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

Statistician's goal: given $X_1, \ldots, X_n$, find an estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ such that $\mathbb{P}_{\hat{\theta}}$ is close to $\mathbb{P}_{\theta^*}$ for the true parameter θ^*.
This means: $\left|\mathbb{P}_{\hat{\theta}}(A) - \mathbb{P}_{\theta^*}(A)\right|$ is **small** for all $A \subset E$.

Definition

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

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

we have a sample space E and model

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

our goal is to estimate true θ^*, the one that generated data $x_1, x_2, \ldots x_n$ iid 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 distributions

$\hat{\theta} \longrightarrow \mathbb{P}_{\hat{\theta}} \longrightarrow$ candidate

$\theta^* \longrightarrow \mathbb{P}_{\theta^*} \longrightarrow$ true

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

we want $\mathbb{P}_{\hat{\theta}}([a,b]) \approx \mathbb{P}_{\theta^*}([a,b])$

$\Rightarrow$ for every event $A \subseteq E$ we want

$$\mathbb{P}_{\hat{\theta}}(A) \approx \mathbb{P}_{\theta^*}(A)$$

$\Rightarrow$ $\left| \mathbb{P}_{\hat{\theta}}(A) - \mathbb{P}_{\theta^*}(A) \right|$ is small $\forall A \subseteq E$

↑↑

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

$$TV(P_\theta, P_{\theta^*}) = \max_{A \subseteq E} \left| P_\theta(A) - P_{\theta^*}(A) \right|$$

$\Downarrow$

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:

$$0.01 \geq TV(P_{\hat{\theta}}, P_{\theta^*}) \geq \max_A \left| P_{\hat{\theta}}(A) - P_{\theta^*}(A) \right|$$

$\Rightarrow$

$$P_{\hat{\theta}}(A) \in \left[P_{\theta^*}(A) \pm 0.01 \right] \quad \forall A$$

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

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

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

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

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 event's. (There are ∞ number of them).

Neco Total variance formula :-

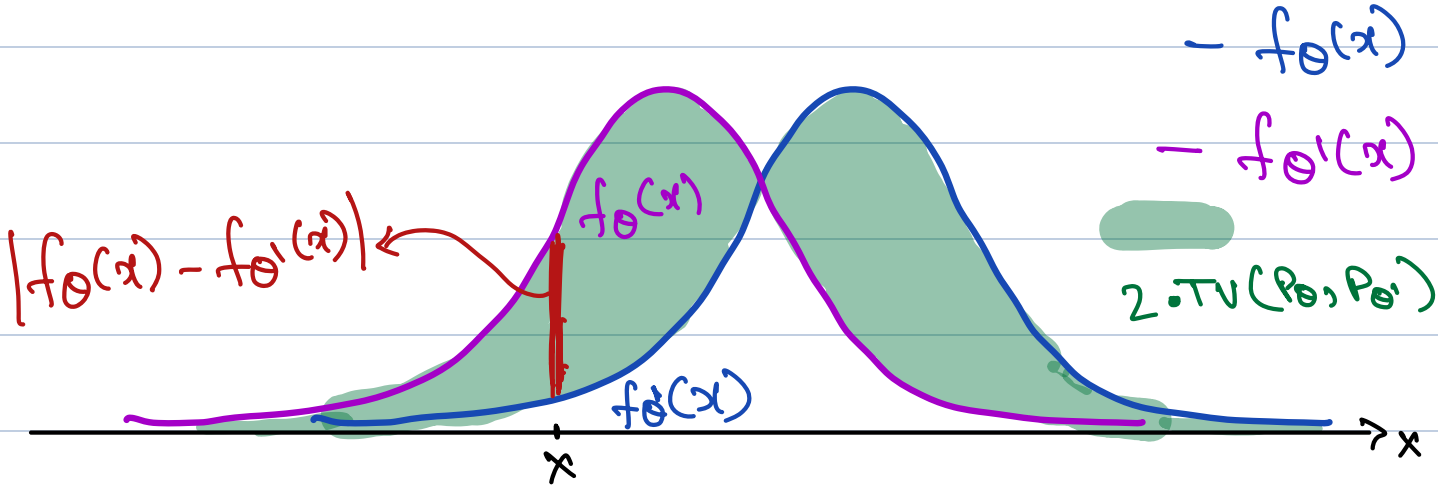
PMF :- $\Rightarrow \quad \mathbb{P}_\theta(X = x) = \mathbb{P}_\theta(x)$

$$\sum_{\forall x} \mathbb{P}_\theta(X = x) = 1 \quad , \mathbb{P}_\theta(x) \geq 0$$

$$\Rightarrow TV(\mathbb{P}_\theta, P_{\theta'}) = \frac{1}{2} \sum_{x \in E} \left| P_\theta(x) - P_{\theta'}(x) \right|$$

for Continous:

$$TV(\mathbb{P}_\theta, P_{\theta'}) = \frac{1}{2} \int_{x \in E} \left| f_\theta(x) - f_{\theta'}(x) \right| dx$$



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 $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ is a simple function of the densities f_θ and $f_{\theta'}$:

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

Total variation distance (4)

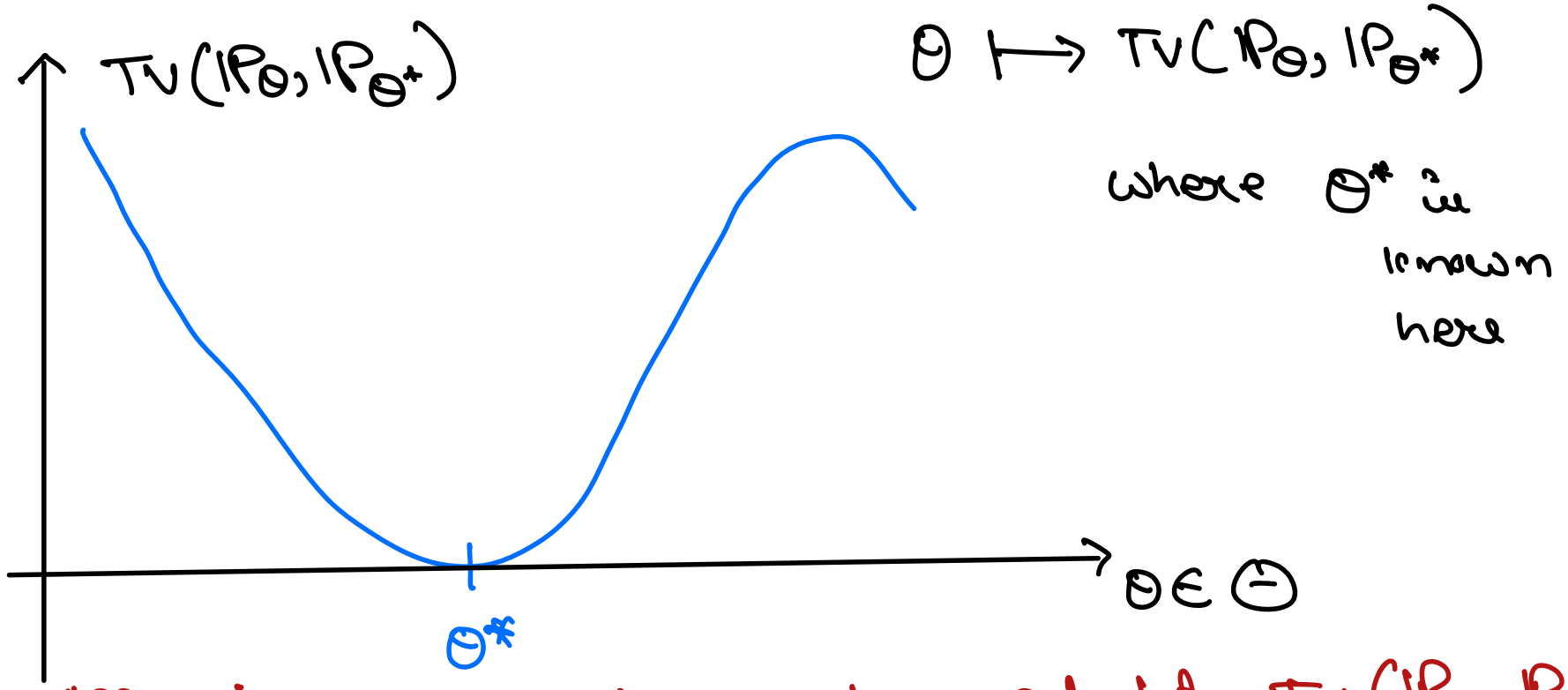
Properties of Total variation:

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

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

Total variation distance (5)

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



Assuming we θ^*, we have calculat $TV(\mathbb{P}_\theta, \mathbb{P}_{\theta^*})$ then the diagram represes 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

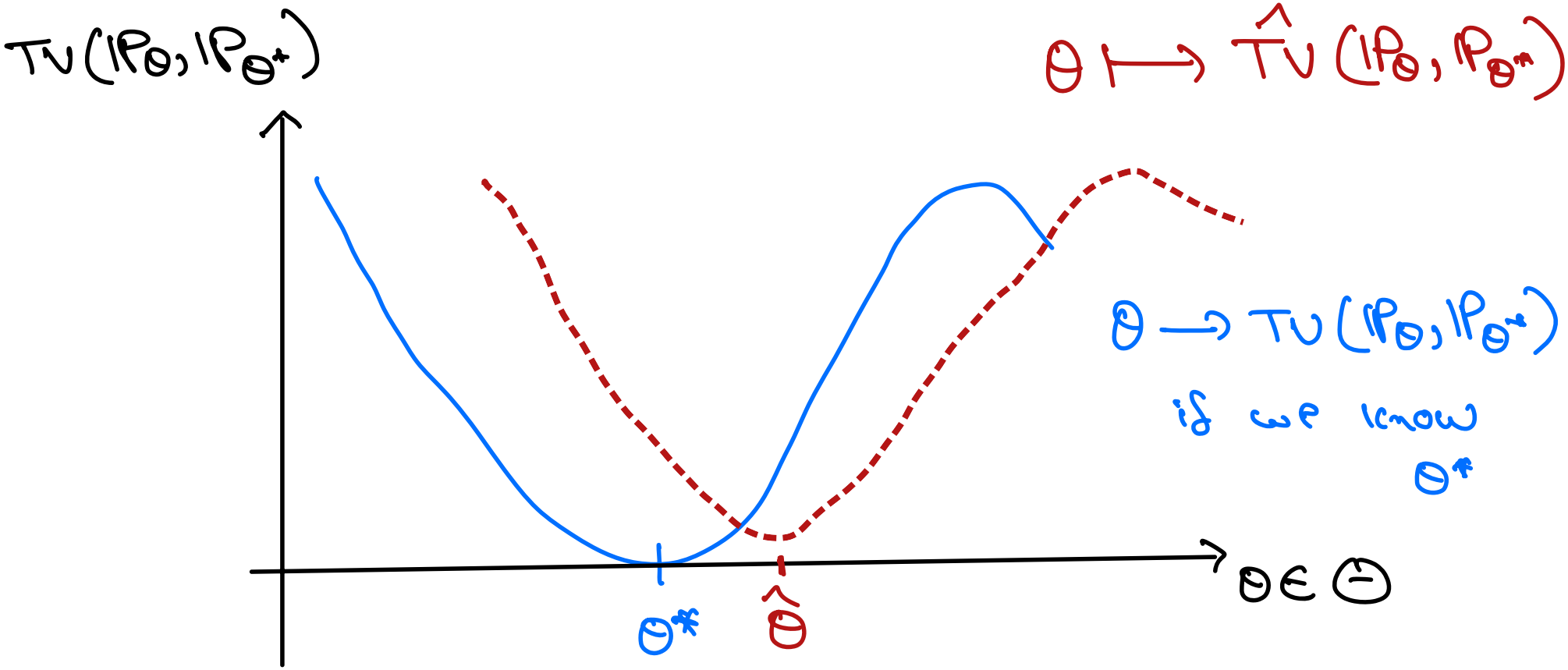
we are toying to minimize the distance b/w $\mathbb{P}_{\hat{\theta}}$, $\mathbb{P}_{\theta^*}$

* we know how to compute $TV(\mathbb{P}_{\hat{\theta}}, \mathbb{P}_{\theta^*})$ 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 $\mathbb{P}_{\theta}$, $\mathbb{P}_{\theta^*}$ for all candidate $\theta \in \Theta$.
* if this is a good Estimate, then when when we are minimizing this Estimate we get something that is closer to $\mathbb{P}_{\theta^*}$

Total variation distance (5)

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



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

we don't know the function $TV(\mathbb{P}_\theta, \mathbb{P}_{\theta^*})$ $\forall \theta \in \Theta$ 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 $TV(\mathbb{P}_\theta, \mathbb{P}_{\theta^*})$ which is $\hat{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta^*})$

The Problem is that its very unclear with How to Build this Estimator of TV i.e $\hat{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta^*})$

* So Building Estimators is typically consists of replacing Expectations by averages But there is no simple way of Expressing the TV as an Expection 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 $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ is defined by

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

Kullback-Leibler (KL) divergence (2)

Properties of KL-divergence:

- $\mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \neq \mathrm{KL}(\mathbb{P}_{\theta'}, \mathbb{P}_\theta)$ in general
- $\mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \geq 0$
- If $\mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = 0$ then $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$ (definite)
- $\mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \not\leq \mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta''}) + \mathrm{KL}(\mathbb{P}_{\theta''}, \mathbb{P}_{\theta'})$ in general

Not a distance.

This is is called a *divergence*.

Asymmetry is the key to our ability to estimate it!

Kullback-Leibler (KL) divergence (3)

$$\mathsf{KL}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta) = \mathbb{E}_{\theta^*}\Big[\log\Big(\frac{p_{\theta^*}(X)}{p_\theta(X)}\Big)\Big]$$

$$= \mathbb{E}_{\theta^*}\big[\log p_{\theta^*}(X)\big] - \mathbb{E}_{\theta^*}\big[\log p_\theta(X)\big]$$

So the function $\theta \mapsto \mathsf{KL}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta)$ is of the form:
"constant" $- \mathbb{E}_{\theta^*}\big[\log p_\theta(X)\big]$

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

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

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

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

HA HA ☺ ⇒ I am a statistician, I can replace Expectation by an average, because I have data from this distribution and try to minimize here.

$$KL(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta)$$

$$= \underbrace{\mathbb{E}_{\theta^*}[\log p_{\theta^*}(x)]}_{\text{constant, does not depend on } \theta \text{ (fixed value) (Negative entropy)}} - \underbrace{\mathbb{E}_{\theta^*}[\log p_\theta(x)]}_{\text{depends on } \theta \text{ when } \theta \text{ changes, this value changes}}$$

function $\theta \longmapsto KL(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta)$

$$= \text{Constant} - \mathbb{E}_{\theta^*}[\log(p_\theta(x))]$$

if we want to compute this we need to know Both θ^*, θ, still

Kullback-Leibler (KL) divergence (4)

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

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

This is the **maximum likelihood principle**.

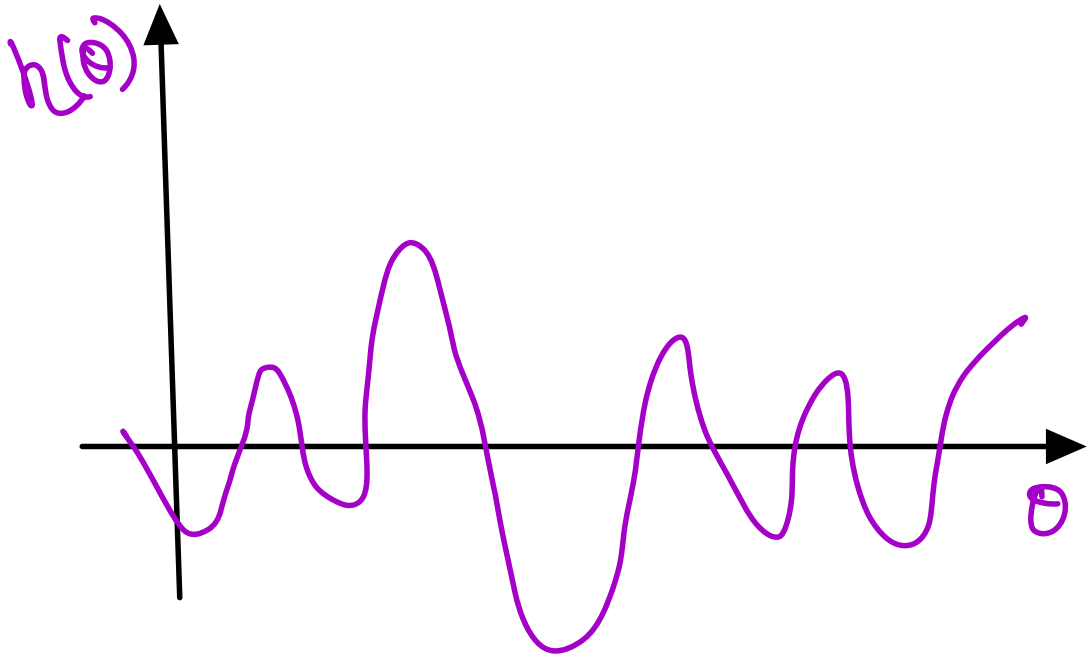
Interlude: maximizing/minimizing functions (1)

Note that

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

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

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

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

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

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 may 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 $\left(E, (\mathbb{P}_\theta)_{\theta\in\Theta}\right)$ be a statistical model associated with a sample of i.i.d. r.v. $X_1, \ldots, X_n$. 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:

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

function of all the data point's & Parameter (Θ)

Likelihood, Discrete case (2)

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

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

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

$$= \left(\frac{p}{1-p}\right)^{x_i} (1-p)^n$$

Likelihood, Discrete case (3)

Example 2 (Poisson model):

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

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

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

Likelihood, Continuous case (1)

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

Definition

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

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

Likelihood, Continuous case (2)

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

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

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

$L(x_1, x_2, \ldots, x_n, \mu, \sigma^2)$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ \frac{-1}{2\sigma^2} \left((n-1)s^2 + n(\bar{x} - \mu)^2 \right) \right\}$$

Maximum likelihood estimator (1)

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

Definition

The *likelihood estimator* of θ is defined as:

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

provided it exists.

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

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

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: $\left(\hat{\mu}_n, \hat{\sigma}_n^2\right) = \left(\bar{X}_n, \hat{S}_n\right)$.

Maximum likelihood estimator (3)

Definition: Fisher information

Define the log-likelihood for one observation as:

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

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

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

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

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

Some intution about how does the Maximum Likelihood perform?

There is something called Fisher information that essentially control's how this MLE Perform's.

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

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

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

$$I(\theta) = \mathbb{E}\left[\nabla \ell(\theta) \nabla \ell(\theta)^T\right] - \mathbb{E}[\nabla \ell(\theta)]\,\mathbb{E}[\nabla \ell(\theta)]^T$$

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

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

This quantity is very critical. what does the 2^{nd} derivative tell us at the maximum?

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

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

⇓

Flat surface

very high 2nd derivative ⇒ very curvy.

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

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

⇒ 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 $\mathbb{P}_\theta$ 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:

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

* 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 $X \sim N(0, \sigma^2)$ 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
$\log f_\theta(x) \to$ r.v
a transformation]

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

Proof

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

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

$$\Downarrow \qquad\qquad \Downarrow$$

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

$$\Downarrow$$

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

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

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

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

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

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

$$= -\int \frac{\partial^2}{\partial\theta^2} f_\theta(x)\, dx + \int \frac{\left\{\frac{\partial}{\partial\theta} f_\theta(x)\right\}^2}{f_\theta(x)}\, dx$$

$= 0$

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

$$\mathrm{Var}(\ell'(\theta)) = \mathbb{E}\left[(\ell'(\theta))^2\right] - \mathbb{E}[\ell'(\theta)]^2$$

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

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

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

$$= \mathbb{E}\left[\ell'(\theta)^2\right]$$

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

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

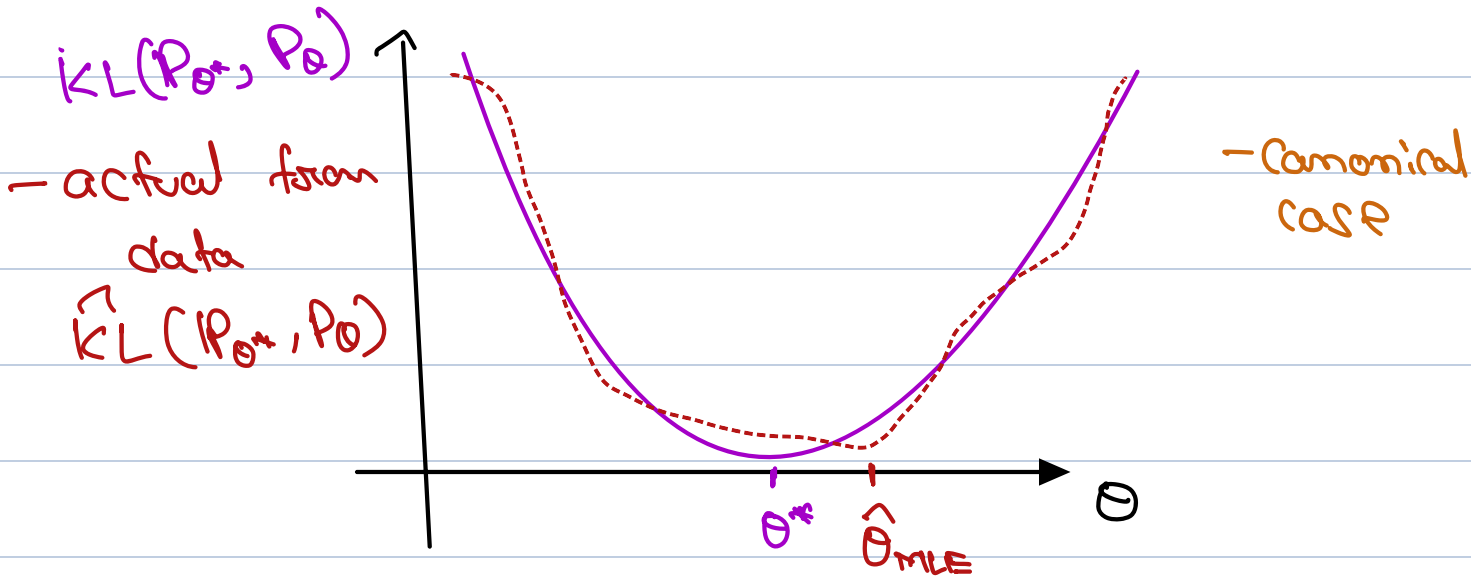
$$\text{Var}(\ell'(\theta)) = -\mathbb{E}\left[\ell''(\theta)\right]$$

Hence Proved

When we are doing MLE → it's an empirical version of trying to minimize the KL divergence.

MLE $\iff$ minimize KL divergence.

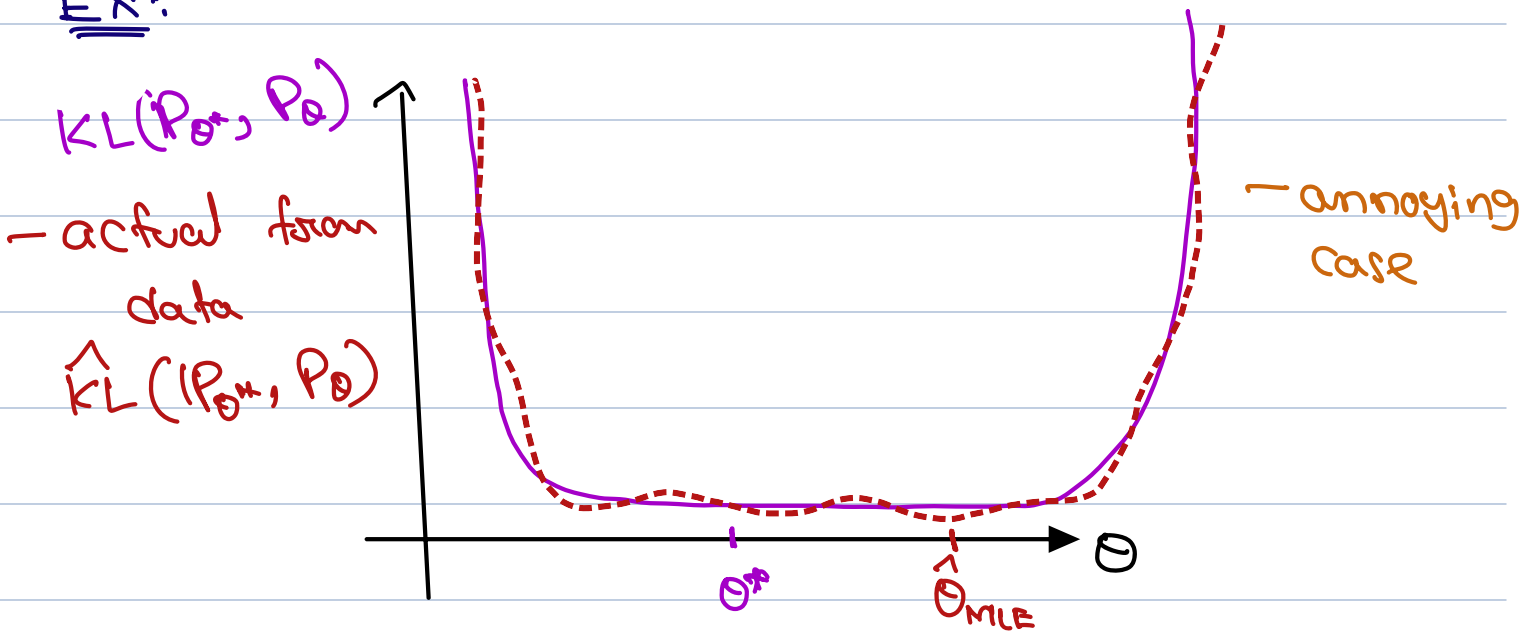
$\iff$ minimize $\widehat{KL}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta)$



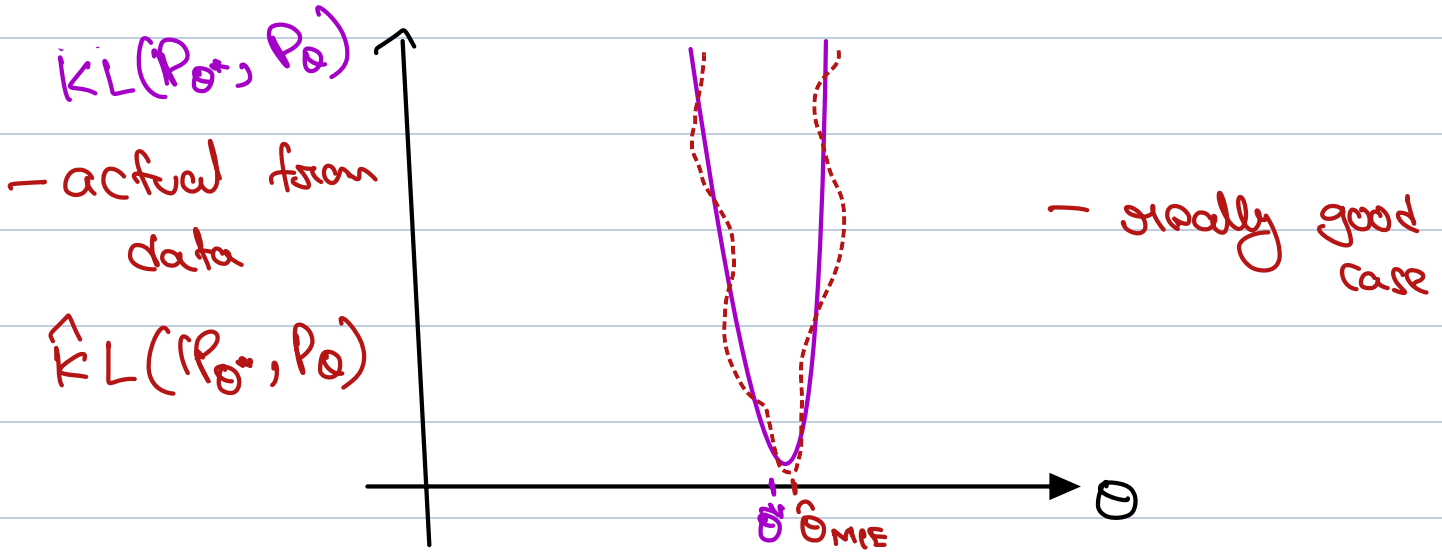
$\Rightarrow$ The more data we have, the more accurate, the closer the dotted line is to the solid line. (The min θ^* is closer to $\hat{\theta}_{MLE}$)

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

Ex:



The fact that the plot is very flat at the BoHam 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 $-\mathbb{E}[\ell''(\theta)]$

(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 Play's a role when assessing the precision of this estimator.

How do we characterize a good estimator? Bias, variance, combine two and form Quadratic risk. $\Rightarrow$ 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 $\Rightarrow$ $I(\theta) \neq 0$

18.650 / 18.6501 Statistics for Applications

Fall 2016