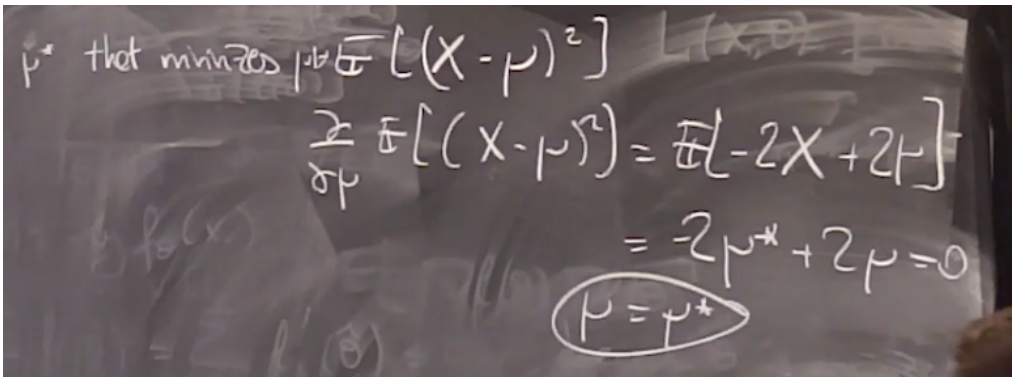
Unit 3 - part 4 M estimation

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 $\mathcal{Q}'(\mu) = \dfrac{d}{d\mu}\mathcal{Q}(\mu)$?

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

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

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,

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

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

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

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

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

$A =$ 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:

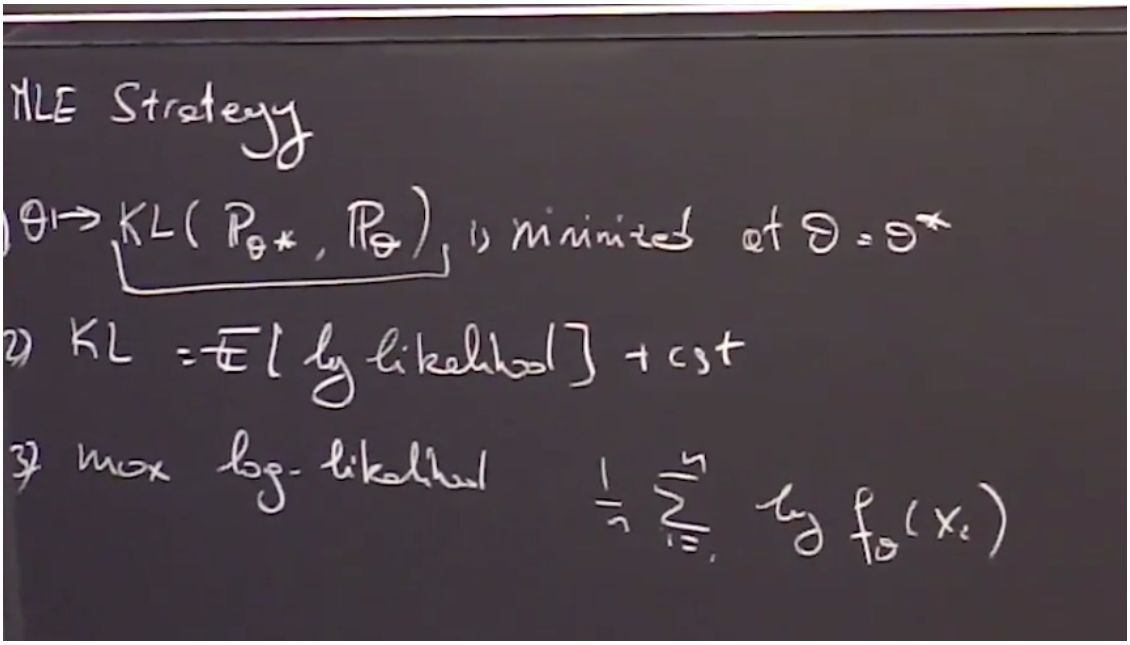
$$\begin{aligned}
F'(\mu) = \frac{d}{d\mu}F(\mu) &= -(1-\alpha)\frac{d}{d\mu}\int_{-\infty}^{\mu} x f(x)\,dx + \alpha\frac{d}{d\mu}\int_{\mu}^{\infty} x f(x)\,dx \\
&\quad + (1-\alpha)\frac{d}{d\mu}\left(\mu\int_{-\infty}^{\mu} f(x)\,dx\right) - \alpha\frac{d}{d\mu}\left(\mu\int_{\mu}^{\infty} f(x)\,dx\right)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

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

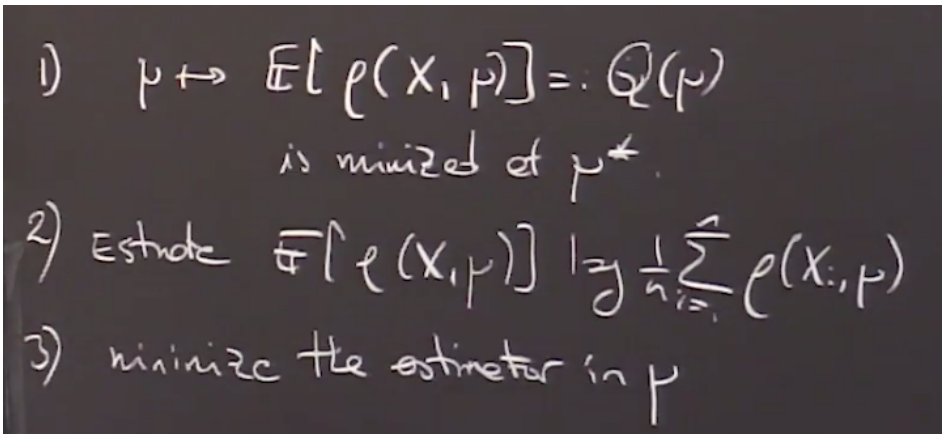
$$\int_{-\infty}^{\mu} f(x)\,dx = \alpha.$$

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.