

18.650
Statistics for Applications

Chapter 3: Maximum Likelihood Estimation

when we do MLE, Likelihood is the function, so we need to maximize the function

Total variation distance (1)

Let $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$ be a statistical model associated with a sample of i.i.d. r.v. X_1, \dots, X_n . Assume that there exists $\theta^* \in \Theta$ such that $X_1 \sim \mathbb{P}_{\theta^*}$: θ^* is the **true** parameter.

Statistician's goal: given X_1, \dots, X_n , find an estimator $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ such that $\mathbb{P}_{\hat{\theta}}$ is close to \mathbb{P}_{θ^*} for the true parameter θ^* .

This means: $|\mathbb{P}_{\hat{\theta}}(A) - \mathbb{P}_{\theta^*}(A)|$ is **small** for all $A \subset E$.

Definition

The *total variation distance* between two probability measures \mathbb{P}_θ and $\mathbb{P}_{\theta'}$ is defined by

$$\text{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \max_{A \subset E} |\mathbb{P}_\theta(A) - \mathbb{P}_{\theta'}(A)|.$$

we have a sample space \mathbb{E} and model

$$P_\Theta \Rightarrow (\mathbb{E}, \{P_\theta\}_{\theta \in \Theta})$$

our goal is to estimate true Θ^* , the one that generated data x_1, x_2, \dots, x_n if x

* But this Θ^* is really a proxy for us to know that we actually understand the distribution itself.

- * The goal of knowing Θ^* is to know what P_{Θ^*} is.
- * Our goal is to come up with the distribution that comes from the family P_Θ that is close to P_{Θ^*} .
- * So, what is it mean for two distributions close to each other. It means that when we compute probabilities on one distribution, you should have probability on the other distribution pretty much.

We have 2 candidate distribution.

$$\hat{\theta} \longrightarrow P_{\hat{\theta}} \rightarrow \text{candidate}$$

$$\theta^* \longrightarrow P_{\theta^*} \rightarrow \text{true}$$

As a statistician we are supposed to come up with good candidate $\hat{\theta}$.

We want

$$P_{\hat{\theta}}([a,b]) \approx P_{\theta^*}([a,b])$$

\Rightarrow for every event $A \in \mathcal{E}$ we want

$$P_{\hat{\theta}}(A) \approx P_{\theta^*}(A)$$

$$\Rightarrow |P_{\hat{\theta}}(A) - P_{\theta^*}(A)| \text{ is small } \forall A \in \mathcal{E}$$

\uparrow
we want this to be true for all event's (σ -measurable sets) in σ -algebra i.e. $\forall A \in \mathcal{E}$, only then we can continue that $\hat{\theta}$ is close to the true unknown

Parameter θ^*

$$TV(P_\Theta, P_{\Theta^*}) = \max_{A \in \mathcal{E}} |P_\Theta(A) - P_{\Theta^*}(A)|$$

worst possible event that

Θ is deviated from Θ^*

so, if you give me two probability measures, P_Θ , $P_{\Theta'}$, and I want to know how close they are

we are finding the worst possible event A that might actually make them differ

\Rightarrow so, if the Total variation $TV(P_\Theta, P_{\Theta'})$ is small, it means that for all possible A 's that you give me, $P_\Theta(A)$ is going to be closer to $P_{\Theta'}(A)$

Ex:

$$0.01 \geq TV(P_\Theta, P_{\Theta^*}) \geq \max_A |P_\Theta(A) - P_{\Theta^*}(A)|$$

$$\Rightarrow P_\Theta(A) \in [P_{\Theta^*}(A) \pm 0.01]$$

$\forall A$

Total variation distance (2)

Assume that E is discrete (i.e., finite or countable). This includes Bernoulli, Binomial, Poisson, ...

Therefore X has a PMF (probability mass function):

$$\mathbb{P}_\theta(X = x) = p_\theta(x) \text{ for all } x \in E,$$

$$p_\theta(x) \geq 0, \quad \sum_{x \in E} p_\theta(x) = 1.$$

The total variation distance between \mathbb{P}_θ and $\mathbb{P}_{\theta'}$ is a simple function of the PMF's p_θ and $p_{\theta'}$:

$$\text{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \frac{1}{2} \sum_{x \in E} |p_\theta(x) - p_{\theta'}(x)| .$$

So, this is maybe not the most convenient way of defining a distance. In reality how are we able to compute the maximum of all possible events.
(There are ∞ number of them).

Now Total variance formula :-

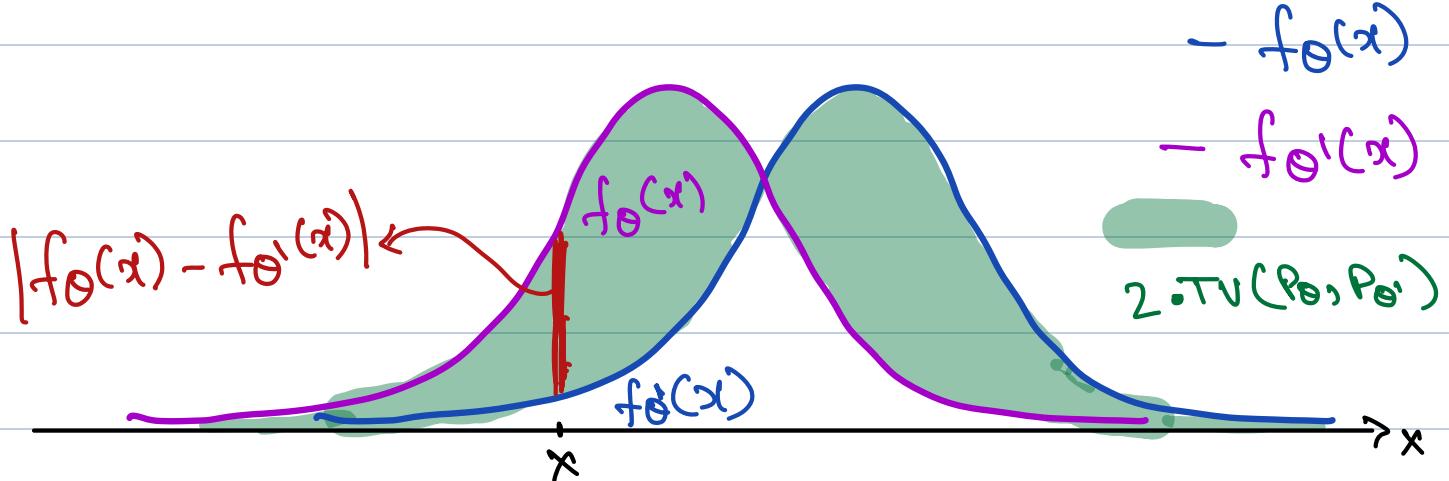
$$\text{PMF:} \Rightarrow P_{\Theta}(x=x) = p_{\Theta}(x)$$

$$\sum_{\forall x} P_{\Theta}(x=x) = 1, \quad p_{\Theta}(x) > 0$$

$$\Rightarrow TV(P_{\Theta}, P_{\Theta'}) = \frac{1}{2} \sum_{x \in E} |p_{\Theta}(x) - p_{\Theta'}(x)|$$

for Continuous:

$$TV(P_{\Theta}, P_{\Theta'}) = \frac{1}{2} \int_{x \in E} |f_{\Theta}(x) - f_{\Theta'}(x)| dx$$



Total variation distance (3)

Assume that E is continuous. This includes Gaussian, Exponential,
....

Assume that X has a density $\mathbb{P}_\theta(X \in A) = \int_A f_\theta(x)dx$ for all $A \subset E$.

$$f_\theta(x) \geq 0, \quad \int_E f_\theta(x)dx = 1.$$

The total variation distance between \mathbb{P}_θ and $\mathbb{P}_{\theta'}$ is a simple function of the densities f_θ and $f_{\theta'}$:

$$\text{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \frac{1}{2} \int_E |f_\theta(x) - f_{\theta'}(x)| dx.$$

$$TV(f(x), g(x)) = \underset{ACE}{\text{more}} |f(x) - g(x)|$$

$$= \frac{1}{2} \int_E |f(x) - g(x)| dx$$

Proof:

$$\frac{1}{2} \int_E |f(x) - g(x)| dx$$

let $A^* = \{x \in E \mid f(x) > g(x)\}$

\downarrow

$$f(x) - g(x) > 0$$

$$\Rightarrow \frac{1}{2} \int_E |f(x) - g(x)| dx = \frac{1}{2} \int_{A^*} (f(x) - g(x)) dx$$

$$+ \frac{1}{2} \int_{(A^*)^c} (g(x) - f(x)) dx$$

we know that

$$\int_E f(x) dx = \int_E g(x) dx = 1$$

$$\Rightarrow \int_E (g(x) - f(x)) dx = 0$$

$$\Rightarrow \int_{A^*} (g(x) - f(x)) dx + \int_{(A^*)^c} (g(x) - f(x)) dx = 0$$

$$\Rightarrow \int_{(A^*)^c} (g(x) - f(x)) dx = - \int_{A^*} (g(x) - f(x)) dx$$

$$= \int_{A^*} (f(x) - g(x)) dx$$

$$\Rightarrow \frac{1}{2} \int_E |f(x) - g(x)| dx = \frac{1}{2} \int_{A^*} (f(x) - g(x)) dx$$

$$+ \frac{1}{2} \int_{A^*} (f(x) - g(x)) dx$$

$$= \int_{A^*} [f(x) - g(x)] dx$$

$$= P_f(A^*) - P_g(A^*) > 0$$

$$= |P_f(A^*) - P_g(A^*)|$$

always

This is
for one
specification

\Rightarrow for all possible A's.

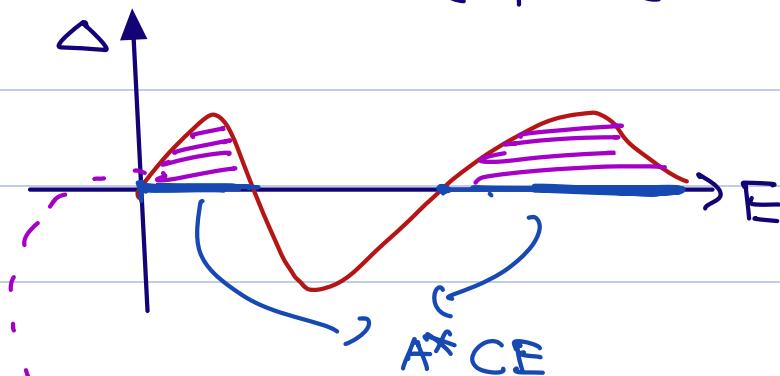
this one needs to be lesser than

more ACE $|P_f(A) - P_g(A)|$

Now we have to show the maximum in
actually event A^*

$$\Delta = f - g \quad \text{defined this function}$$

difference b/w two densities.



total area needs to
be 0 $\Rightarrow \int f - g = 0$

$|P_f(A^*) - P_g(A^*)|$ this is maximum

$\Rightarrow \max_{ACE} |P_\theta - P_{\theta^*}| = \frac{1}{2} \int_E |P_\theta - P_{\theta^*}| dx$



Total variation distance (4)

\mathbb{P}_θ , $\forall \theta \in \Theta$

distance \Leftrightarrow metric

Properties of Total variation:

$(\mathbb{P}_\theta, TV) \rightarrow$
metric
space

- ▶ $TV(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = TV(\mathbb{P}_{\theta'}, \mathbb{P}_\theta)$ (symmetric)
- ▶ $TV(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \geq 0$
- ▶ If $TV(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = 0$ then $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$ (definite)
- ▶ $TV(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \leq TV(\mathbb{P}_\theta, \mathbb{P}_{\theta''}) + TV(\mathbb{P}_{\theta''}, \mathbb{P}_{\theta'})$ (triangle inequality)

These imply that the total variation is a *distance* between probability distributions.

so we have this actual total variance
distance b/w Probability distributions
and, Here is now a Statistical
strategy to Implement our goal.

\Rightarrow goal : Our goal was to SPit out
 $\hat{\theta}$ which was close to θ^* , such that
 $P_{\hat{\theta}}$ is close to P_{θ^*} .

So hopefully ,we are trying to minimize
the total variance distance b/w $P_{\hat{\theta}}$ &
 P_{θ^*} . But we cannot do that ,
Because Just by this fact, we don't know
 θ^*

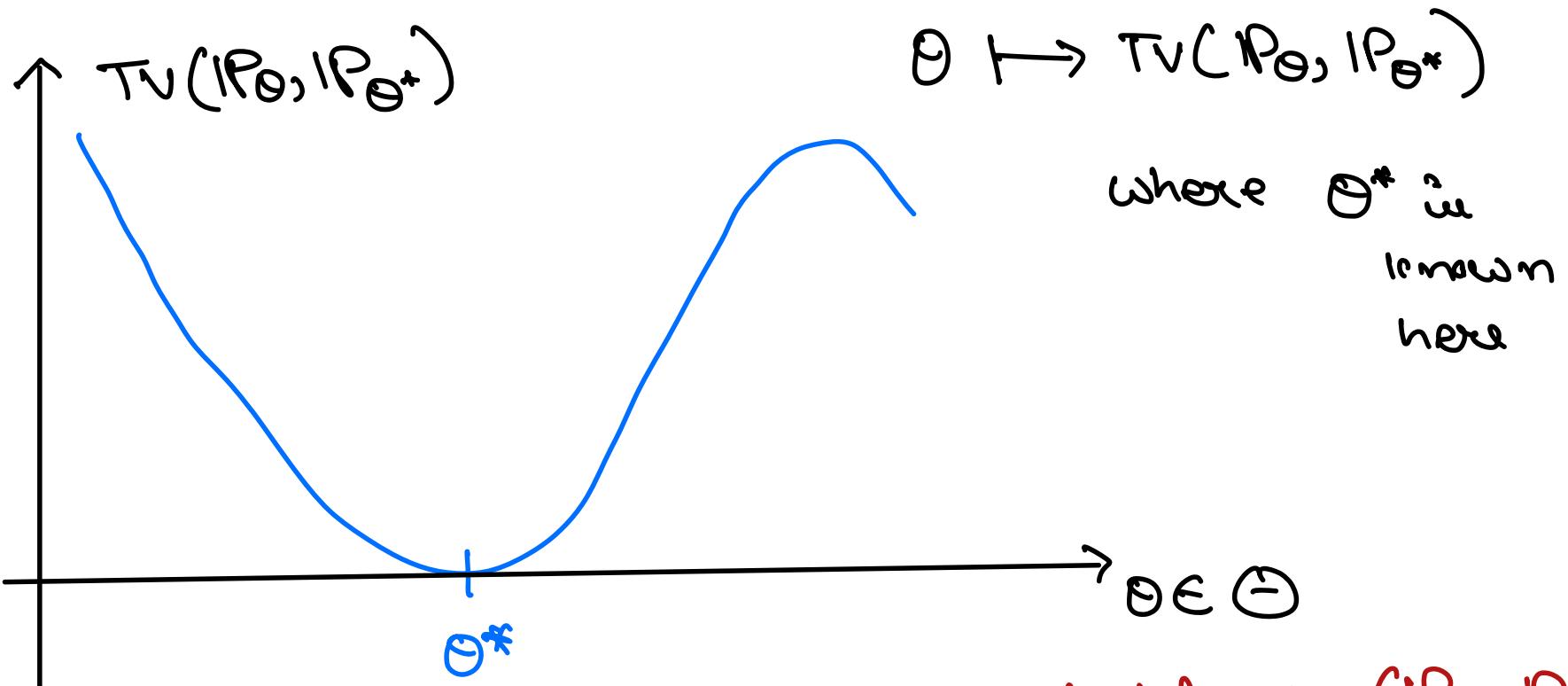
We know how to compute T.V if we have
known Both $P_{\hat{\theta}}$, P_{θ^*} (ie $\hat{\theta}, \theta^*$),
But here one of the argument's (θ^*)
is not known to us, so we need to
estimate it.

Strategy



Total variation distance (5)

An estimation strategy: Build an estimator $\widehat{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta^*})$ for all $\theta \in \Theta$. Then find $\hat{\theta}$ that minimizes the function $\theta \mapsto \widehat{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta^*})$.



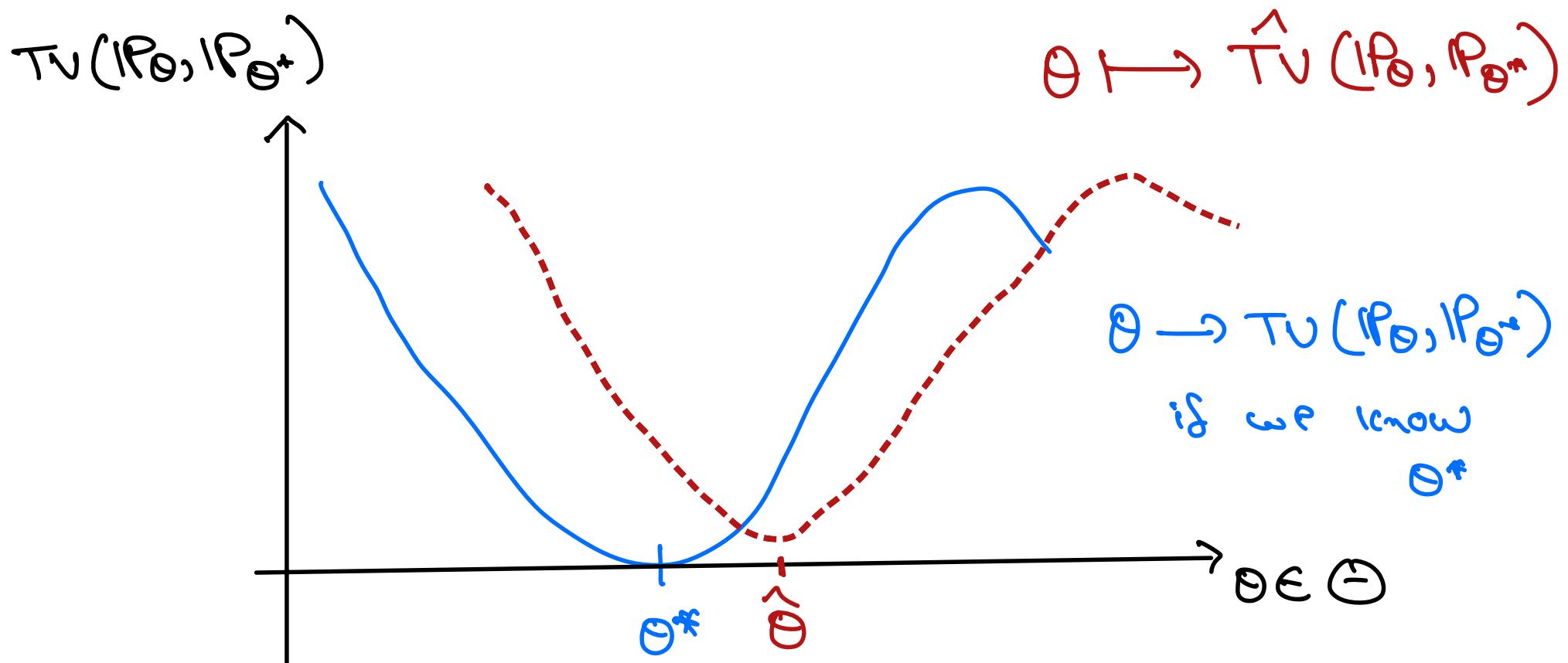
ASSUMING we θ^* , we have calculated $TV(\mathbb{P}_\theta, \mathbb{P}_{\theta^*})$
then the diagram expresses distance for all $\theta \in \Theta$
 \Rightarrow Then $\min TV(\mathbb{P}_\theta, \mathbb{P}_{\theta^*}) = 0$ at $\theta = \theta^*$, But in reality
we don't know this θ^* , so we cannot construct this graph

we are trying to minimize the distance b/w
 $\text{IP}_{\hat{\Theta}}$, IP_{Θ^*}

- * we know how to compute $\text{TV}(\text{IP}_{\hat{\Theta}}, \text{IP}_{\Theta^*})$
if we know both $\hat{\Theta}, \Theta^*$, But here Θ^*
is unknown. Θ^* is not known to us.
 Θ^* we need to estimate
- * so, let's build an estimator of the TV
distance b/w IP_{Θ} , IP_{Θ^*} for all candidate
 $\Theta \in \mathcal{O}$.
- * if this is a good estimate, then when
when we are minimizing this Estimate
we get something that is closer to
 IP_{Θ^*}

Total variation distance (5)

An estimation strategy: Build an estimator $\widehat{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta^*})$ for all $\theta \in \Theta$. Then find $\hat{\theta}$ that *minimizes* the function $\theta \mapsto \widehat{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta^*})$.



problem: Unclear how to build $\widehat{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta^*})$!

we don't know the function $TU(P_0, P_{\theta^*})$

$\forall \theta \in \Theta$ because we don't know the value of θ^* (true parameter)

\Rightarrow so, we are going to estimate this distance function from data, The more the data, the better this estimator of this function $TU(P_0, P_{\theta^*})$ which is $\hat{TU}(P_0, P_{\theta^*})$

The Problem is that it's very unclear with how to build this Estimator of TU i.e $\hat{TU}(P_0, P_{\theta^*})$

* So Building Estimator's is typically consists of replacing Expectation's by average's BUT there is no simple way of Expressing the TU as an Expectation w.r.t θ^* of anything

Kullback-Leibler (KL) divergence (1)

There are **many** distances between probability measures to replace total variation. Let us choose one that is more convenient.

Definition

The *Kullback-Leibler (KL) divergence* between two probability measures \mathbb{P}_θ and $\mathbb{P}_{\theta'}$ is defined by

$$\text{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \begin{cases} \sum_{x \in E} p_\theta(x) \log \left(\frac{p_\theta(x)}{p_{\theta'}(x)} \right) & \text{if } E \text{ is discrete} \\ \int_E f_\theta(x) \log \left(\frac{f_\theta(x)}{f_{\theta'}(x)} \right) dx & \text{if } E \text{ is continuous} \end{cases}$$

Kullback-Leibler (KL) divergence (2)

Properties of KL-divergence:

- ① ► $\text{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \neq \text{KL}(\mathbb{P}_{\theta'}, \mathbb{P}_\theta)$ in general
- $\text{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \geq 0$
- If $\text{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = 0$ then $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$ (definite)
- $\text{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \nleq \text{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta''}) + \text{KL}(\mathbb{P}_{\theta''}, \mathbb{P}_{\theta'})$ in general

Not a distance.

This is called a *divergence*.

Asymmetry is the key to our ability to estimate it! ②

Kullback-Leibler (KL) divergence (3)

$$\begin{aligned}\text{KL}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta) &= \mathbb{E}_{\theta^*} \left[\log \left(\frac{p_{\theta^*}(X)}{p_\theta(X)} \right) \right] \\ &= \mathbb{E}_{\theta^*} [\log p_{\theta^*}(X)] - \mathbb{E}_{\theta^*} [\log p_\theta(X)]\end{aligned}$$

So the function $\theta \mapsto \text{KL}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta)$ is of the form:
“constant” – $\mathbb{E}_{\theta^*} [\log p_\theta(X)]$

Can be estimated: $\mathbb{E}_{\theta^*}[h(X)] \rightsquigarrow \frac{1}{n} \sum_{i=1}^n h(X_i)$ (by LLN)

$$\widehat{\text{KL}}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta) = \text{“constant”} - \frac{1}{n} \sum_{i=1}^n \log p_\theta(X_i)$$

$$KL(P_{\Theta^*}, P_\Theta) = \mathbb{E}_{\Theta^*} \left[\log \left(\frac{P_{\Theta^*}(x)}{P_\Theta(x)} \right) \right]$$

Expectation w.r.t the true distribution from which my data is actually drawn of the log of this station

HA HA ☺ \Rightarrow I am a statistician, I

can replace Expectation by an average,

because I have data from this distribution and try to minimize here.

$$KL(P_{\Theta^*}, P_\Theta)$$

$$= \underbrace{\mathbb{E}_{\Theta^*} \left[\log P_{\Theta^*}(x) \right]}_{\text{constant, does not depend on } \Theta \text{ (fixed value)}}$$

(negative entropy)

$$- \underbrace{\mathbb{E}_{\Theta^*} \left[\log P_\Theta(x) \right]}_{\text{depends on } \Theta}$$

when Θ changes, this value changes

$$\text{function } \Theta \mapsto KL(P_{\Theta^*}, P_\Theta)$$

$$= \text{Constant} - \mathbb{E}_{\Theta^*} \left[\log (P_\Theta(x)) \right]$$

if we want to compute this we need to know both Θ^*, Θ , still

Kullback-Leibler (KL) divergence (4)

$$\widehat{\text{KL}}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta) = \text{"constant"} - \frac{1}{n} \sum_{i=1}^n \log p_\theta(X_i)$$

$$\begin{aligned}\min_{\theta \in \Theta} \widehat{\text{KL}}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta) &\Leftrightarrow \min_{\theta \in \Theta} -\frac{1}{n} \sum_{i=1}^n \log p_\theta(X_i) \\ &\Leftrightarrow \max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \log p_\theta(X_i) \\ &\Leftrightarrow \max_{\theta \in \Theta} \sum_{i=1}^n \log p_\theta(X_i) \\ &\Leftrightarrow \max_{\theta \in \Theta} \prod_{i=1}^n p_\theta(X_i)\end{aligned}$$

This is the **maximum likelihood principle**.

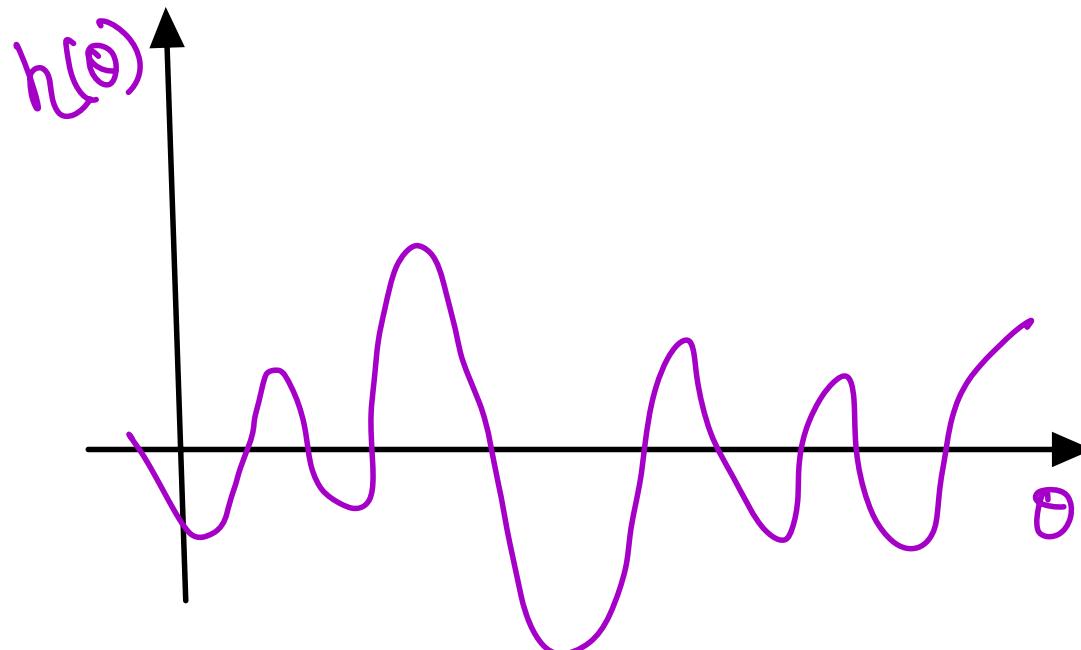
Interlude: maximizing/minimizing functions (1)

Note that

$$\min_{\theta \in \Theta} -h(\theta) \Leftrightarrow \max_{\theta \in \Theta} h(\theta)$$

In this class, we focus on **maximization**.

Maximization of arbitrary functions can be difficult:



Example: $\theta \mapsto \prod_{i=1}^n (\theta - X_i)$

Interlude: maximizing/minimizing functions (2)

Definition

A function twice differentiable function $h : \Theta \subset \mathbb{R} \rightarrow \mathbb{R}$ is said to be *concave* if its second derivative satisfies

$$h''(\theta) \leq 0, \quad \forall \theta \in \Theta$$

It is said to be *strictly concave* if the inequality is strict: $h''(\theta) < 0$

Moreover, h is said to be (strictly) *convex* if $-h$ is (strictly) concave, i.e. $h''(\theta) \geq 0$ ($h''(\theta) > 0$).

Examples:

- ▶ $\Theta = \mathbb{R}, h(\theta) = -\theta^2,$
- ▶ $\Theta = (0, \infty), h(\theta) = \sqrt{\theta},$
- ▶ $\Theta = (0, \infty), h(\theta) = \log \theta,$
- ▶ $\Theta = [0, \pi], h(\theta) = \sin(\theta)$
- ▶ $\Theta = \mathbb{R}, h(\theta) = 2\theta - 3$

Interlude: maximizing/minimizing functions (3)

More generally for a *multivariate* function: $h : \Theta \subset \mathbb{R}^d \rightarrow \mathbb{R}$, $d \geq 2$, define the

- ▶ *gradient vector*: $\nabla h(\theta) = \begin{pmatrix} \frac{\partial h}{\partial \theta_1}(\theta) \\ \vdots \\ \frac{\partial h}{\partial \theta_d}(\theta) \end{pmatrix} \in \mathbb{R}^d$

▶ *Hessian matrix*:

$$\nabla^2 h(\theta) = \begin{pmatrix} \frac{\partial^2 h}{\partial \theta_1 \partial \theta_1}(\theta) & \cdots & \frac{\partial^2 h}{\partial \theta_1 \partial \theta_d}(\theta) \\ & \ddots & \\ \frac{\partial^2 h}{\partial \theta_d \partial \theta_d}(\theta) & \cdots & \frac{\partial^2 h}{\partial \theta_d \partial \theta_d}(\theta) \end{pmatrix} \in \mathbb{R}^{d \times d}$$

h is concave $\Leftrightarrow x^\top \nabla^2 h(\theta) x \leq 0 \quad \forall x \in \mathbb{R}^d, \theta \in \Theta.$

h is strictly concave $\Leftrightarrow x^\top \nabla^2 h(\theta) x < 0 \quad \forall x \in \mathbb{R}^d, \theta \in \Theta.$

Examples:

- ▶ $\Theta = \mathbb{R}^2$, $h(\theta) = -\theta_1^2 - 2\theta_2^2$ or $h(\theta) = -(\theta_1 - \theta_2)^2$
- ▶ $\Theta = (0, \infty)$, $h(\theta) = \log(\theta_1 + \theta_2)$,

Interlude: maximizing/minimizing functions (4)

Strictly concave functions are easy to maximize: if they have a maximum, then it is **unique**. It is the unique solution to

$$h'(\theta) = 0,$$

or, in the multivariate case

$$\nabla h(\theta) = 0 \in \mathbb{R}^d.$$

There are many algorithms to find it numerically: this is the theory of “convex optimization”. In this class, often a **closed form formula** for the maximum.

Likelihood, Discrete case (1)

Let $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$ be a statistical model associated with a sample of i.i.d. r.v. X_1, \dots, X_n . Assume that E is discrete (i.e., finite or countable).

Definition

The *likelihood* of the model is the map L_n (or just L) defined as:

$$\begin{aligned} L_n : E^n \times \Theta &\rightarrow \mathbb{R} \\ (x_1, \dots, x_n, \theta) &\mapsto \mathbb{P}_\theta[X_1 = x_1, \dots, X_n = x_n]. \end{aligned}$$

function of all the data point's & parameter (θ)

Likelihood, Discrete case (2)

Example 1 (Bernoulli trials): If $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Ber}(p)$ for some $p \in (0, 1)$:

- ▶ $E = \{0, 1\}$;
- ▶ $\Theta = (0, 1)$;
- ▶ $\forall (x_1, \dots, x_n) \in \{0, 1\}^n, \quad \forall p \in (0, 1),$

$$\begin{aligned} L(x_1, \dots, x_n, p) &= \prod_{i=1}^n \mathbb{P}_p[X_i = x_i] \\ &= \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} \\ &= p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i}. \\ &= \left(\frac{p}{1-p}\right)^{\sum_{i=1}^n x_i} (1-p)^n \end{aligned}$$

Likelihood, Discrete case (3)

Example 2 (Poisson model):

If $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Poiss}(\lambda)$ for some $\lambda > 0$:

- ▶ $E = \mathbb{N}$;
- ▶ $\Theta = (0, \infty)$;
- ▶ $\forall (x_1, \dots, x_n) \in \mathbb{N}^n, \quad \forall \lambda > 0,$

$$\begin{aligned} L(x_1, \dots, x_n, p) &= \prod_{i=1}^n \mathbb{P}_\lambda[X_i = x_i] \\ &= \prod_{i=1}^n e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \\ &= e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^n x_i}}{x_1! \dots x_n!}. \end{aligned}$$

Likelihood, Continuous case (1)

Let $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$ be a statistical model associated with a sample of i.i.d. r.v. X_1, \dots, X_n . Assume that all the \mathbb{P}_θ have density f_θ .

Definition

The *likelihood* of the model is the map L defined as:

$$\begin{aligned} L &: E^n \times \Theta && \rightarrow \mathbb{R} \\ &(x_1, \dots, x_n, \theta) && \mapsto \prod_{i=1}^n f_\theta(x_i). \end{aligned}$$

Likelihood, Continuous case (2)

Example 1 (Gaussian model): If $X_1, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, for some $\mu \in \mathbb{R}, \sigma^2 > 0$:

- ▶ $E = \mathbb{R}$;
- ▶ $\Theta = \mathbb{R} \times (0, \infty)$
- ▶ $\forall (x_1, \dots, x_n) \in \mathbb{R}^n, \quad \forall (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$,

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

$$\begin{aligned} L(x_1, x_2, \dots, x_n, \mu, \sigma^2) \\ = \frac{1}{(2\pi\sigma^2)^n} \exp\left\{-\frac{1}{2\sigma^2} \left[(x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots + (x_n - \mu)^2 \right]\right\} \end{aligned}$$

Maximum likelihood estimator (1)

Let X_1, \dots, X_n be an i.i.d. sample associated with a statistical model $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$ and let L be the corresponding likelihood.

Definition

The *likelihood estimator* of θ is defined as:

$$\hat{\theta}_n^{MLE} = \operatorname{argmax}_{\theta \in \Theta} L(X_1, \dots, X_n, \theta),$$

provided it exists.

Remark (log-likelihood estimator): In practice, we use the fact that

$$\hat{\theta}_n^{MLE} = \operatorname{argmax}_{\theta \in \Theta} \log L(X_1, \dots, X_n, \theta).$$

Maximum likelihood estimator (2)

Examples

- ▶ Bernoulli trials: $\hat{p}_n^{MLE} = \bar{X}_n$.
- ▶ Poisson model: $\hat{\lambda}_n^{MLE} = \bar{X}_n$.
- ▶ Gaussian model: $(\hat{\mu}_n, \hat{\sigma}_n^2) = (\bar{X}_n, \hat{S}_n)$.

Maximum likelihood estimator (3)

Definition: Fisher information

Define the log-likelihood for one observation as:

$$\ell(\theta) = \log L_1(X, \theta), \quad \theta \in \Theta \subset \mathbb{R}^d$$

Assume that ℓ is a.s. twice differentiable. Under some regularity conditions, the *Fisher information* of the statistical model is defined as:

$$I(\theta) = \mathbb{E}[\nabla \ell(\theta) \nabla \ell(\theta)^\top] - \mathbb{E}[\nabla \ell(\theta)] \mathbb{E}[\nabla \ell(\theta)]^\top = -\mathbb{E}[\nabla^2 \ell(\theta)].$$

If $\Theta \subset \mathbb{R}$, we get:

$$I(\theta) = \text{var}[\ell'(\theta)] = -\mathbb{E}[\ell''(\theta)]$$

Some intuition about how does the Maximum Likelihood Perform?

There is something called Fisher information that essentially controls how this MLE performs.

\Rightarrow Fisher information is essentially a 2nd derivative or hessian. (In multidimensional)

$l(\theta) \rightarrow$ Log Likelihood for one-observation.

$$l(\theta) = \log L_1(x, \theta) \quad \theta \in \Theta \subset \mathbb{R}^d$$

$$\begin{aligned} I(\theta) &= \mathbb{E}[\nabla l(\theta) \nabla l(\theta)^T] - \mathbb{E}[\nabla l(\theta)] \mathbb{E}[\nabla l(\theta)]^T \\ &= -\mathbb{E}[\nabla^2 l(\theta)] \end{aligned}$$

why do we care about this Fisher Information quantity? Fisher - Father of modern statistics.

This quantity is very intuitive. what does the 2nd derivative tell us at the maximum?

It's telling us how curved the function is at maximum.

if 2nd derivative is 0, i.e $H(0) = 0$



Flat surface

very high 2nd derivative \Rightarrow very curvy.

Very Curvy means is that we are very robust to the estimation error (replacing Expectation with average). if we are extremely curvy, we can move a little bit, but the maximum is not gonna move much

\Rightarrow The flatter it is, the more sensitive to fluctuation's the argmax gonna be, the curvy it is, the less sensitive it is.

\Rightarrow A good model is one that has a larger fisher information. (more curvy)

Maximum likelihood estimator (4)

Theorem

Let $\theta^* \in \Theta$ (the *true* parameter). Assume the following:

1. The model is identified.
2. For all $\theta \in \Theta$, the support of \mathbb{P}_θ does not depend on θ ;
3. θ^* is not on the boundary of Θ ;
4. $I(\theta)$ is invertible in a neighborhood of θ^* ;
5. A few more technical conditions.

Then, $\hat{\theta}_n^{MLE}$ satisfies:

- ▶ $\hat{\theta}_n^{MLE} \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \theta^*$ w.r.t. \mathbb{P}_{θ^*} ;
- ▶ $\sqrt{n} \left(\hat{\theta}_n^{MLE} - \theta^* \right) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N} \left(0, I(\theta^*)^{-1} \right)$ w.r.t. \mathbb{P}_{θ^*} .

* Fisher information matrix tells us how much information about the Θ is in our model.

* if our model is very well-parameterized, then we will have a lot of information.

let $X \sim N(\theta, \sigma^2)$ σ^2 unknown, we can parameterize our model by $\sigma, \sigma^2, \sigma^4, \sigma^{24}$ we could parameterize it by whatever we want, then we would have a simple transformation. But we could say some of them are less informative using Fisher information.

Let $\theta \in \mathbb{R}$ (single parameter)

$$l(\theta) = \log L_1(x, \theta) \quad (\text{only one observation})$$

$$= \log f_\theta(x)$$

x - random variable
 $\log f_\theta(x) \rightarrow \text{g.i.v}$
a transformation

$$I(\theta) = \text{Var}_{\theta}(\ell'(\theta)) = -E_{\theta}[\ell''(\theta)]$$

Proof

Assume that X has density $f_{\theta}(x)$

$$\int f_{\theta}(x) dx = 1 \implies \frac{\partial^2}{\partial \theta^2} \int f_{\theta}(x) dx = 0$$

↓

$$\frac{\partial}{\partial \theta} \int f_{\theta}(x) dx = 0$$

↓

$$\int \frac{\partial^2}{\partial \theta^2} f_{\theta}(x) dx = 0$$

$$\boxed{\int \frac{\partial}{\partial \theta} f_{\theta}(x) dx = 0}$$

$$\ell''(\theta) = \frac{\partial^2}{\partial \theta^2} \log f_{\theta}(x) = \frac{\partial}{\partial \theta} \left[\frac{f'_{\theta}(x)}{f_{\theta}(x)} \right]$$

$$= \frac{f_{\theta}(x) f''_{\theta}(x) - {f'_{\theta}(x)}^2}{{f_{\theta}(x)}^2}$$

$$-E[\ell''(\theta)] = -E \left[\frac{f_{\theta}(x) f''_{\theta}(x) - {f'_{\theta}(x)}^2}{{f_{\theta}(x)}^2} \right]$$

$$= - \int \frac{f_0(x) f_0''(x) - f_0'(x)^2}{f_0(x)^2} \cdot f_0(x) dx$$

$$= - \int f_0''(x) dx + \int \frac{f_0'(x)^2}{f_0(x)} dx$$

$$= - \int \frac{\partial^2}{\partial \theta^2} f_0(x) dx + \int \frac{\left[\frac{\partial}{\partial \theta} f_0(x) \right]^2}{f_0(x)} dx$$

\approx 0

$$-E[\ell'(\theta)] = \int \frac{\left[\frac{\partial}{\partial \theta} f_0(x) \right]^2}{f_0(x)} dx$$

$$\text{var}(\ell'(\theta)) = E[(\ell'(\theta))^2] - E[\ell'(\theta)]^2$$

$$E[\ell'(\theta)] = E\left[\frac{\frac{\partial}{\partial \theta} f_0(x)}{f_0(x)} \right] = \int \frac{\frac{\partial}{\partial \theta} f_0(x)}{f_0(x)} f_0(x) dx$$

$$= \int \frac{\partial}{\partial \theta} f_0(x) dx = 0$$

$$\text{Var}(\ell'(\theta)) = E\left[\ell'(\theta)^2\right] - E[\ell'(\theta)]^2 \stackrel{=} {=} 0$$

$$= I E\left[\ell'(\theta)^2\right]$$

$$= \int \frac{\left[\frac{\partial}{\partial \theta} f_\theta(x) \right]^2}{f_\theta(x)^2} f_\theta(x) dx$$

$$= \int \frac{\left[\frac{\partial}{\partial \theta} f_\theta(x) \right]^2}{f_\theta(x)} dx$$

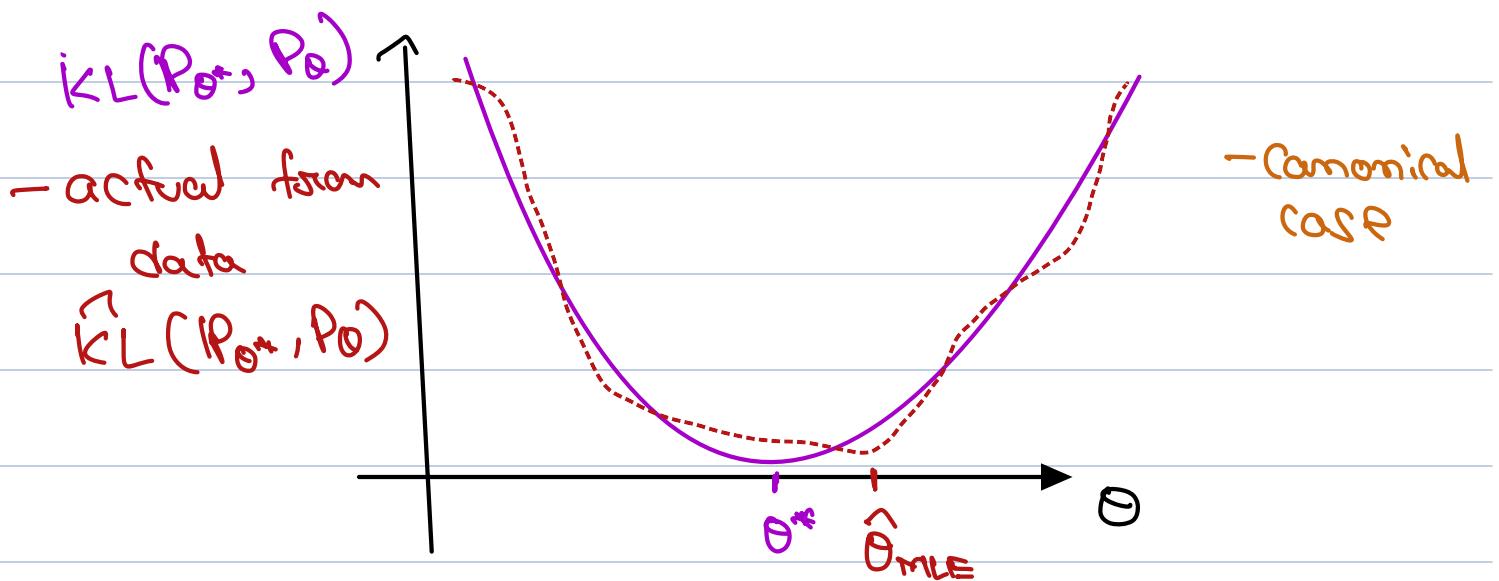
$$\text{Var}(\ell'(\theta)) = - I E\left[\ell''(\theta)\right]$$

Hence Proved

when we are doing MLE \rightarrow its an empirical version of trying to minimize the KL divergence.

MLE \iff minimize KL divergence.

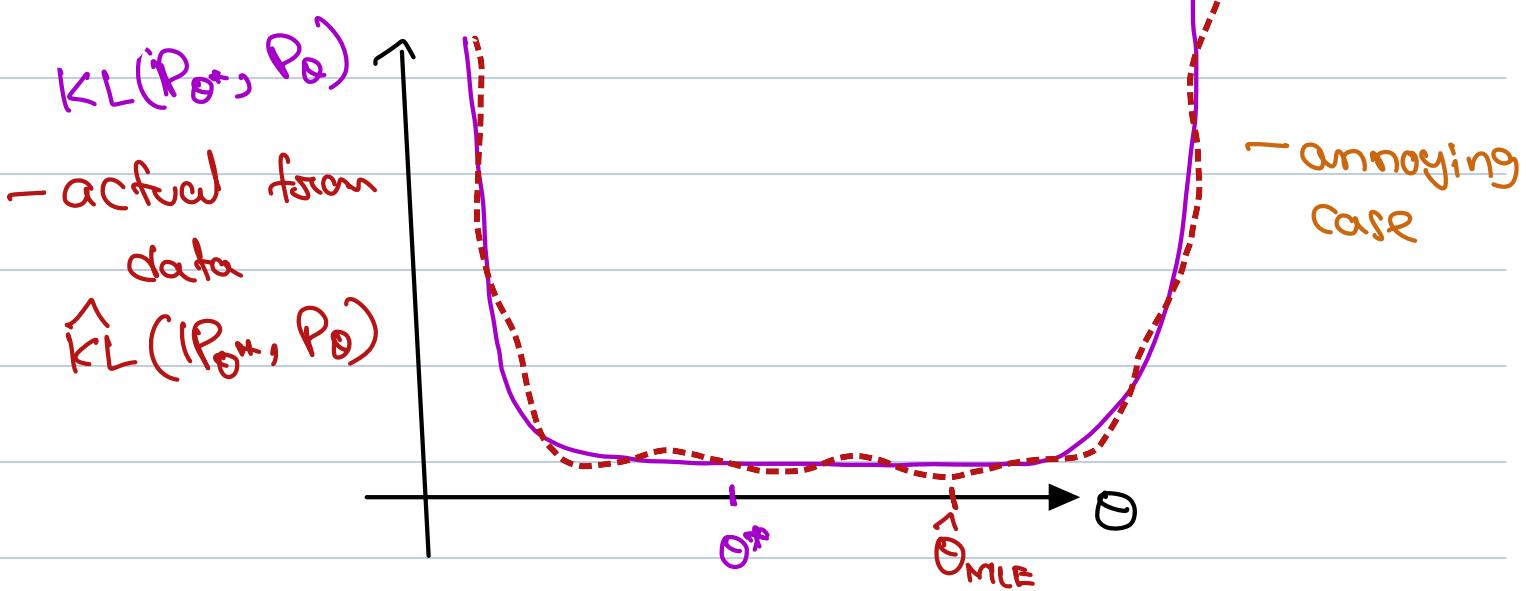
\iff minimize $\hat{KL}(P_{\theta^*}, P_{\theta})$



\Rightarrow The more data we have, the more accurate, the closer the dotted line is to the solid line. (The min θ^* is closer to $\hat{\theta}_{MLE}$)

\Rightarrow But there could be really many diagram's for some distributions.

Ex:



The fact that the plot is very flat at the bottom makes our requirements for being close to U-shaped solid curve much much more stringent,



what is the quantity that measures how curved the function is at a given point? The second derivative.

\Rightarrow The Fisher information $-E[\ell''(\theta)]$
(flipped diagram's)

The Fisher information telling us how curved our likelihood around the maximum \Rightarrow therefore telling us how good, how robust our MLE is.

This Fisher information plays a role when assessing the precision of this estimator.

How do we characterize a good estimator?
Bias, variance, combine two and form quadratic metric \Rightarrow Essentially one of these bias or variance is gonna be worse if our function $I(\theta)$ is flatter, meaning our fisher information is smaller.

$I(\theta)$ to be invertable?

if $\theta \in \mathbb{R}$ (single parameter) $I(\theta)$ is to be invertable $\Rightarrow I(\theta) \neq 0$

MIT OpenCourseWare

<https://ocw.mit.edu>

18.650 / 18.6501 Statistics for Applications

Fall 2016

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.