mathcal{Q}(\mu)$?

Hint: Use the product rule and the fundamental theorem of calculus.

- ○ 1
- ◉ $\int_{-\infty}^{\mu} f(x)\ dx - \int_{\mu}^{\infty} f(x)\ dx$
- ○ $4\mu f(\mu) + \int_{-\infty}^{\mu} f(x)\ dx - \int_{\mu}^{\infty} f(x)\ dx$
- ○ $4\mu f(\mu) + 1$

✔

Using your response from the previous question and the definition of median, what is $\mathcal{Q}'(\text{med}(X))$?

- ◉ 0
- ○ 1
- ○ $4\text{med}(X)\ f(\text{med}(X)) + 1$
- ○ Cannot be determined.

✔

Solution:

For the first question, we have

$$\begin{aligned}\mathbb{E}\left[|X-\mu|\right] &= \int_{-\infty}^{\infty} |x-\mu| f(x)\ dx \\ &= \int_{\mu}^{\infty} (x-\mu) f(x)\ dx + \int_{-\infty}^{\mu} (-x+\mu) f(x)\ dx \\ &= \int_{\mu}^{\infty} x f(x)\ dx - \int_{-\infty}^{\mu} x f(x)\ dx - \mu\left(\int_{\mu}^{\infty} f(x)\ dx - \int_{-\infty}^{\mu} f(x)\ dx\right)\end{aligned}$$

Therefore, "$\int_{\mu}^{\infty} x f(x)\ dx - \int_{-\infty}^{\mu} x f(x)\ dx - \mu\left(\int_{\mu}^{\infty} f(x)\ dx - \int_{-\infty}^{\mu} f(x)\ dx\right)$" is the correct answer to the first question.

For the second question, we differentiate the previous answer term by term. We have, by the fundamental theorem of calculus and the product rule that

$$\begin{aligned}\frac{d}{d\mu}\left(\int_{\mu}^{\infty} x f(x)\ dx\right) &= -\mu f(\mu) \\ \frac{d}{d\mu}\left(-\int_{-\infty}^{\mu} x f(x)\ dx\right) &= -\mu f(\mu) \\ \frac{d}{d\mu}\left(-\mu\left(\int_{\mu}^{\infty} f(x)\ dx - \int_{-\infty}^{\mu} f(x)\ dx\right)\right) &= -\int_{\mu}^{\infty} f(x)\ dx + \int_{-\infty}^{\mu} f(x)\ dx + 2\mu f(\mu).\end{aligned}$$

Adding these terms, we have cancellations, yielding

$$\frac{d}{d\mu}\mathcal{Q}(\mu) = -\int_{\mu}^{\infty} f(x)\ dx + \int_{-\infty}^{\mu} f(x)\ dx.$$

Therefore, the correct response to the second question is "$\int_{-\infty}^{\mu} f(x)\ dx - \int_{\mu}^{\infty} f(x)\ dx$".

For the third question, by definition, the median $\operatorname{med}(X)$ of X is a real number that satisfies $P(X > \operatorname{med}(X)) = P(X < \operatorname{med}(X))$. Therefore,

$$\mathcal{Q}'(\operatorname{med}(X)) = \int_{-\infty}^{\operatorname{med}(X)} f(x)\ dx - \int_{\operatorname{med}(X)}^{\infty} f(x)\ dx = P(X < \operatorname{med}(X)) - P(X > \operatorname{med}(X)) = 0.$$

The correct response is "0".

Quantile as a Minimizer

7 points possible (graded)

Recall from the lecture that the **check function** is defined as

$$C_\alpha(x) = \begin{cases} -(1-\alpha)x & \text{if } x < 0 \\ \alpha x & \text{if } x \geq 0. \end{cases}$$

Assume that X is a continuous random variable with density $f : \mathbb{R} \to \mathbb{R}$. Define the **$\alpha$-quantile** of X to be $Q_X(\alpha) \in \mathbb{R}$ such that

$$\mathbf{P}(X \leq Q_X(\alpha)) = \alpha.$$

(Here we have used a different convention of the definition of the quantile function from before, where for a standard normal distribution, q_α is such that $P(X>q_\alpha) = \alpha$.)

Just like for the median, whether Q_α is unique depends on the distribution.

In this problem, you will convince yourself that any α-quantile of X satisfies

$$Q_X(\alpha) = \operatorname{argmin}_{\mu \in \mathbb{R}} \mathbb{E}[C_\alpha(X-\mu)].$$

First, compute $\mathbb{E}[C_\alpha(X-\mu)]$. Answer by entering the coefficients A, B, C, D in terms of α and μ in the expression below:

$$\begin{aligned} \mathbb{E}[C_\alpha(X-\mu)] = {} & A\int_{-\infty}^{\mu} x f(x)\,dx + B\int_{\mu}^{\infty} x f(x)\,dx \\ & + C\int_{-\infty}^{\mu} f(x)\,dx + D\int_{\mu}^{\infty} f(x)\,dx. \end{aligned}$$

$A =$ alpha-1 ✔ **Answer:** alpha-1 $B =$ alpha ✔ **Answer:** alpha

$\alpha - 1$ α

$C =$ alpha-1 ✖ **Answer:** -(alpha-1)*mu $D =$ alpha+mu ✖ **Answer:** -alpha*mu

$\alpha - 1$ $\alpha + \mu$

Second, let $F(\mu) = \mathbb{E}[C_\alpha(X - \mu)]$ denote the expression obtained in the question above. Find $F'(\mu)$. Answer by entering the coefficients E, G, H, in terms of α and μ below:

$$F'(\mu) = (\mathbb{E}[C_\alpha(X - \mu)])' = E + G(\mu f(\mu)) + H\int_{-\infty}^{\mu} f(x)\,dx.$$

$E =$ mu ✖ **Answer:** -alpha

μ

$G =$ 0 ✔ **Answer:** 0

0

$H =$ 1 ✔ **Answer:** 1

1

Finally, set $F'(\mu) = 0$ to find $\operatorname{argmin}_{\mu\in\mathbb{R}} F(\mu) = \operatorname{argmin}_{\mu\in\mathbb{R}} \mathbb{E}[C_\alpha(X - \mu)]$. (There is no answer box for this question.)

Solution:

Given the check function center about μ:

$$C_\alpha(x - \mu) = \begin{cases} -(1-\alpha)(x-\mu) & \text{if } x < \mu \\ \alpha(x-\mu) & \text{if } x \geq \mu, \end{cases}$$

compute $F(\mu) = \mathbb{E}[C_\alpha(X - \mu)]$:

$$\begin{aligned} F(\mu) = \mathbb{E}[C_\alpha(X-\mu)] &= -\int_{-\infty}^{\mu}(1-\alpha)(x-\mu)f(x)\,dx + \int_{\mu}^{\infty}\alpha(x-\mu)f(x)\,dx \\ &= -(1-\alpha)\int_{-\infty}^{\mu} xf(x)\,dx + \alpha\int_{\mu}^{\infty} xf(x)\,dx \\ &\quad + (1-\alpha)\mu\int_{-\infty}^{\mu} f(x)\,dx - \alpha\mu\int_{\mu}^{\infty} f(x)\,dx. \end{aligned}$$

Then, the derivative of F with respect to μ is:

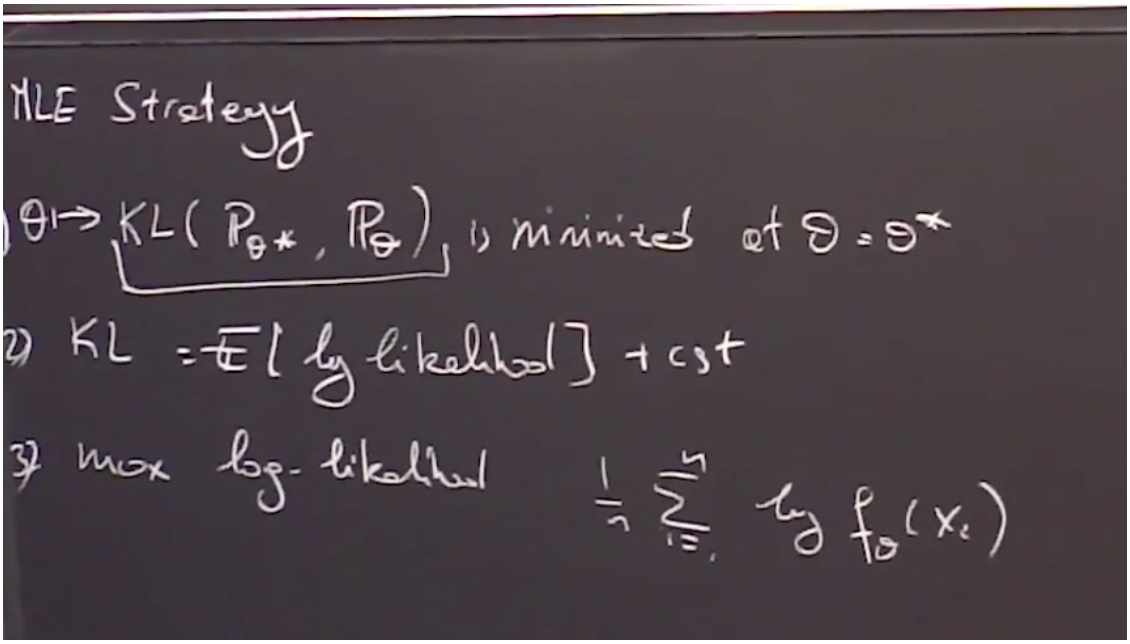
$$\begin{aligned}
F'(\mu) = \frac{d}{d\mu}F(\mu) &= -(1-\alpha)\frac{d}{d\mu}\int_{-\infty}^{\mu} x f(x)\,dx + \alpha\frac{d}{d\mu}\int_{\mu}^{\infty} x f(x)\,dx \\
&\quad + (1-\alpha)\frac{d}{d\mu}\left(\mu\int_{-\infty}^{\mu} f(x)\,dx\right) - \alpha\frac{d}{d\mu}\left(\mu\int_{\mu}^{\infty} f(x)\,dx\right) \\
&= -(1-\alpha)(\mu f(\mu)) + \alpha(-\mu) f(\mu)) \\
&\quad + (1-\alpha)\left(\int_{-\infty}^{\mu} f(x)\,dx + \mu f(\mu)\right) - \alpha\left(\int_{\mu}^{\infty} f(x)\,dx - \mu f(\mu)\right) \\
&= (1-\alpha)\int_{-\infty}^{\mu} f(x)\,dx - \alpha\int_{\mu}^{\infty} f(x)\,dx \\
&= (1-\alpha)\left(\int_{-\infty}^{\mu} f(x)\,dx\right) - \alpha\left(1 - \int_{-\infty}^{\mu} f(x)\,dx\right) \\
&= \left(\int_{-\infty}^{\mu} f(x)\,dx\right) - \alpha.
\end{aligned}$$

Setting $F'(\mu) = 0$ yields

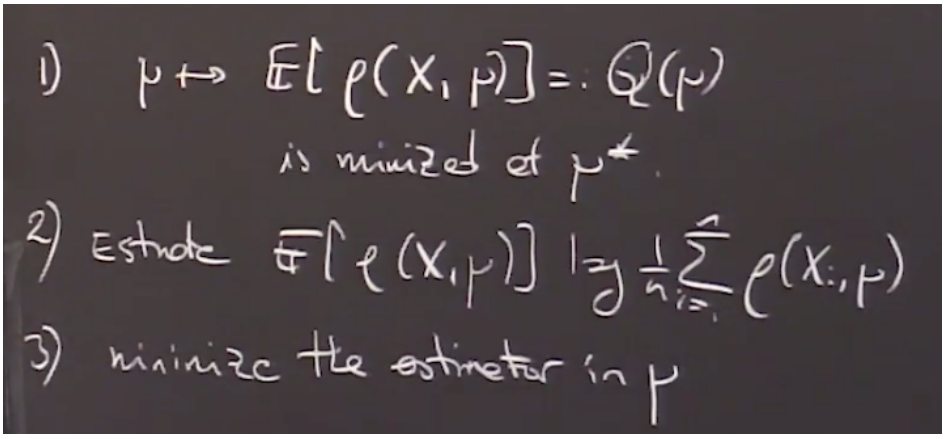
$$\int_{-\infty}^{\mu} f(x)\,dx = \alpha.$$

Hence, $\operatorname{argmin}_{\mu\in\mathbb{R}} F(\mu)$ is an α-quantile of X.

MLE strategy



M estimation strategy



Concept check: Defining M-estimators

1/1 point (graded)

Suppose we have access to a distribution $\mathbf{P}$ which has an unknown parameter μ^* that we would like to estimate from samples $X_1, \ldots, X_n \overset{iid}{\sim} \mathbf{P}$. Suppose we have a **loss function** $\rho(x, \mu)$ with the property that

$$\mu^* = \operatorname{argmin}_{\mu \in \mathbb{R}} \mathbb{E}_{X \sim \mathbf{P}} [\rho(X, \mu)] .$$

What commonly used statistical trick is used to define an M-estimator? (Refer to the slides.)

- ◯ Using the KL divergence instead of TV distance.
- ◯ The method of moments.
- ◉ Replacing expectations with averages. ✔

Solution:

The correct response is "Replacing expectations with averages." Indeed, we have that the equation

$$\mu^* = \text{argmin}_{\mu \in \mathbb{R}} \mathbb{E}_{X \sim \mathbf{P}} \left[\rho \left(X, \mu \right) \right]$$

becomes

$$\widehat{\mu} = \text{argmin}_{\mu \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \left[\rho \left(X_i, \mu \right) \right]$$

upon replacing the expectation by an average over the sample. Here, $\widehat{\mu}$ is precisely the M-estimator associated with $\rho \left(x, \mu \right)$.

The response "Using the KL divergence instead of TV distance." is incorrect. Rather, the KL divergence was used specifically in the context of maximum likelihood estimation. It does not play a role in the context of M-estimation.

The response "The method of moments." is also incorrect. The method of moments is a tool for parameter estimation which is distinct from M-estimation. The method of moments is not what is used to define an M-estimator.

The $\mathbf{J}$ and $\mathbf{K}$ matrices:

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be i.i.d. random vector in $\mathbb{R}^k$ with some unknown distribution $\mathbf{P}$ with some associated parameter $\vec{\mu}^* \in \mathbb{R}^d$ on some sample space E. Let $Q \left(\vec{\mu} \right) = \mathbb{E} \left[\rho \left(\mathbf{X}, \vec{\mu} \right) \right]$ for some function $\rho : E \times \mathcal{M} \to \mathbb{R}$, where $\mathcal{M}$ is the set of all possible values of the unknown true parameter $\vec{\mu}^*$.

Then the matrices $\mathbf{J}$ and $\mathbf{K}$ are defined as

$$\mathbf{J} = \mathbb{E} \left[\mathbf{H} \rho \right] = \mathbb{E} \left[\begin{pmatrix} \frac{\partial^2 \rho}{\partial \mu_1 \partial \mu_1} (\mathbf{X}_1, \vec{\mu}) & \cdots & \frac{\partial^2 \rho}{\partial \mu_1 \partial \mu_d} (\mathbf{X}_1, \vec{\mu}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \rho}{\partial \mu_d \partial \mu_1} (\mathbf{X}_1, \vec{\mu}) & \cdots & \frac{\partial^2 \rho}{\partial \mu_d \partial \mu_d} (\mathbf{X}_1, \vec{\mu}) \end{pmatrix} \right] \qquad (d \times d)$$

$$\mathbf{K} = \text{Cov} \left[\nabla \rho \left(\mathbf{X}_1, \vec{\mu} \right) \right] = \text{Cov} \left[\begin{pmatrix} \frac{\partial \rho}{\partial \mu_1} (\mathbf{X}_1, \vec{\mu}) \\ \vdots \\ \frac{\partial \rho}{\partial \mu_d} (\mathbf{X}_1, \vec{\mu}) \end{pmatrix} \right] \qquad (d \times d) .$$

In one dimension, i.e. $d = 1$, the matrices reduce to the following:

$$J \left(\mu \right) = \mathbb{E} \left[\frac{\partial^2 \rho}{\partial \mu^2} (X_1, \mu) \right]$$

$$K \left(\mu \right) = \text{Var} \left[\frac{\partial \rho}{\partial \mu} (X_1, \mu) \right]$$

Concept Check: M-estimators vs. Maximum Likelihood Estimation

1/1 point (graded)

Let ρ denote a loss function, and let $X_1, \ldots, X_n \overset{iid}{\sim} \mathbf{P}$. Let $\widehat{\mu}$ denote the M-estimator for some unknown parameter $\mu^* = \operatorname{argmin}_{\mu \in \mathbb{R}} \mathbb{E}\left[\rho\left(X_1, \mu\right)\right] \in \mathbb{R}$ associated with $\mathbf{P}$. (Here we are assuming that μ^* is a one-dimensional parameter.)

Consider the following functions

$$J\left(\mu\right) = \mathbb{E}\left[\frac{\partial^2 \rho}{\partial \mu^2}\left(X_1, \mu\right)\right]$$
$$K\left(\mu\right) = \operatorname{Var}\left[\frac{\partial \rho}{\partial \mu}\left(X_1, \mu\right)\right]$$

Which of the following statements are true? (Choose all that apply.)

- [ ] It is always true that $J\left(\mu\right) = K\left(\mu\right)$.
- [x] $J\left(\mu\right) = K\left(\mu\right)$ when ρ is the negative log-likelihood– in this case, both of these functions are equal to the Fisher information.
- [x] Under some technical conditions, the functions $J\left(\mu\right)$ and $K\left(\mu\right)$ determine the asymptotic variance of the M-estimator $\widehat{\mu}$.

✔

Solution:

- The response "It is always true that $J\left(\mu\right) = K\left(\mu\right)$." is incorrect. In general, the functions $J\left(\mu\right)$ and $K\left(\mu\right)$ will not be equal to each other. For example, if the loss function is given in terms of Huber's loss (as we will see later in this lecture), $J\left(\mu\right) \neq K\left(\mu\right)$.
- The choice "$J\left(\mu\right) = K\left(\mu\right)$ when ρ is the negative log-likelihood– in this case, both of these functions are equal to the Fisher information." is correct. In the special case where $\rho\left(x, \mu\right)$ is defined to be the negative log-likelihood of the statistical model, then it is true that $J\left(\mu\right) = K\left(\mu\right)$. This was derived in Lecture 11.
- The choice "Under some technical conditions, the functions $J\left(\mu\right)$ and $K\left(\mu\right)$ determine the asymptotic variance of the M-estimator $\widehat{\mu}$." is correct. This is content of the theorem on the slide "Asymptotic Normality," which shows that the asymptotic variance of $\widehat{\mu}_n$, assuming some hypotheses, is given by $J(\mu^*)^{-1} K\left(\mu^*\right) J(\mu^*)^{-1}$.

Remark on signs:
Let us match the signs in the definition of $\mathbf{J}$ and $\mathbf{K}$ with those in the definition of Fisher information. For maximum likelihood estimation,

$$\rho_n\left(\theta\right) := \rho\left(\mathbf{X}_1, \ldots, \mathbf{X}_n, \theta\right) = -\ell_n\left(\theta\right) \qquad \text{where } \ell_n\left(\theta\right) = \ln L_n\left(\mathbf{X}_1, \ldots, \mathbf{X}_n, \theta\right).$$

For this particular loss function ρ, the $\mathbf{J}$ and $\mathbf{K}$ matrices are

$$\mathbf{J} = \mathbb{E}\left[\mathbf{H}\rho_1\left(\theta\right)\right] = -\mathbb{E}\left[\mathbf{H}\ell_1\left(\theta\right)\right]$$
$$\mathbf{K} = \operatorname{Cov}\left[\nabla\rho_1\left(\theta\right)\right] = \operatorname{Cov}\left[-\nabla\ell_1\left(\theta\right)\right] = \operatorname{Cov}\left[\nabla\ell_1\left(\theta\right)\right] \qquad \left(\operatorname{Cov}\left[\mathbf{Y}\right] = \operatorname{Cov}\left[-\mathbf{Y}\right] \text{ for any random vector } \mathbf{Y}.\right.$$

Both of these matrices equals the Fisher information matrix.

Asymptotic normality of the M-estimators

1/3 points (graded)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathbf{P}$. Let $\rho(x, \mu)$ denote a loss function satisfying

$$\mu^* = \operatorname{argmin}_{\mu \in \mathbb{R}} \mathbb{E}\left[\rho(X_1, \mu)\right]$$

where $\mu^* \in \mathbb{R}$ is some unknown one-dimensional parameter associated with $\mathbf{P}$ that we would like to estimate. Let

$$\begin{aligned} J(\mu) &= \mathbb{E}\left[\frac{\partial^2 \rho}{\partial \mu^2}(X_1, \mu)\right] \\ K(\mu) &= \operatorname{Var}\left[\frac{\partial \rho}{\partial \mu}(X_1, \mu)\right] \end{aligned}$$

You construct the M-estimator $\widehat{\mu}_n$ associated ρ.

Assuming that the conditions for the asymptotic normality of this M-estimator hold, we have

$$\sqrt{n}\frac{\widehat{\mu}_n - \mu^*}{\sqrt{J(\mu^*)^{-2} K(\mu^*)}} \xrightarrow[n \to \infty]{(d)} Q$$

for some distribution Q.

What is Q?

- ◯ Poisson with mean 1.
- ◯ Exponential with mean 1.
- ◉ Standard normal.
- ◯ $\mathcal{N}(0, \sigma^2)$ for some unknown parameter σ^2.

✔

Let q_α denote the α-quantile of the distribution Q. For what value of q_α is it true that

$$\mu^* \in \left[\widehat{\mu}_n - q_\alpha \sqrt{\frac{J(\mu^*)^{-2} K(\mu^*)}{n}}, \widehat{\mu}_n + q_\alpha \sqrt{\frac{J(\mu^*)^{-2} K(\mu^*)}{n}} \right]$$

with probability 95% as $n \to \infty$?

$q_\alpha =$ 1.614 ✖ **Answer:** 1.96

Let

$$\mathcal{I} := \left[\widehat{\mu}_n - q_\alpha \sqrt{\frac{J(\mu^*)^{-2} K(\mu^*)}{n}}, \widehat{\mu}_n + q_\alpha \sqrt{\frac{J(\mu^*)^{-2} K(\mu^*)}{n}} \right]$$

denote the interval in the previous question.

Is $\mathcal{I}$ an asymptotic confidence interval for μ^* of confidence level 95%?

- (●) Yes, because the previous question solves for q_α so that this holds.
- ( ) Yes, because of the asymptotic normality of $\widehat{\mu}_n$.
- ( ) No, because we did not define a statistical model for this problem.
- ( ) No, because the endpoints of $\mathcal{I}$ depend on the true parameter. ✔

✖

Solution:

For the first question, the correct response is "Standard normal." Referring to the theorem regarding the asymptotic normality of the M-estimators, we see that the asymptotic variance of $\widehat{\mu_n}$ is $J(\mu^*)^{-2}K(\mu^*)$. Hence,

$$\sqrt{n}\frac{\widehat{\mu_n}-\mu^*}{\sqrt{J(\mu^*)^{-2}K(\mu^*)}} \xrightarrow[(d)]{n\to\infty} \mathcal{N}(0,1)\,.$$

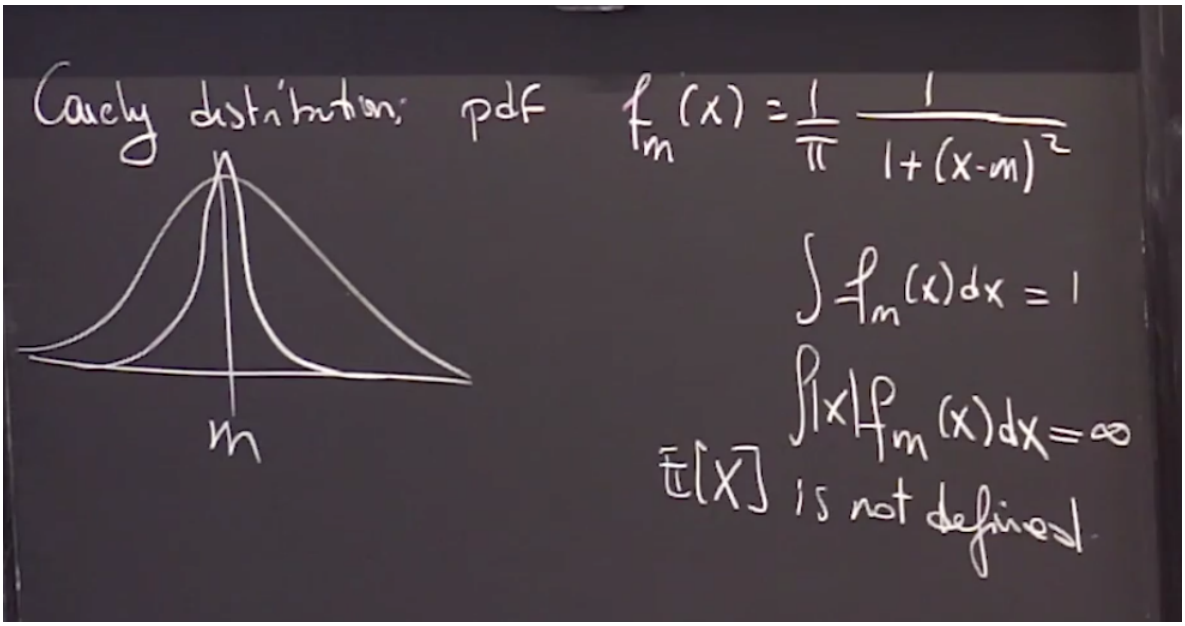
For the second question, the correct response is "1.96". By the previous equation,

$$P\left(\sqrt{n}\left|\frac{\widehat{\mu}_n-\mu^*}{\sigma}\right| \geq q_{0.025}\right) = 1-P\left(\mu^* \in \left[\widehat{\mu}_n - q_{0.025}\sqrt{\frac{J(\mu^*)^{-2}K(\mu^*)}{n}}, \widehat{\mu}_n + q_{0.025}\sqrt{\frac{J(\mu^*)^{-2}K(\mu^*)}{n}}\right]\right) =$$

where $q_{0.025} = 1.96$ is the 2.5%-quantile of a standard Gaussian.

For the third question, the correct response is "No, because the endpoints of $\mathcal{I}$ depend on the true parameter." By definition, the endpoints of a confidence interval should be estimators, and this is not the case for $\mathcal{I}$ because $K^{-1}(\mu^*)$ and $J(\mu^*)$ depend on the true parameter.

slide 54
Cauchy distribution
looks like gaussian but the tails decay at 1/x^2 not exponentially



E[x] not defined so method of moments can't be use
can replace the mean with the median

Robust Statistics and the Median

4/4 points (graded)

In this problem, you will see how some estimators are more resilient to corruptions or mistakes in the data than others. Such estimators are referred to as **robust** .

Researchers in a lab observe the following data set

$$\{0.5, 1.8, -2.3, 0.86, 0.32\}.$$

In reality, these numbers are generated from the standard normal distribution $\mathcal{N}(0, 1)$, but as in most statistical problems, this distribution is unknown to the researchers. Their goal is to estimate the mean of this unknown distribution, and they will try two different statistics for doing so: the sample mean and the median.

What is the sample mean $\overline{X}_5$ of the data set?

0.236 ✔

What is the median M_5 of the data set? (For a discrete data set, we define the median to be the middle number when the data set is sorted from smallest to largest.)

0.5 ✔

Suppose now that one of the lab assistants makes a typo, and instead records the data set

$$\{0.5, 1.8, -23, 0.86, 0.32\}.$$

(Note that the number -2.3 from the sample has been erroneously changed to -23.)

What is the sample mean of the new, corrupted data set?

-3.904 ✔ **Answer:** -3.9

What is the median of the new, corrupted data set?

0.5 ✔ **Answer:** 0.5

Solution:

The mean of the first data set is

$$\frac{-2.3 + 0.5 + 1.8 + 0.86 + 0.32}{5} \approx 0.236$$

From smallest to largest, the data set reads

$$-2.3, 0.32, 0.5, 0.86, 1.8$$

, so 0.5 is the median.

The mean of the second data set is

$$\frac{-23 + 0.5 + 1.8 + 0.86 + 0.32}{5} \approx -3.90.$$

From smallest to largest, the data set reads,

$$-23, 0.32, 0.5, 0.86, 1.8$$

so 0.5 is still the median.

Remark: This simple example illustrates how the median is in general a more *robust* estimator than the mean. That is, errors or corruptions in the data set have a limited effect on how much the median changes. However, the same is not true for the mean.

Cauchy distribution I

2/2 points (graded)

The **Cauchy distribution** is a continuous distribution with a parameter m, known as the **location parameter** , and with density given by

$$f_m(x) = \frac{1}{\pi} \frac{1}{1 + (x - m)^2}.$$

Suppose X is a random variable distributed as the Cauchy distribution.

What is $\mathbb{E}[X]$?

- ◯ 0
- ◯ $\frac{1}{2}$
- ◯ 1
- ◉ Does not exist. ✔

Recall that the **median** of a continuous distribution is any number M such that $P(X > M) = P(X < M) = 1/2$. For the Cauchy distribution, it turns out that the median is unique.

If the location parameter is set to be $m = 1/2$, what is $\text{med}(X)$?

1/2 ✔ **Answer:** 0.5

Solution:

The correct answer to the first question is "Does not exist". To show this, let us temporarily set the location parameter to be $m = 0$. If we were to try to compute the mean, we would write down the integral

$$\int_{-\infty}^{\infty} \frac{1}{\pi} \cdot \frac{x}{1+x^2}\, dx.$$

However, this improper integral does not converge. The antiderivative is

$$\frac{1}{2\pi} \ln\left(1+x^2\right),$$

which is unbounded as $|x| \to \infty$.

The answer to the second question is "1/2". This is because

$$P\left(X > 1/2\right) = \int_{1/2}^{\infty} \frac{1}{\pi} \cdot \frac{1}{1+(x-1/2)^2}\, dx = -\int_{1/2}^{-\infty} \frac{1}{\pi} \cdot \frac{1}{1+(-y+1/2)^2}\, dy = P\left(X < 1/2\right).$$

The third equation follows from making the substitution $x = -y + 1$.

Cauchy distribution II

1/1 point (graded)

As in the previous problem, let X denote a random variable distributed as the Cauchy distribution with location parameter m.

Which of the following are true about the random variable $X - m$? (Choose all that apply.)

- [x] The expectation (first moment) of $X - m$ is not defined.
- [x] $X - m$ is distributed as a Cauchy random variable with location parameter set to be 0.
- [x] $X - m$ is **symmetric** in the sense that $X - m$ and $m - X$ both have the same distribution.
- [ ] The method of moments can be used to estimate the location parameter m.

✔

Solution:

Let us examine the choices in order.

- "The expectation (first moment) of $X - m$ is not defined." is correct. As we showed in a previous problem, the improper integral

$$\int_{-\infty}^{\infty} \frac{1}{\pi} \cdot \frac{x}{1 + (x - m)^2} \, dx$$

  does not converge, so the expectation of X is not defined.

- "$X - m$ is distributed as a Cauchy random variable with location parameter set to be 0." is correct. We show that $X - m$ has the same cdf as a Cauchy random variable Y with location parameter set to be 0. Indeed

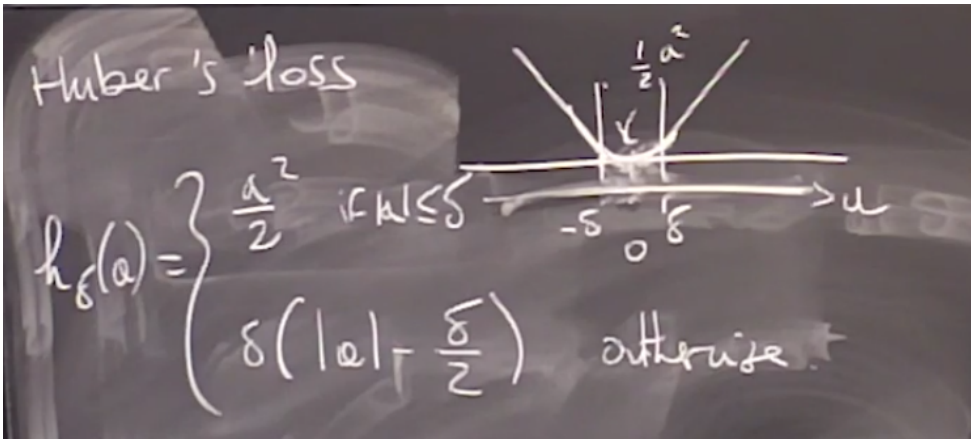
$$P(X - m < t) = \int_{-\infty}^{t+m} \frac{1}{\pi} \cdot \frac{1}{1 + (x - m)^2} \, dx = \int_{-\infty}^{t} \frac{1}{\pi} \cdot \frac{1}{1 + y^2} \, dy = P(Y < t).$$

  Here we made the substitution $y = x - m$.

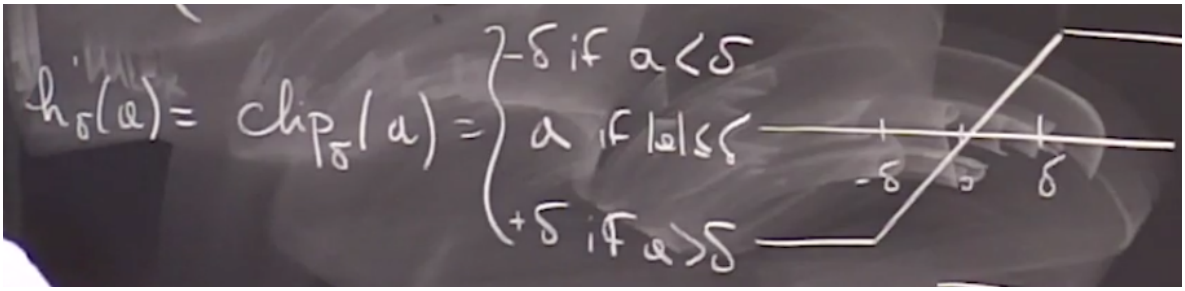
- "$X - m$ is **symmetric** in the sense that $X - m$ and $m - X$ both have the same distribution." is correct. By the previous question, $X - m$ has a density given by $f(x) = \frac{1}{\pi} \frac{1}{1+x^2}$. This is an even function, so it follows that $X - m$ and $m - X$ have the same distribution.

- "The method of moments can be used to estimate the location parameter m." is incorrect. Since the moments of a Cauchy random variable do not exist, the method of moments cannot be used for parameter estimation for this family of distributions.

Huber's loss (not in slides)
wrote a book on robustness in statistics
problem with absolute is that it is a V
can't differentiate at the point (bottom of V)
so instead replaces by a curve

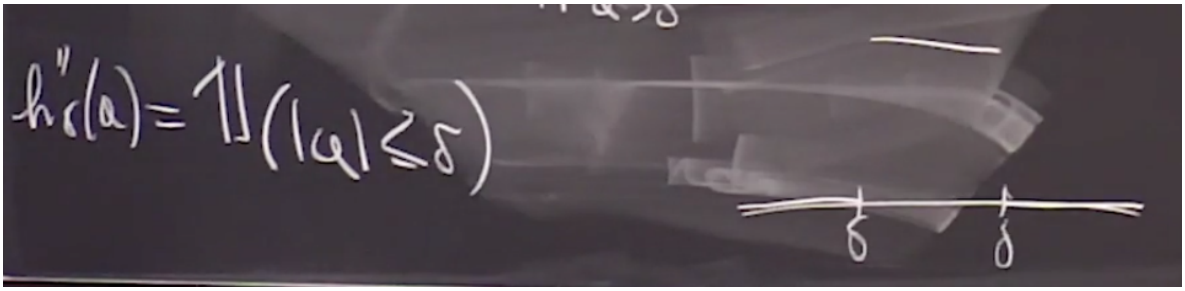


this loss function is differentiable
this is the first derivative



is ‘clipping’ a to be within -delta and delta

this is the second derivative
not continuous but not important

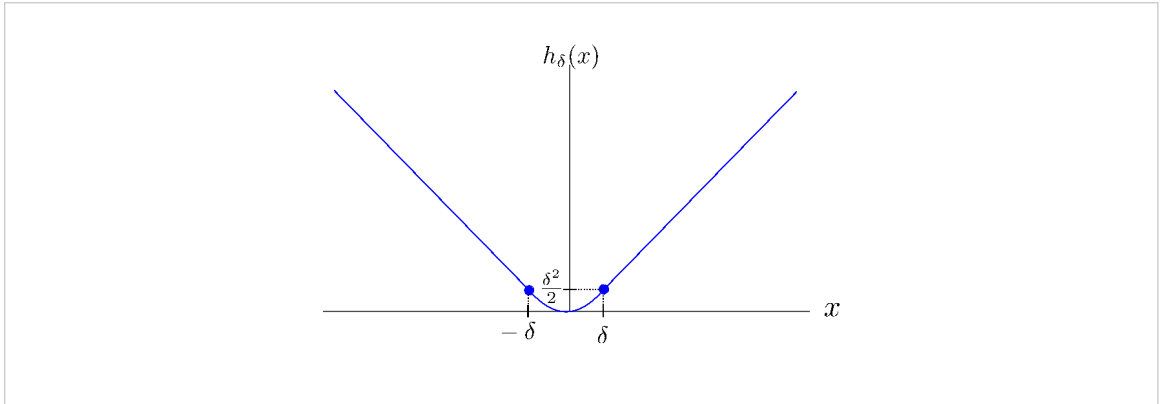


Huber's Loss

3/3 points (graded)

Huber's loss is defined to be

$$h_\delta(x) = \begin{cases} \frac{x^2}{2} & \text{if } |x| \le \delta \\ \delta(|x| - \delta/2) & \text{if } |x| > \delta \end{cases}$$



Let k denote the smallest integer such that the $\frac{d^k}{dx^k} h_\delta(x)$ is **not** a continuous function.

What is k?

2 ✔ **Answer:** 2

The function $\frac{d^k}{dx^k} h_\delta(x)$ is discontinuous at two points $x_1, x_2 \in \mathbb{R}$ where $x_1 < x_2$.

What are x_1 and x_2 in terms of δ?

$x_1 =$ -delta ✔ **Answer:** -delta

$-\delta$

$x_2 =$ delta ✔ **Answer:** delta

δ

STANDARD NOTATION

Solution:

Observe that

$$\frac{\partial h_\delta}{\partial x}(x) = \begin{cases} x & \text{if } |x| < \delta \\ \delta & \text{if } x > \delta \\ -\delta & \text{if } x < -\delta, \end{cases}$$

which is a continuous function. However, the next derivative

$$\frac{\partial^2 h_\delta}{\partial^2 x}(x) = \begin{cases} 1 & \text{if } |x| < \delta \\ 0 & \text{if } |x| > \delta \end{cases}$$

has discontinuities at $x = \pm\delta$. In particular, $\frac{\partial^2 h_\delta}{\partial^2 x}(\pm\delta)$ is not defined. Therefore, for the first question, we conclude that $k = 2$. For the second question, $x_1 = -\delta$ and $x_2 = \delta$.

Comparing Huber's Loss and the absolute value function

0/1 point (graded)

Recall Huber's loss $h_\delta(x)$ as defined in the previous problem. The absolute value function is defined to be $|x|$.

Which of the following statements are true? (Choose all that apply.)

- ☐ Both Huber's loss and the absolute value are differentiable everywhere.
- ☐ For $x > 0$ sufficiently large, both Huber's loss and the absolute value are both linear functions. ✔
- ☑ In the intervals where $h_\delta(x)$ is a linear function, both Huber's loss and the absolute value function have the same slope.
- ☐ Both Huber's loss and the absolute value function are convex. ✔

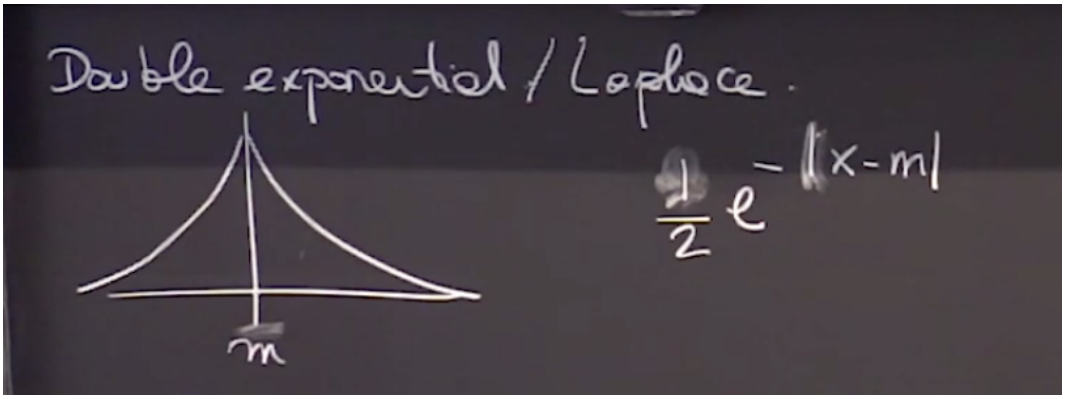
✖

Solution:

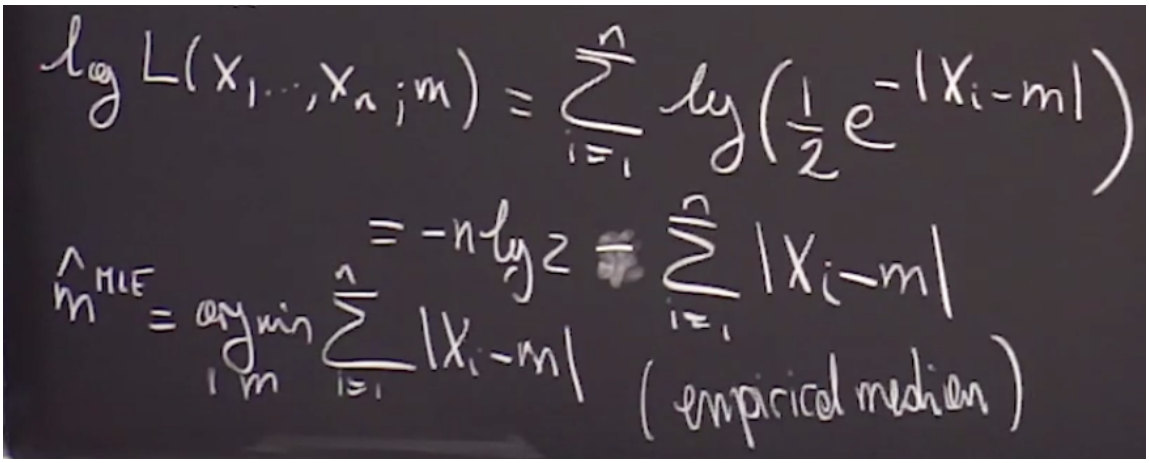
We examine the choices in order.

- "Both Huber's loss and the absolute value are differentiable everywhere." is incorrect. It is true that Huber's loss is differentiable everywhere. However, $|x|$ is not differentiable at $x = 0$.
- "For $x > 0$ sufficiently large, both Huber's loss and the absolute value are both linear functions." is correct. This is certainly true for the absolute function, as $|x| = x$ if $x > 0$. Moreover, if $x > \delta$, then we have $h_\delta(x) = \delta(x - \delta/2)$ which is also a linear function.
- "In the intervals where $h_\delta(x)$ is a linear function, both Huber's loss and the absolute value function have the same slope." is incorrect. For example, if $x > \delta$, then $|x|$ has slope $+1$. However, $h_\delta(x)$ has slope δ, which is not necessarily equal to 1.
- "Both Huber's loss and the absolute value function are convex." is correct. This is evident from the graphs of $|x|$ and $h_\delta(x)$.

Applying Huber's loss to the Laplace distribution

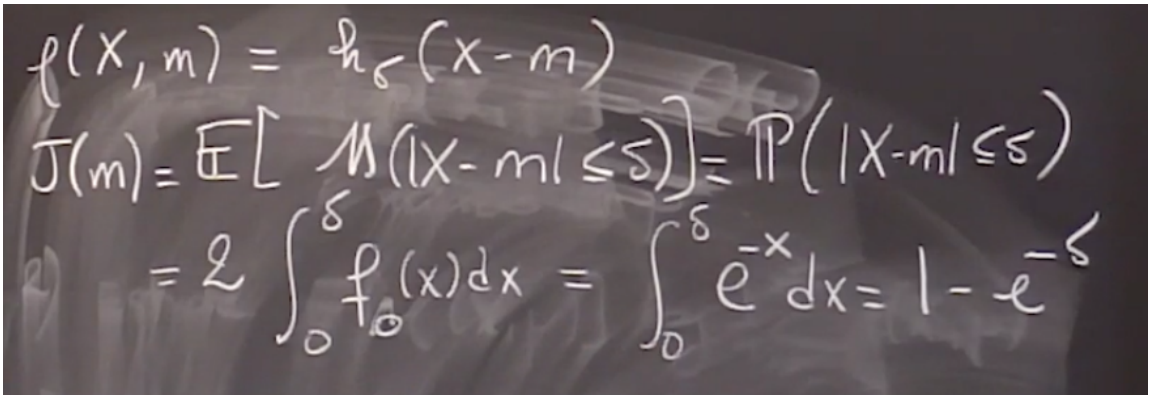


MLE



Huber's loss is very close to this, especially as we let delta go to 0

first derivative (second further down)



2x because of the symmetry on each side between 0 and delta

The Laplace distribution

2/3 points (graded)

The **Laplace distribution** (also known as the **double-exponential distribution**) is a continuous distribution with **location parameter** $m \in \mathbb{R}$ and density given by

$$f_m(x) = \frac{1}{2} e^{-|x-m|}.$$

Let X denote a Laplace random variable with location parameter set to be $m = 0$.

What is $\mathbb{E}[X]$?

0 ✔ **Answer:** 0

Does the variance $\sigma^2 = \mathbb{E}[(X - \mathbb{E}[X])^2]$ exist?

- (●) Yes
- ( ) No

✔

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

Hint: The function $x^k e^{-|x|}$ is integrable, i.e. $\int_{-\infty}^{\infty} x^k e^{-|x|} dx$ is finite for all k.

- [x] The distribution of X is symmetric in the sense that X and $-X$ have the same distribution. ✔
- [x] The function $\ln f_m(x)$ has a continuous first derivative.
- [ ] For any integer $k > 0$, the k-th moment $\mathbb{E}[X^k]$ exists. ✔

✖

Solution:

For the first question, observe that the function $xe^{-|x|}$ is odd and also integrable. Therefore,

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} \frac{1}{2} x e^{-|x|}\, dx = 0.$$

For the second question, the function $x^2 e^{-|x|}$ is integrable. Hence,

$$\mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] = \int_{-\infty}^{\infty} \frac{1}{2} x^2 e^{-|x|}\, dx$$

For the third question, we examine the choices in order.

- "The distribution of X is symmetric in the sense that X and $-X$ have the same distribution." is correct. This is because the density $\frac{1}{2}e^{-|x|}$ is an even function.
- "The function $\ln f_m(x)$ has a continuous first derivative." is incorrect. This is because $\ln f_m(x) = -|x - m|$, which is not differentiable at $x = m$.
- "For any integer $k > 0$, the k-th moment $\mathbb{E}[X^k]$ exists." is correct. In general, the function $x^k e^{-|x|}$ is integrable on $\mathbb{R}$, so the k-th moment $\mathbb{E}[X^k]$ exists for all $k > 0$.

The Sample Median

4 points possible (graded)

Let $S = x_1 < x_2 < \ldots < x_n$ denote a sorted list of numbers. We define the **elementary median** $\text{med}_e(S)$ to be

$$\text{med}_e(S) := \begin{cases} x_{\lceil n/2 \rceil} & \text{if } n \text{ is odd} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}) & \text{if } n \text{ is even} \end{cases}$$

In other words, when n is odd, the median is the middle number when the set S is sorted from smallest to largest. If n is even, we can just define the median to be the average of both middle numbers. This definition is likely familiar from prior math classes.

A more advanced definition, useful for statistical purposes, is to define the **sample median** $\text{med}_s(S)$ of a sample $S := X_1, X_2, \ldots, X_n$ to be

$$\text{med}(S) := \text{argmin}_m \sum_{i=1}^{n} |X_i - m|\,.$$

While the elementary median is unique, this is not always the case for the **sample median**, as you will see in the next few questions.

While the elementary median is unique, this is not always the case for the **sample median**, as you will see in the next few questions.

Consider the data set $S = \{1, 2, 3\}$.

What is the **elementary median** of S?

2 ✔ **Answer:** 2

What is the **sample median** of S?
Hint: Use computational software to graph the objective function.

2 ✔ **Answer:** 2

Now consider the data set $T = \{1, 2, 3, 4\}$.

What is the **elementary median** of T?

2.5 ✔ **Answer:** 2.5

Which of the following describes **all** of the sample medians of T? (*Hint:* Use computational software to graph the objective function)

- ○ 2.5
- ○ Any number in the open interval $(2, 3)$.
- ◉ Any number in the closed interval $[2, 3]$.
- ○ None of the above.

✔

Solution:

For the first question, the elementary median of S is 2. For the second question, we want to find m such that

$$F(m) = |1 - m| + |2 - m| + |3 - m|$$

is as small as possible. Note that $F(m)$ is a piecewise linear function with discontinuities in its first derivative precisely at the points $(1, 3)$, $(2, 2)$, and $(3, 3)$ (one corresponding to each summand of the form $|x - m|$). Drawing the line segments connecting these three points, we see that the minimizer of $F(m)$ is at $m = 2$. Therefore, for the second question, the sample median is 2.

For the third question, the elementary median is 2.5. For the fourth question, we want to find m such that

$$G(m) = |1 - m| + |2 - m| + |3 - m| + |4 - m|$$

is as small as possible. As before, $G(m)$ is a piecewise linear function. Its discontinuities are at the points $(1, 6)$, $(2, 4)$, $(3, 4)$, and $(4, 6)$. Hence, $G(m)$ is a horizontal line segment on the interval $[2, 3]$, and $G(m)$ has positive slope otherwise. Hence, the correct response to the fourth question is "Any number in the closed interval $[2, 3]$.".

Remark: In general, one can show that for an ordered sample $S = x_1 < \ldots < x_n$ that

$$\text{med}(S) = \begin{cases} x_{\lceil n/2 \rceil} & \text{if } n \text{ is odd} \\ \text{Any number in the interval } \ [x_{n/2}, x_{n/2+1}] & \text{if } n \text{ is even} \end{cases}$$

Maximum Likelihood Estimator for the Laplace Distribution

2/2 points (graded)

Consider a Laplace statistical model $(\mathbb{R}, \{P_m\}_{m\in\mathbb{R}})$ where P_m denotes the Laplace distribution with location parameter m. Let $X_1, \ldots, X_n \overset{iid}{\sim} P_{m^*}$ denote a sample from a Laplace distribution with unknown parameter m^*. Recall that the density of P_m is given by

$$f_m(x) = \frac{1}{2}e^{-|x-m|}.$$

What is the log-likelihood $\ell_n(X_1, \ldots, X_n; m)$ for this statistical model?

- (selected) $-n\ln(2) - \sum_{i=1}^n |X_i - m|$
- $-\sum_{i=1}^n |X_i - m|$
- $-n\ln(2) - \sum_{i=1}^n (X_i - m)^2$
- $-n\ln(2) + \sum_{i=1}^n |X_i - m|$

✔

Recall that the maximum likelihood estimator $\widehat{m}_n^{\text{MLE}}$ is given by

$$\widehat{m}_n^{\text{MLE}} = \text{argmin}_{m\in\mathbb{R}} -\ell_n(X_1, \ldots, X_n; m).$$

Suppose you observe the sample

$$S = 0.5, 1.2, 0.6, -0.7, -0.2.$$

What is the value of the MLE for m^* for this data set? *Hint:* Use the previous question, in particular the remark at the end of the solution.

0.5 ✔ **Answer:** 0.5

Solution:

For the first question, the likelihood for n observations is given by

$$\prod_{i=1}^{n} f_m(X_i) = \frac{1}{2^n} \prod_{i=1}^{n} e^{-|x-m|}.$$

Therefore,

$$\ell_n(X_1, \ldots, X_n; m) = -n \ln(2) - \sum_{i=1}^{n} |X_i - m|.$$

For the second question, we need to minimize the quantity

$$5 \ln(2) + |m - 0.5| + |m - 1.2| + |m - 0.6| + |m + 0.2| + |m + 0.7|.$$

with respect to m. As stated in the previous quantity, any m that minimizes the above is a sample median of the data set S. Since S is odd, we have that $\widehat{m}_n^{\text{MLE}} = 0.5$.

Concept Question: Maximum Likelihood Estimator for the Laplace distribution

1/1 point (graded)

As in the previous problem, let $\widehat{m}_n^{\text{MLE}}$ denote the MLE for an unknown parameter m^* of a Laplace distribution.

Can we apply the theorem for the asymptotic normality of the MLE to $\widehat{m}_n^{\text{MLE}}$? (You must choose the correct answer that also has the correct explanation.)

- ○ No, because the log-likelihood is not concave.
- ◉ No, because the log-likelihood is not twice-differentiable, so the Fisher information does not exist.
- ○ Yes, because the log-likelihood is concave.
- ○ Yes, because the other technical conditions required to apply the theorem are satisfied.

✔

Solution:

We examine the choices in order.

- "No, because the log-likelihood is not concave." is incorrect. A sum of concave functions is concave, and for any constant c, the function $x \to -|x - c|$ is concave. Therefore, the log-likelihood

$$\ell_n(X_1, \ldots, X_n; m) = -n \ln(2) - \sum_{i=1}^{n} |X_i - m|$$

  is also concave. Hence, the reasoning for this response is incorrect.

- "No, because the log-likelihood is not twice-differentiable, so the Fisher information does not exist." is correct. This is because $\ell_n(X_1, \ldots, X_n; m)$ has discontinuities in its first derivative with respect to m at $m = X_i$ for $i = 1, \ldots, n$.

- "Yes, because the log-likelihood is concave." is incorrect. Although the log-likelihood is concave, the Fisher information does not exist, as discussed in the analysis of the previous two responses.

- "Yes, because the other technical conditions required to apply the theorem are satisfied." is incorrect. The remaining technical conditions are not enough to guarantee asymptotic normality since the Fisher information does not exist, as discussed above.

Applying Huber's loss to a Laplace distribution I

1/2 points (graded)

As above, let m^* denote an unknown parameter for a Laplace distribution. In this problem, we will use the principles of M-estimation and the smoothness of Huber's loss function to construct an asymptotically normal estimator for m^*. Let P_m denote the Laplace distribution with location parameter m.

Recall Huber's loss is defined as

$$h_\delta(x) = \begin{cases} \frac{x^2}{2} & \text{if } |x| < \delta \\ \delta(|x| - \delta/2) & \text{if } |x| > \delta \end{cases}.$$

As computed in lecture, the derivative of Huber's loss is the **clip function** :

$$\text{clip}_\delta(x) := \frac{d}{dx} h_\delta(x) = \begin{cases} \delta & \text{if } x > \delta \\ x & \text{if } -\delta \le x \le \delta \\ -\delta & \text{if } x < -\delta \end{cases}$$

Find the value of

$$\frac{\partial}{\partial m} \mathbb{E}_{X \sim P_{m^*}} [h_\delta(X - m)] \Big|_{m=m^*}.$$

Hint: You are allowed to switch the derivative and expectation.

0 ✔

In the framework of M-estimation, our loss function is not Huber's loss itself, but rather

$$\rho(x, m) := h_\delta(x - m)$$

Recall the functions

$$J(m) = \mathbb{E}\left[\frac{\partial^2 \rho}{\partial m^2}(X_1, m)\right]$$
$$K(m) = \mathrm{Var}\left[\frac{\partial \rho}{\partial m}(X_1, m)\right]$$

Do the functions K and J exist for a Laplace statistical model?

- ○ No, because the log-likelihood is not twice-differentiable.
- ○ No, because $J(m)$ exists but $K(m)$ does not.
- ○ Yes, because the Fisher information is well-defined for a Laplace statistical model.
- ◉ Yes, because the function $\rho(x, m)$ as defined above is twice-differentiable.

✔

Solution:

The answer to the first question is 0. To see this, observe that

$$\begin{aligned}\frac{\partial}{\partial m}\mathbb{E}_{X\sim P_m^*}\left[h_\delta\left(X-m\right)\right] &= \mathbb{E}_{X\sim P_m^*}\left[\frac{\partial}{\partial m}h_\delta\left(X-m\right)\right] \\ &= \frac{1}{2}\int_{-\infty}^{\infty}\text{clip}_\delta\left(x-m\right)e^{-|x-m^*|}\,dx \\ &= \frac{1}{2}\left(-\delta\int_{m+\delta}^{\infty}e^{-|x-m^*|}\,dx+\delta\int_{-\infty}^{-\delta+m}e^{-|x-m^*|}\,dx+\int_{-\delta+m}^{\delta+m}\left(x-m\right)e^{-|x-m^*|}dx\right).\end{aligned}$$

Applying the change of variables $y=x-m$, we have

$$=\frac{1}{2}\left(-\delta\int_{\delta}^{\infty}e^{-|y+m-m^*|}\,dy+\delta\int_{-\infty}^{-\delta}e^{-|y+m-m^*|}\,dy+\int_{-\delta}^{\delta}ye^{-|y+m-m^*|}\,dy\right).$$

Setting $m=m^*$, we have

$$\frac{\partial}{\partial m}\mathbb{E}_{X\sim P_m^*}\left[h_\delta\left(X\right)\right]\Big|_{m=m^*}=\frac{1}{2}\left(-\delta\int_{\delta}^{\infty}e^{-|y|}\,dy+\delta\int_{-\infty}^{-\delta}e^{-|y|}\,dy+\int_{-\delta}^{\delta}ye^{-|y|}\,dy\right)=0.$$

Remark: The function $m\mapsto\mathbb{E}_{X\sim P_m^*}\left[h_\delta\left(X\right)\right]$ is strictly convex, so this means the loss function has a unique critical point, and this is where the minimum is attained. The above calculation guarantees that the minimum is at $m=m^*$, the value of the true parameter.

"No, because the log-likelihood is not twice-differentiable.","No, because $J\left(m\right)$ exists but $K\left(m\right)$ does not.","Yes, because the Fisher information is well-defined for a Laplace statistical model.","Yes, because the function $\rho\left(x,m\right)$ as defined above is twice-differentiable."

For the second question, we consider the responses in order.

- "No, because the log-likelihood is not twice-differentiable." is incorrect. In the problem "Huber's loss" on the page "Robust Statistics and Huber's Loss", we showed that $\rho\left(x,m\right)=h_\delta\left(x-m\right)$ is twice-differentiable with respect to m.

- "No, because $J\left(m\right)$ exists but $K\left(m\right)$ does not." is also incorrect. Both $K\left(m\right)$ and $J\left(m\right)$ exist because ρ is twice-differentiable, and its derivatives are integrable.

- "Yes, because the Fisher information is well-defined for a Laplace statistical model." is incorrect. The Fisher information does not exist for a Laplace statistical model, as was shown in the problem "Concept Question: Maximum Likelihood Estimator for the Laplace Distribution" on the page "Applying Huber's loss to the Laplace distribution."

- "Yes, because the function $\rho\left(x,m\right)$ as defined above is twice-differentiable." is the correct answer. In a previous problem, we showed that $\rho\left(x,m\right)=h_\delta\left(x-m\right)$ is twice-differentiable with respect to m.

Applying Huber's loss to a Laplace distribution II

1/1 point (graded)

We use the same set-up as in the previous problem. Recall that m^* is an unknown location parameter for a Laplace distribution.

The M-estimator $\widehat{m}$ for m^* associated to the loss function $\rho(x, m) = h_\delta(x - m)$ is given by

$$\widehat{m} = \text{argmin}_{m \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} h_\delta(X_i - m).$$

Consider the slide *Asymptotic normality* in Unit 3. Suppose that the technical conditions in 3. of slide are satisfied. Also, note that for any fixed x, the function $m \mapsto h_\delta(x - m)$ is strictly convex. (You can see this by observing $h_\delta(x - m) = h_\delta(m - x)$, so the graph of $h_\delta(x - m)$ as a function of m for a fixed x is the same as the graph of $h_\delta(x - m)$ as a function of x for a fixed m.) Finally, consider the calculation of $J(m)$ from lecture.

Can we apply the theorem on the slide *Asymptotic normality* to conclude that $\widehat{m}$ is asymptotically normal?

(Choose the correct answer, 'Yes' or 'No', that also has a correct explanation.)

- No, because m^* is not the unique minimizer of the function $m \mapsto \mathbb{E}_{X \sim P_{m^*}}[\rho(X, m)]$.
- No, because $J(m)$ is not invertible.
- Yes, because m^* is the unique minimizer of the function $m \mapsto \frac{1}{n}\sum_{i=1}^{n} h_\delta(X_i - m)$ and $J(m)$ is invertible.
- **Yes, because m^* is the unique minimizer of the function $m \mapsto \mathbb{E}_{X \sim P_{m^*}}[\rho(X, m)]$ and $J(m)$ is invertible.** ✔

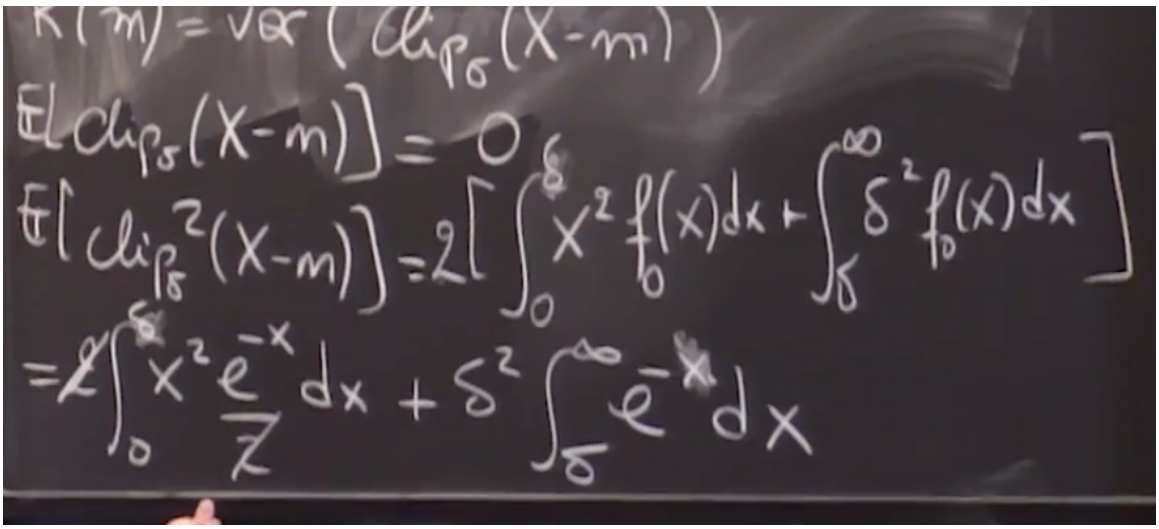
Solution:

"Yes, because m^* is the unique minimizer of the function $m \mapsto \mathbb{E}_{X \sim P_{m^*}}[\rho(X, m)]$ and $J(m)$ is invertible." is the correct answer. As shown in the Remark in the solution to the previous problem, m^* is the unique minimizer of the loss function $m \mapsto \mathbb{E}_{X \sim P_{m^*}}[\rho(X, m)]$. Moreover,

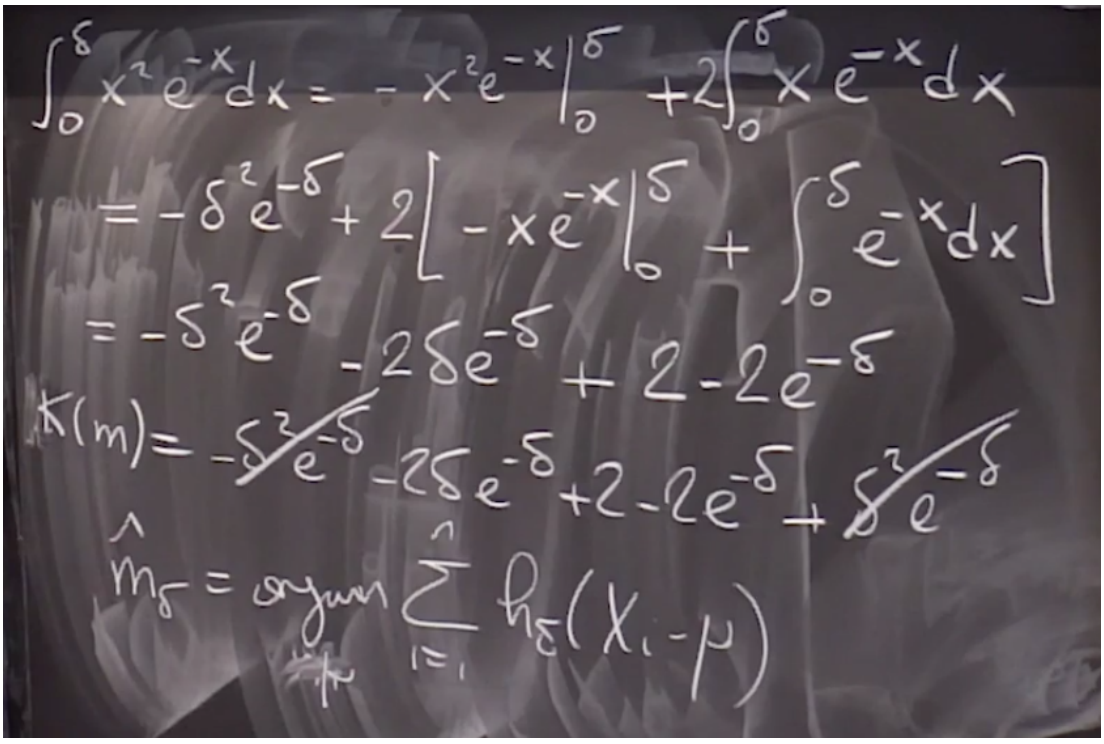
$$J(m) = 1 - e^{-\delta}$$

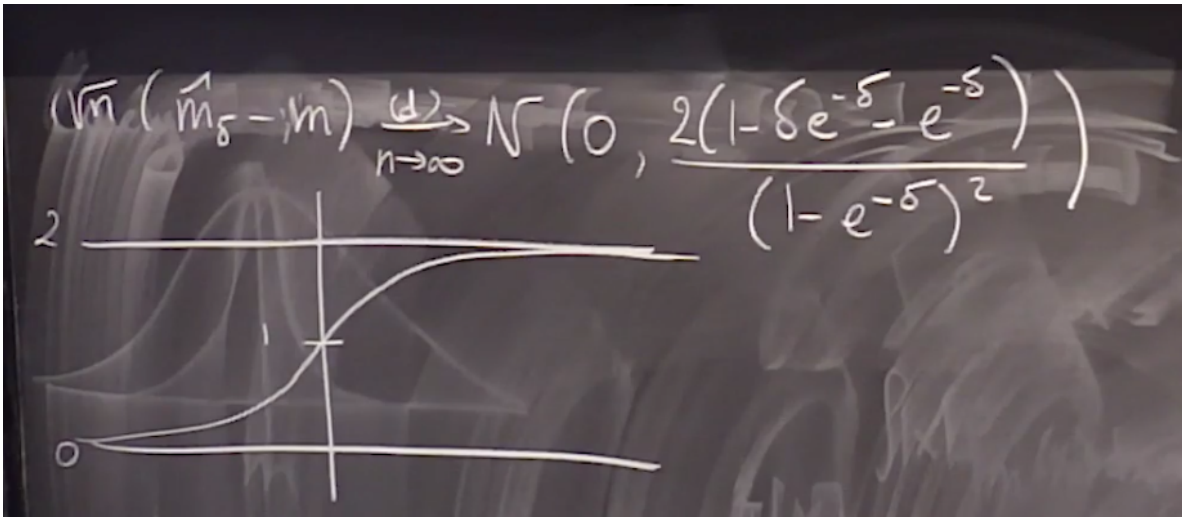
as was shown in the lecture. Hence $J(m)$ is invertible. Assuming that the required technical conditions are satisfied, we conclude that the estimator $\widehat{m}$ is asymptotically normal.

second derivative

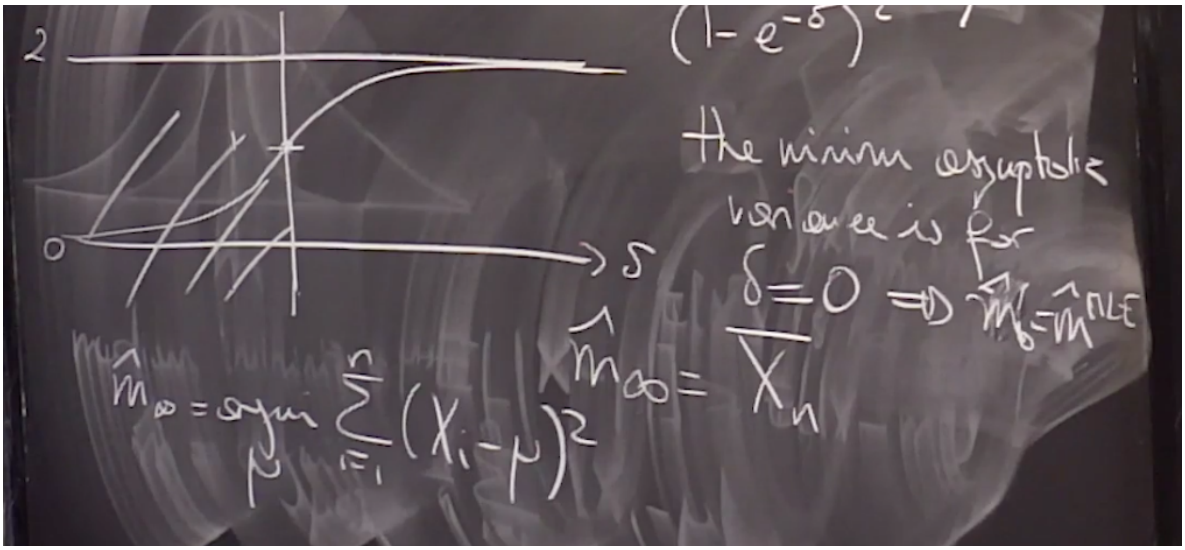


integration by parts because of truncation at delta





not well defined at delta=0 but by getting very close to 0 it becomes 1
and at delta = inf the asymptotic variance becomes 2
so it's bounded between 0 and 2



Asymptotic Variance of the M-estimator for a Laplace distribution

2/2 points (graded)

We use the same statistical set-up from the previous three questions. As before, m^* denotes the location parameter for a Laplace distribution, and $X_1, \ldots, X_n \overset{iid}{\sim} \text{Lap}(m^*)$. Recall the M-estimator

$$\widehat{m}(\delta) = \text{argmin}_{m \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} h_\delta (X_i - m),$$

where now we emphasize the dependence on the parameter $\delta \in (0, \infty)$.

In lecture, we showed that

$$\sqrt{n}(\widehat{m}(\delta) - m^*) \xrightarrow[n \to \infty]{(d)} N(0, g(\delta)).$$

where

$$g(\delta) = \frac{2(1 - \delta e^{-\delta} - e^{-\delta})}{(1 - e^{-\delta})^2}.$$

We can extend g to be a continuous function with domain $[0, \infty]$ by setting $g(0) = 1$ and $g(\infty) = 2$.

Where is the minimum of g attained on $[0, \infty]$? (You may use computational software.)

(If applicable type **inf** for ∞.)

0 ✔ **Answer:** 0

0

Where is the maximum of g attained on $[0, \infty]$? (You may use computational software.)

(If applicable type **inf** for ∞.)

inf ✔ **Answer:** inf

inf

STANDARD NOTATION

Solution:

One can see by graphing that $g(\delta)$ is an increasing function on $[0, \infty]$. Hence, the minimum is attained at $\delta = 0$, and the maximum is attained at $\delta = \infty$. Therefore, the correct response to the first question is "0", and the correct response to the second question is "I". Below we justify this rigorously.

If we are able to show that

$$g'(\delta) \geq 0,$$

for $\delta \in [0, \infty)$, then the result follows. By the quotient rule for derivatives,

$$g'(\delta) = 2 \cdot \left(\frac{\delta e^{-\delta}}{(1 - e^{-\delta})^2} - \frac{2(1 - \delta e^{-\delta} - e^{-\delta}) e^{-\delta}}{(1 - e^{-\delta})^3} \right) = 2 \cdot \frac{\delta e^{-\delta} - 2e^{-\delta} + \delta e^{-2\delta} + 2e^{-2\delta}}{(1 - e^{-\delta})^3}.$$

The denominator is positive for $\delta \in [0, \infty]$, so it suffices to show that the numerator is nonnegative. Let $\tilde{g}(\delta) = \delta - 2 + \delta e^{-\delta} + 2e^{-\delta}$ denote the numerator of the above divided by $e^{-\delta}$. Observe that $\tilde{g}(\delta) \geq 0$ if and only if

$$h(\delta) := e^{\delta}(\delta - 2) + \delta + 2 \geq 0.$$

Since $h(0) = 0$, if we can show that $h'(\delta) \geq 0$ for $\delta \in [0, \infty)$, then this implies $h(\delta)$ is increasing, and hence, $h(\delta) \geq 0$ for $\delta \in [0, \infty)$. Therefore $\tilde{g} \geq 0$ as well, which would suffice to prove what we want.

Observe that

$$h'(\delta) = e^{x}(x - 2) + e^{x} + 1 = xe^{x} - e^{x} + 1.$$

Since $h'(0) = 0$, we would be done if we can show that $h''(\delta) \geq 0$ because

$$\begin{aligned} h''(\delta) &\geq 0 \Rightarrow \\ h'(\delta) &\geq 0 \Rightarrow \\ h(\delta) &\geq 0 \Rightarrow \\ \tilde{g}(\delta) &\geq 0 \Rightarrow \\ g'(\delta) &\geq 0 \end{aligned}$$

on the interval $[0, \infty)$. Finally, $h''(\delta) = \delta e^{-\delta} \geq 0$, so we have shown analytically that $g(\delta)$ is an increasing function on $[0, \infty)$, as desired.

Extreme Values of Huber's loss I

1/1 point (graded)

If $\delta = \infty$, it makes sense to extend the definition of Huber's loss to be

$$h_{\infty}(x) = \frac{x^2}{2}.$$

Setting $\delta = \infty$, we have

$$\widehat{m}(\infty) = \operatorname{argmin}_{m \in \mathbb{R}} \frac{1}{2n} \sum_{i=1}^{n} (X_i - m)^2.$$

What is another name for $\widehat{m}(\infty)$?
Hint: You may use the fact that the objective function is strictly convex.

- (●) The sample average.
- ( ) The sample median.
- ( ) The sample average divided by 2.
- ( ) The sample median divided by 2.

✔

Solution:

The correct response is "The sample average.". We will show this analytically. Let us differentiate and find the value of m that is a critical point of the function

$$F(m) := \frac{1}{2n} \sum_{i=1}^{n} (X_i - m)^2.$$

Observe that

$$F'(m) = -\frac{1}{n} \sum_{i=1}^{n} (X_i - m).$$

Setting $m = \frac{1}{n} \sum_{i=1}^{n} X_i$, we see that $F'(m) = 0$. By strict convexity, this implies that the sample average is the unique global minimizer of $F(m)$.

Extreme values of Huber's loss II

1/1 point (graded)

Note that for all $\delta > 0$,

$$\begin{aligned} \widehat{m}(\delta) &= \operatorname{argmin}_{m \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} h_\delta (X_i - m) \\ &= \operatorname{argmin}_{m \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \frac{h_\delta (X_i - m)}{\delta} \end{aligned}$$

Moreover, for all $x \in \mathbb{R}$,

$$\lim_{\delta \to 0^+} \frac{h_\delta (x)}{\delta} = |x|.$$

Therefore, it makes sense to define

$$\widehat{m}(0) = \operatorname{argmin}_{m \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} |X_i - m|.$$

What is another name for $\widehat{m}(0)$?

- ◯ The sample average.
- ◉ The sample median.
- ◯ The true mean.
- ◯ The true median.

✔

Solution:

The correct response is "The sample median." This is a direct consequence of the definition of the sample median from the problem "The Sample Median" on the page "Applying Huber's Loss to the Laplace Distribution."

Concept Check: Methods of Estimation I

0/1 point (graded)

Which of the following estimators are defined in terms of an optimization problem? (Choose all that apply.)

- ☑ Maximum likelihood estimator. ✔
- ☐ Method of moments estimator.
- ☑ M-estimator. ✔

Solution:

The correct responses are "Maximum likelihood estimator." and "M-estimator." The MLE is defined by maximizing the log-likelihood, and an M-estimator is defined by minimizing a loss function. However, the method of moments estimator is constructed by solving a system of equations, so this response is not correct.

Concept Check: Methods of Estimation II

1/1 point (graded)

All three method of estimation studied in this unit: maximum likelihood estimation, the method of moments, and M-estimation, lead to asymptotically normal estimators if certain technical conditions are satisfied.

In general, an asymptotically normal estimator $\widehat{\theta}_n$ can be used to construct a confidence interval for an unknown parameter.

What quantity related to the estimator $\widehat{\theta}$ determines the length of an asymptotic confidence interval at level 95%? (Assume that you use the plug-in method and that n is very large.)

- (●) The asymptotic variance of $\widehat{\theta}_n$.
- ( ) The rate of convergence of $\widehat{\theta}_n$ to the normal distribution $\mathcal{N}(0, 1)$.
- ( ) The mean of $\widehat{\theta}_n$.

✔

Solution:

The correct response is "The asymptotic variance of $\widehat{\theta}_n$," as we demonstrate below. Consider an asymptotically normal estimator $\widehat{\theta_n}$, which satisfies

$$\sqrt{n}(\widehat{\theta_n} - \theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \sigma^2)$$

for some asymptotic variance $\sigma^2 > 0$. Let $q_{\alpha/2}$ denote the $\alpha/2$-quantile of a standard Gaussian. Then we have that

$$P\left(\sqrt{n}\frac{\left|\widehat{\theta_n} - \theta\right|}{\sigma} \geq q_{\alpha/2}\right) \xrightarrow{n\to\infty} \alpha$$

which implies that

$$P\left(\theta \notin \left[\widehat{\theta_n} - q_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \widehat{\theta_n} + q_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]\right) \xrightarrow{n\to\infty} \alpha.$$

Therefore, using the plug-in method, we have that

$$P\left(\theta \notin \left[\widehat{\theta_n} - q_{\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}}, \widehat{\theta_n} + q_{\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}}\right]\right) \xrightarrow{n\to\infty} \alpha,$$

Setting $\alpha = 0.05$, we have that

$$\mathcal{I} := \left[\widehat{\theta_n} - q_{\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}}, \widehat{\theta_n} + q_{\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}}\right]$$

If n is very large, we have that $\widehat{\sigma}_n \approx \sigma$, so the length of $\mathcal{I}$ is approximately $2q_{0.025}\sigma/\sqrt{n}$. That is, the length depends only on the $\alpha/2$ quantile, the sample size, and the asymptotic variance. Therefore, "The rate of convergence of $\widehat{\theta}_n$ to the normal distribution $\mathcal{N}(0, 1)$." and "The mean of $\widehat{\theta}_n$." are incorrect responses.