# 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. $X_1, \dots, X_n$. Assume that there exists $\theta^* \in \Theta$ such that $X_1 \sim \mathbb{P}_{\theta^*}$: $\theta^*$ 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 $\mathbb{P}_{\theta^*}$ for the true parameter $\theta^*$.
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

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

we have a sample spac $E$ and model $P_\theta$ $\Rightarrow$ $\left(E, \{\mathbb{P}_\theta\}_{\theta \in \Theta}\right)$

our goal is to estimate true $\theta^*$, the one that generated data $x_1, x_2, \dots x_n \overset{iid}{\sim} x$

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

* The Goal of knowing $\theta^*$ is to know what $\mathbb{P}_{\theta^*}$ is.

* Our goal is to come up with the distribution that comes from the family $P_\theta$ that is close to $\mathbb{P}_{\theta^*}$.

* So, what is it mean for two distribution's close to each other. it mean's that when we compute probabilities on one distribution, you should have probability on the other distribution pretty much.

we have 2 candidate distribution.

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

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

As a statistian 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 $A \subset E$ we want

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

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

$\Uparrow$

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

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

$\Downarrow$

worst possible event that $\theta$ is deviated from $\theta^*$

so, if you give me two Probability measures, $\mathbb{P}_\theta$, $\mathbb{P}_{\theta'}$, 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 $TV(\mathbb{P}_\theta, \mathbb{P}_{\theta'})$ is small, it mean's that for all possible A's that you give me, $\mathbb{P}_\theta(A)$ is going to be closer to $\mathbb{P}_{\theta'}(A)$

Ex:

$$0.01 \geq TV(\mathbb{P}_{\hat\theta}, \mathbb{P}_{\theta^*}) \geq \max_{A} \left| \mathbb{P}_{\hat\theta}(A) - \mathbb{P}_{\theta^*}(A) \right|$$

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

# 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 $p_\theta$ and $p_{\theta'}$:

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

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 event's. (There are $\infty$ number of them).

**New Total variance formula :-**

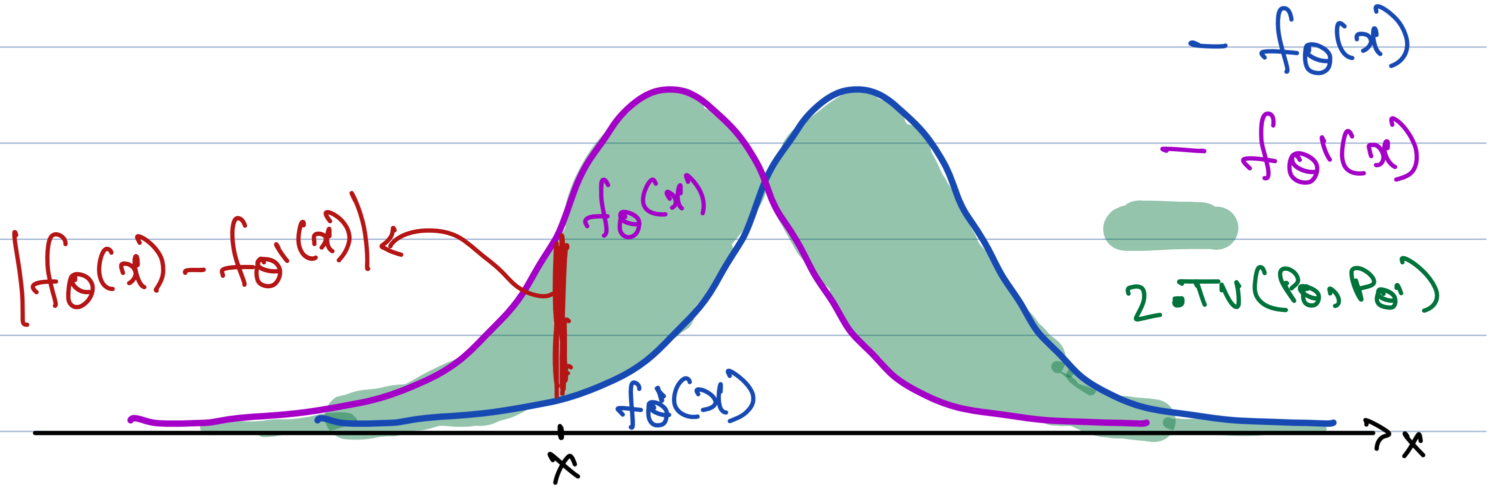
PMF :- $\Rightarrow$ $\mathbb{P}_\theta(X=x) = \mathbb{P}_\theta(x)$

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

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

for continuous:

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



— $f_\theta(x)$
— $f_{\theta'}(x)$
🟩 $2 \cdot TV(P_\theta, P_{\theta'})$
$|f_\theta(x) - f_{\theta'}(x)|$
$f_\theta(x)$
$f_{\theta'}(x)$
x
x

# 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 $f_\theta$ and $f_{\theta'}$:

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

# Total variation distance (4)

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

distance is metric

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

Properties of Total variation:

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

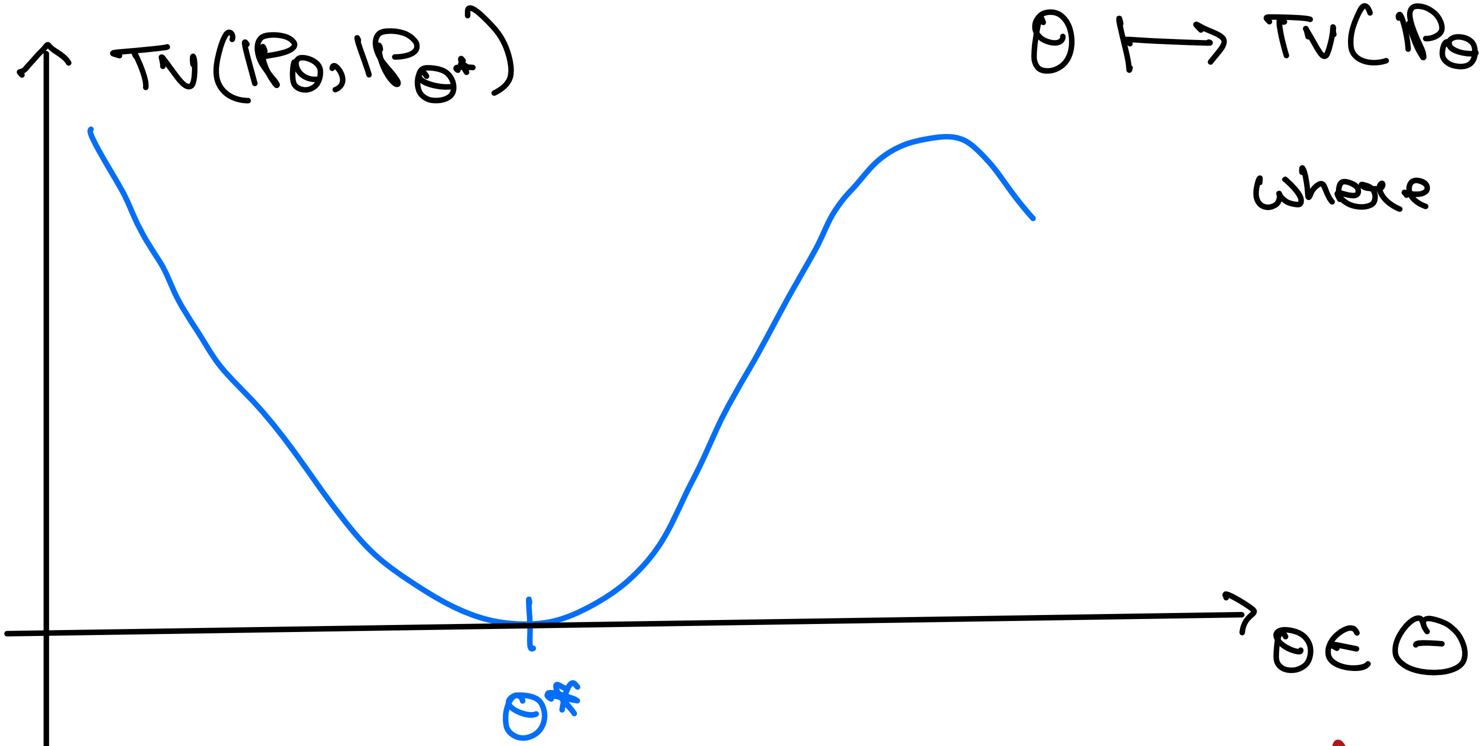
These imply that the total variation is a *distance* between probability distributions.

# 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 $\theta \mapsto \widehat{\mathrm{TV}}(\mathbb{P}_\theta, \mathbb{P}_{\theta^*})$.

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

where $\theta^*$ is known here



Assuming we $\theta^*$, we have calculat $TV(\mathbb{P}_\theta, \mathbb{P}_{\theta^*})$ then the diagram represe 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 $\theta^*$, so we cannot construct this graph we are torying to minimize the distance b/w $\mathbb{P}_{\hat{\theta}}$, $\mathbb{P}_{\theta^*}$

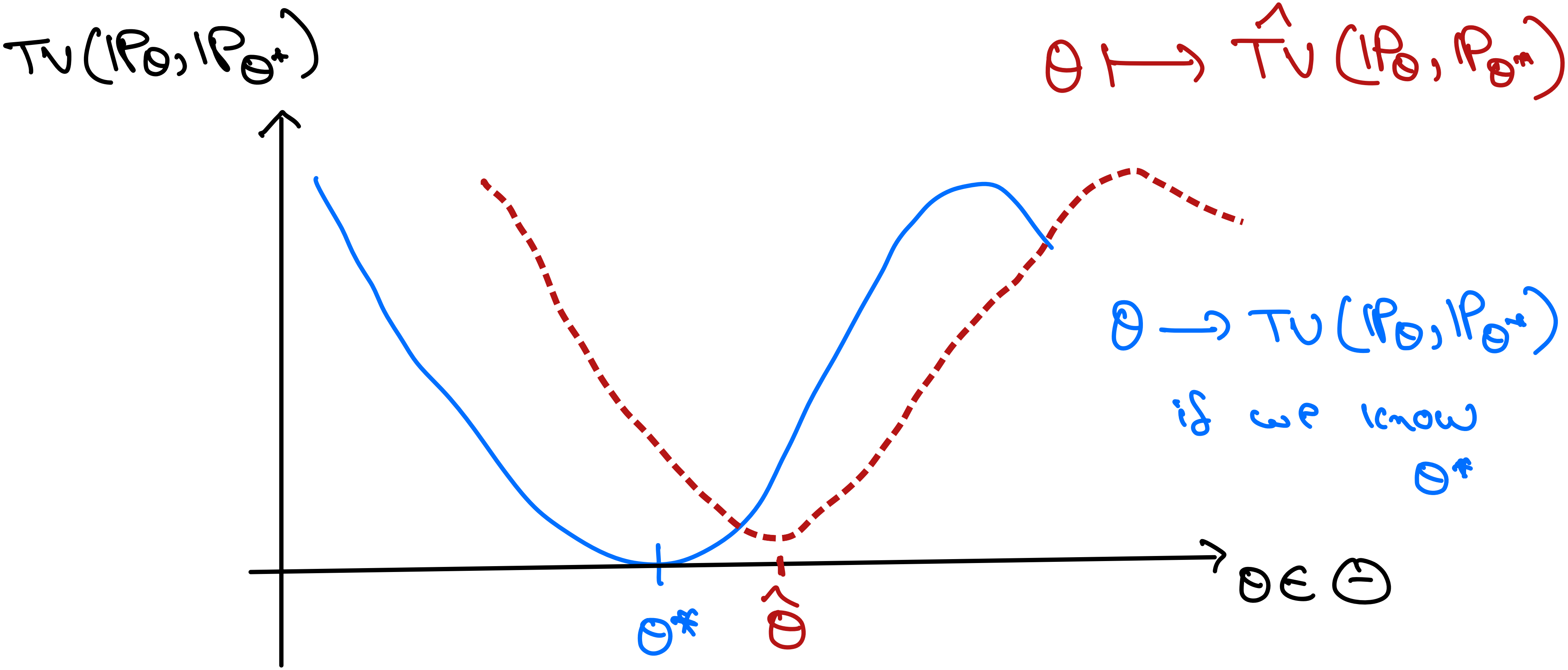
* we know How to compute $TV(\mathbb{P}_{\hat{\theta}}, \mathbb{P}_{\theta^*})$ if we know Both $\hat{\theta}, \theta^*$, But here $\theta^*$ is unknown. $\theta^*$ is not known to us. $\theta^*$ we need to Estimate

* so, Let's Build an Estimator of the TV distance b/w $\mathbb{P}_{\theta}$, $\mathbb{P}_{\theta^*}$ foor 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 $\theta^*$ (true Parameter)

$\Rightarrow$ so, we are going to estimate this distance function from data, The more the date, the better this estimator of this function $TV(\mathbb{P}_\theta, \mathbb{P}_{\theta^*})$ which is $\widehat{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta^*})$

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

* So Building Estimator's is typically consitst of replacing Expectation's by average's But there is no Simple way of Expressing the TV as an Expection w.r.t $\theta^*$ 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:

- 1 $\mathsf{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \neq \mathsf{KL}(\mathbb{P}_{\theta'}, \mathbb{P}_\theta)$ in general
- $\mathsf{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \geq 0$
- If $\mathsf{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = 0$ then $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$ (definite)
- $\mathsf{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \not\leq \mathsf{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta''}) + \mathsf{KL}(\mathbb{P}_{\theta''}, \mathbb{P}_{\theta'})$ in general

**Not a distance**.

This is is called a *divergence*.

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

# 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: $\mathbb{E}_{\theta^*}[h(X)] \rightsquigarrow \dfrac{1}{n}\sum_{i=1}^{n} h(X_i)$ (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 form which my data is actually drawn of the log of this ration

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.

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

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

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

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

if we want to compute this we need to know both $\theta^*, \theta$, 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 $\big(E, (\mathbb{P}_\theta)_{\theta\in\Theta}\big)$ 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:

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

# Likelihood, Discrete case (2)

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

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

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

# Likelihood, Discrete case (3)

**Example 2 (Poisson model):**

If $X_1, \ldots, X_n \overset{iid}{\sim}$ 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 $f_\theta$.

## Definition

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

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

# Likelihood, Continuous case (2)

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

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

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

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

# 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 $\ell$ is a.s. twice differentiable. Under some regularity conditions, the *Fisher information* of the statistical model is defined as:

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

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

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

# 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 $\theta$;
3. $\theta^*$ is not on the boundary of $\Theta$;
4. $I(\theta)$ is invertible in a neighborhood of $\theta^*$;
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. $\mathbb{P}_{\theta^*}$.