

# Parametric Models and Maximum Likelihood Estimation

Motonobu Kanagawa

Introduction to Statistics, EURECOM

March 18, 2024

# Outline

- 1 Estimation in Parametric Models
- 2 Maximum Likelihood Estimation
- 3 MLE as Kullback-Leibler Divergence Minimization
- 4 Consistency of MLE
- 5 Conclusions and Further Readings

# Density Estimation Problem

- Let  $P$  be an **unknown** probability distribution on a measurable set  $\mathcal{X} \subset \mathbb{R}^d$ .
- Assume that  $P$  has a **probability density function**  $p : \mathcal{X} \rightarrow \mathbb{R}$ .
- Given i.i.d. data  $X_1, \dots, X_n$  from the unknown  $P$ , we are interested in **estimating the density function**  $p$ .
- This is the task of **density estimation**.

## Notation

We may write  $X_1, \dots, X_n \sim p$  (i.i.d.) with the density function  $p$ .



# Density Estimation Problem

- There are mainly two approaches to this problem: **parametric** and **nonparametric**.

## Parametric approach

- Define a model of a **finite degree of freedom** for the unknown density  $p$ .
- This is called a **parametric model**, and indexed by a finite number of **parameters**.
- Assumptions of the model are often made on the **shape** of the unknown density  $p$ .
- Density estimation is done by **estimating the parameters** from the data  $X_1, \dots, X_n \sim p$ .

# Density Estimation Problem

## Nonparametric approach

- Define a model with **infinite degree of freedom**.
- **Increase the complexity** of the model **as more data become available**.
- Assumptions of the model are often made on the **smoothness** of the unknown density  $p$ .
- e.g., kernel density estimation [Silverman, 1986].

- In this course we'll only focus on the **parametric approach** (while the nonparametric approach is also important).

# Parameter Approach to Density Estimation

- In the parametric approach, we define a **parametric model** for the unknown density function  $p$  generating data  $X_1, \dots, X_n$ .

## Parametric Model

- Let  $\Theta$  be a set of **parameter vectors** (e.g.,  $\Theta \subset \mathbb{R}^q$ ).
- For each  $\theta \in \Theta$ , define a **probability density function**  $p_\theta : \mathcal{X} \rightarrow [0, \infty)$ .
- A **parametric model** is defined as the **set** of such density functions:

$$\mathcal{P}_\Theta := \{p_\theta \mid \theta \in \Theta\}.$$

# Parametric Approach to Density Estimation

## Remarks on the Term “Parametric Models”

- The parametric model can be seen as a function  $f : \mathcal{X} \times \Theta \rightarrow [0, \infty)$  such that

$$f(x, \theta) := p_\theta(x), \quad x \in \mathcal{X}, \theta \in \Theta.$$

We may say that  $f$  is a parametric model.

- Alternatively, regarding  $\theta \in \Theta$  as a variable, we also say  $p_\theta$  is a parametric model for simplicity.

# Parametric Approach to Density Estimation

- The parametric model  $\mathcal{P}_\Theta$  should be designed so that the unknown density  $p$  belongs to  $\mathcal{P}_\Theta$ , i.e.,  $p \in \mathcal{P}_\Theta$ ;

- $p \in \mathcal{P}_\Theta = \{p_\theta \mid \theta \in \Theta\}$  is equivalent to the existence of some  $\theta^* \in \Theta$  such that

$$p = p_{\theta^*} \in \mathcal{P}_\Theta.$$

- We may call such  $\theta^*$  the true parameter (vector).

- Therefore the model  $\mathcal{P}_\Theta$  should reflect our knowledge/belief about the unknown  $p$ .



# Parametric Approach to Density Estimation

- If  $p \in \mathcal{P}_\Theta = \{p_\theta \mid \theta \in \Theta\}$ , we say that the model  $\mathcal{P}_\Theta$  is **correctly specified**.
  - In this case, estimation of the unknown density  $p = p_{\theta^*}$  can be done by **estimating the true parameter  $\theta^*$**  from the data  $X_1, \dots, X_n$ .
- If  $p \notin \mathcal{P}_\Theta = \{p_\theta \mid \theta \in \Theta\}$ , we say that the model  $\mathcal{P}_\Theta$  is **misspecified**.

## Example: Gaussian Models

- Recall that the density function of a **Gaussian distribution** on  $\mathcal{X} = \mathbb{R}$  is given by

$$p_{\text{gauss}}(x; \mu, \sigma^2) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$$

where

- $\mu \in \mathbb{R}$  is the **mean** of  $p_{\text{gauss}}$
- $\sigma^2 > 0$  is the **variance** of  $p_{\text{gauss}}$ .

## Example: Gaussian Density Models

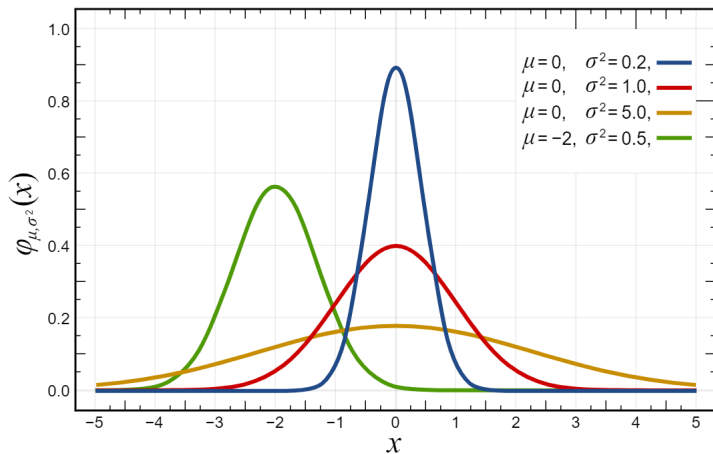


Figure 1: Gaussian density functions; From Wikipedia “Normal distribution.”

## Example: Gaussian Models

There are several ways to define a probabilistic model.

### 1. Parametrizing the mean

- Assume that we know/believe that the **variance** of the unknown density  $p$  is  $\sigma^2$ .
- Then we can define a parametric model  $p_\theta$  by **treating the mean  $\mu$  as a parameter  $\theta$** :

$$p_\theta(x) := p_{\text{gauss}}(x; \theta, \sigma^2).$$

- In this case, the parameter set may be defined as  $\Theta := [-a, a] \subset \mathbb{R}$  for some  $a > 0$ .

### Remarks

- Note that the **definition** of the parameter set  $\Theta$  is a **part of the model**.
- e.g., the choice of the interval  $[-a, a]$  implicitly represents our belief that the mean  $\mu$  of  $p$  should satisfy  $\mu \in [-a, a]$ .

## Example: Gaussian Models

### 2. Parametrizing both the mean and variance

- We can treat **both mean  $\mu$  and variance  $\sigma^2$**  as parameters.
- In this case, we can define a parametric model  $p_\theta$  as

$$P_\theta(x) := p_{\text{gauss}}(x; \theta_1, \theta_2).$$

where

$$\theta := (\theta_1, \theta_2) \in \Theta \subset \mathbb{R} \times (0, \infty).$$

- The parameter set may be defined as

$$\Theta := [-a, a] \times [b, c] \subset \mathbb{R} \times (0, \infty)$$

for some  $a, b, c > 0$ .

## Example: Gaussian Models

- By using the Gaussian model  $p_\theta$ , we implicitly makes several assumptions about the unknown  $p$ :

### Assumptions about the true $p$ made in the Gaussian model

- 1 There is only **one mode** (or the “bump”) in the density  $p$ .
- 2  $X \sim p$  may take an **arbitrarily large value**, but with an **exponentially small probability**.
- 3  $X \sim p$  takes both **positive** and **negative** values.
- 4 All the **moments** of  $p$  exist:  $-\infty < \mathbb{E}_{X \sim p}[X^k] < \infty$  for all  $k \in \mathbb{N}$ .

*$n = 1$  mean*  
 *$n = 2$  -*

- Gaussian models have been widely used in practice.

— This is because there are several **mathematically and computationally convenient** properties (we'll see this soon).

## Example: Gaussian Mixture Models

- Assume instead that we know/believe that there **two bumps** in the true density  $p$ .

- Then the use of the above Gaussian model might be inappropriate.
- We can instead consider a **two-component Gaussian mixture model**:

$$p_{\theta}(x) := \frac{1}{2}p_{\text{gauss}}(x; \theta_1, \theta_2) + \frac{1}{2}p_{\text{gauss}}(x; \theta_3, \theta_4), \quad x \in \mathbb{R}$$

where

$$\theta := (\theta_1, \theta_2, \theta_3, \theta_4) \in \Theta \subset \mathbb{R} \times (0, \infty) \times \mathbb{R} \times (0, \infty).$$

# Outline

- 1 Estimation in Parametric Models
- 2 Maximum Likelihood Estimation**
- 3 MLE as Kullback-Leibler Divergence Minimization
- 4 Consistency of MLE
- 5 Conclusions and Further Readings



# Maximum Likelihood Estimation

- **Maximum likelihood estimation (MLE)** is a classic but still widely used approach to **estimating the parameter of a parametric model**, advocated by [Fisher, 1922].
- The approach defines an estimator of the true parameter  $\theta^*$  (in the correctly specified case) as a maximizer of the **likelihood function**.

# Notation

In this lecture, we will use the notation

$$\arg \max_{\theta \in \Theta} A(\theta) = \left\{ \theta^* \in \Theta \mid A(\theta^*) = \max_{\theta \in \Theta} A(\theta) \right\}$$

as a **set** of elements in  $\Theta$  that maximize the objective function  $A(\theta)$ .

- Thus, if there are **multiple** maximizers of  $A(\theta)$ , then  $\arg \max_{\theta \in \Theta} A(\theta)$  consists of **multiple** elements.

( $\arg \min_{\theta \in \Theta} A(\theta)$  is defined in a similar way.)

# Likelihood Function

- Let  $X_1, \dots, X_n \in \mathcal{X} \subset \mathbb{R}^d$  be i.i.d. data.

## Likelihood Function

For a parametric model  $\mathcal{P}_\Theta := \{p_\theta(x) \mid \theta \in \Theta\}$ , the **likelihood function**  $\ell_n : \Theta \rightarrow [0, \infty)$  for the data  $X_1, \dots, X_n$  is defined by:

$$\ell_n(\theta) := \prod_{i=1}^n p_\theta(X_i), \quad \theta \in \Theta.$$

## Remarks

- $\ell_n(\theta)$  is a function of the parameter vector  $\theta \in \Theta$  (with  $X_1, \dots, X_n$  being fixed).
- $\ell_n(\theta)$  is **not** a **probability density function** of  $\theta \in \Theta$ .

In fact, its integral may not be 1:

$$\int \ell_n(\theta) d\theta = \int \left( \prod_{i=1}^n p_\theta(X_i) \right) d\theta \neq 1.$$

# Maximum Likelihood Estimation (MLE)

- Let  $X_1, \dots, X_n \sim p$  be i.i.d. data from the unknown density function  $p$ .
- Let  $\ell_n(\theta) := \prod_{i=1}^n p_\theta(X_i)$  be the likelihood function.

## Maximum Likelihood Estimation (MLE)

- Assume that there exists a **true parameter**  $\theta^* \in \Theta$  such that  $p = p_{\theta^*}$  (i.e., the correctly specified case  $p \in \mathcal{P}_\Theta = \{p_\theta \mid \theta \in \Theta\}$ ).
- MLE defines an **estimate**  $\hat{\theta}_n$  of the **true parameter**  $\theta^* \in \Theta$  as a solution to the following **optimization problem**:

$$\hat{\theta}_n \in \arg \max_{\theta \in \Theta} \ell_n(\theta) := \left\{ \theta' \in \Theta \mid \ell_n(\theta') = \max_{\theta \in \Theta} \ell_n(\theta) \right\}$$

- i.e., the estimate  $\hat{\theta}_n$  is a **maximizer of the likelihood function**:

$$\ell_n(\hat{\theta}_n) = \max_{\theta \in \Theta} \ell_n(\theta).$$

# Maximum Likelihood Estimation: Intuition

- We may interpret a parametric model  $p_\theta$  as a **conditional probability density function** on  $\mathcal{X}$  given  $\theta \in \Theta$ :

$$p(x | \theta) := p_\theta(x), \quad x \in \mathcal{X}, \theta \in \Theta.$$

- Thus, the likelihood function may be interpreted as the **conditional joint probability density** of i.i.d. observations  $X_1, \dots, X_n$ :

$$\ell_n(\theta) = \prod_{i=1}^n p_\theta(X_i) = \prod_{i=1}^n p(X_i | \theta).$$

- Note that the **product form** is due to the **independence** assumption of  $X_1, \dots, X_n$ .

- Thus the MLE may be interpreted as searching for the parameter vector  $\theta^*$  that **maximizes the conditional probability (density) of the data**  $X_1, \dots, X_n$ .
- This interpretation of the likelihood function becomes important in Bayesian inference (we'll see this in a coming lecture)

## MLE as Maximizing the Log Likelihood Function

- MLE can be equivalently defined as a maximizer of the **log likelihood**:

$$\hat{\theta} \in \arg \max_{\theta \in \Theta} \ell_n(\theta) = \arg \max_{\theta \in \Theta} \log \ell_n(\theta)$$

- This is because the logarithm is a **monotonically increasing function**:

$$\log(t) > \log(s) \iff t > s > 0.$$

- The log likelihood function is often easier to work with in practice, because the **product** becomes the **sum**.

$$\log \ell_n(\theta) = \log \prod_{i=1}^n p_{\theta}(X_i) = \sum_{i=1}^n \log p_{\theta}(X_i).$$

- We'll also see the use of log likelihood leads to a **deeper understanding of MLE** [Akaike, 1998].

## Example: MLE with a Gaussian Density Model

- Consider a Gaussian density model on  $\mathcal{X} = \mathbb{R}$ , with a parametrized mean  $\mu = \theta$  and a fixed variance  $\sigma^2 > 0$ :

$$p_{\theta}(x) := p_{\text{gauss}}(x; \theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \theta)^2}{2\sigma^2}\right)$$

- Assume that i.i.d. data  $X_1, \dots, X_n$  are given.

## Example: MLE with a Gaussian Density Model

- Then the log likelihood function is given as

$$\begin{aligned}\log \ell_n(\theta) &:= \sum_{i=1}^n \log p_{\text{gauss}}(X_i; \theta, \sigma^2) \\&= \sum_{i=1}^n \log \left( \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(X_i - \theta)^2}{2\sigma^2} \right) \right) \\&= \sum_{i=1}^n \left( \log \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{(X_i - \theta)^2}{2\sigma^2} \right) \\&= n \log \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right) - \sum_{i=1}^n \frac{(X_i - \theta)^2}{2\sigma^2}.\end{aligned}$$

↙ least  
squares



## Example: MLE with a Gaussian Density Model

- To obtain the maximizer, compute the derivative w.r.t.  $\theta$  and equate it to 0:

$$\frac{d \log \ell_n(\theta)}{d\theta} = \sum_{i=1}^n \frac{(X_i - \theta)}{\sigma^2} = 0.$$

- Solving this leads to the maximum likelihood estimator for the mean:

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n X_i.$$

- This is the **empirical average** of  $X_1, \dots, X_n$ !

### Exercise

- Think about **why** the empirical average is obtained as MLE for the Gaussian model.  
(Hint: recall that the empirical average can be given as a solution to the **least-squares problem**).

## Example: MLE with a Gaussian Density Model

### Exercise

- Consider the Gaussian model with both mean  $\mu = \theta_1$  and variance  $\sigma^2 = \theta_2$  parametrized:

$$p_{\theta}(x) := p_{\text{gauss}}(x; \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} \exp\left(-\frac{(x - \theta_1)^2}{2\theta_2}\right)$$

- Show that the MLE for  $\theta = (\theta_1, \theta_2)$  with i.i.d. observations  $X_1, \dots, X_n$  is given by

$$\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2) = \left( \frac{1}{n} \sum_{i=1}^n X_i, \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\theta}_1)^2 \right).$$

*Handwritten notes:* An orange arrow points from the word "mean" to  $\hat{\theta}_1$ . Another orange arrow points from the word "variance" to  $\hat{\theta}_2$ .

# Illustration

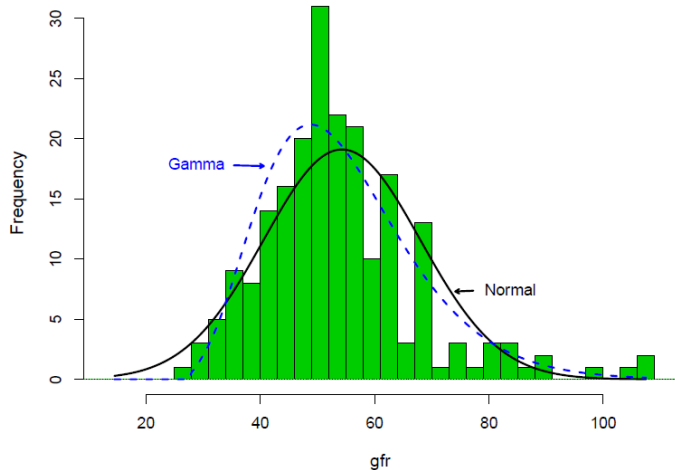


Figure 2: [Efron and Hastie, 2016, Fig 4.1]

# Maximum Likelihood Estimation (MLE)

- In general, this optimization problem of MLE has no analytical solution (e.g., consider Gaussian mixture models)
- In that case, one needs to use **numerical optimization**. e.g.,
  - Gradient descent (see the Optim course for details.)
  - Expectation-Maximization (EM) algorithm.
- In this lecture, we'll study **statistical properties** of MLE, **assuming that we can obtain the maximizer**  $\hat{\theta}_n$ .

✓  
adam  
of  
hunter

# Outline

- 1 Estimation in Parametric Models
- 2 Maximum Likelihood Estimation
- 3 MLE as Kullback-Leibler Divergence Minimization**
- 4 Consistency of MLE
- 5 Conclusions and Further Readings

# MLE as Kullback-Leibler (KL) Divergence Minimization

from Info

- Here we'll see an interpretation of MLE as searching for the parameter  $\theta \in \Theta$  that **minimizes the KL divergence** between the **true density  $p$**  and the **model density  $p_\theta$**  [Akaike, 1998].
- This interpretation is very important, because it provides an understanding of the MLE in the **misspecified case  $p \notin \mathcal{P}_\Theta$** :
- To describe this, we'll look at the definition and properties of the KL divergence.

# Kullback-Leibler (KL) Divergence

- KL divergence quantifies the **discrepancy** between two probability density functions.

## Kullback-Leibler (KL) Divergence

- Let  $p$  and  $q$  be probability density functions on  $\mathcal{X} \subset \mathbb{R}^d$  such that  $p(x)/q(x) < \infty$  for all  $x \in \mathcal{X}$ .
- Then the KL divergence between  $p$  and  $q$  is defined as

$$\begin{aligned} KL(p||q) &:= \int p(x) \log \frac{p(x)}{q(x)} dx \\ &= \int p(x) \log p(x) dx - \int p(x) \log q(x) dx. \end{aligned}$$

## Intuition: KL Divergence as a Discrepancy Measure

- If  $KL(p||q)$  is **large**, then  $p$  and  $q$  are **very different**;
- If  $KL(p||q)$  is **small**, then  $p$  and  $q$  are **similar**.

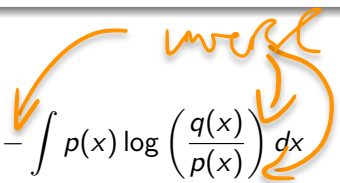
# Properties of the KL Divergence

## Nonnegativity

- The KL divergence only take **non-negative** values: for any density functions  $p$  and  $q$ ,

$$KL(p||q) \geq 0.$$

- This can be seen as follows:



$$\begin{aligned}
 KL(p||q) &= \int p(x) \log \left( \frac{p(x)}{q(x)} \right) dx = - \int p(x) \log \left( \frac{q(x)}{p(x)} \right) dx \\
 &\geq - \log \left( \int p(x) \frac{q(x)}{p(x)} dx \right) \\
 &= - \log \left( \int q(x) dx \right) = - \log(1) = 0
 \end{aligned}$$

where the inequality follows from **Jensen's inequality** and  $\log(t)$  being a convex function of  $t > 0$  (see e.g., [Berger, 1985, Sec 1.8]).



# Properties of the KL Divergence

## KL Divergence as a Discrepancy

- $KL(p||q) = 0$  if and only if  $p = q$  (almost everywhere).

-“if” part can be shown easily: If  $p = q$ , we have

$$KL(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx = \int p(x) \log(1) dx = 0.$$

## Asymmetry of the KL Divergence

- That the KL divergence is not symmetric: in general,

$$KL(p||q) \neq KL(q||p)$$

- Therefore, KL divergence is not a distance (metric) between probability density functions. (A distance measure needs to be symmetric).

# Properties of the KL Divergence

- KL divergence has its origin in [Information Theory](#).
- Indeed, the KL divergence can be written as

$$\begin{aligned}
 KL(p||q) &:= \int p(x) \log \frac{p(x)}{q(x)} dx \\
 &= - \underbrace{\int p(x) \log \frac{1}{p(x)} dx}_{\text{Entropy of } p} + \underbrace{\int p(x) \log \frac{1}{q(x)} dx}_{\text{Cross Entropy of } p \text{ and } q}
 \end{aligned}$$

difference between entropy

- For details see e.g. [Gray, 2011] and the InfoTheo course.

## Example: KL Divergence between Gaussians

- Consider the KL divergence between two **Gaussian densities**  $p$  and  $q$  on  $\mathcal{X} := \mathbb{R}$ .

### KL Divergence between Univariate Gaussians

- Let  $p(x) := p_{\text{gauss}}(x; \mu_p, \sigma_p^2)$  with mean  $\mu_p \in \mathbb{R}$  and variance  $\sigma_p^2 > 0$ ;
- Let  $q(x) := p_{\text{gauss}}(x; \mu_q, \sigma_q^2)$  with mean  $\mu_q \in \mathbb{R}$  and variance  $\sigma_q^2 > 0$ .
- Then the KL divergence between  $p$  and  $q$  is given by

$$KL(p\|q) = \frac{1}{2} \left( \frac{(\mu_p - \mu_q)^2}{\sigma_q^2} + \log \left( \frac{\sigma_q^2}{\sigma_p^2} \right) + \frac{\sigma_p^2}{\sigma_q^2} - 1 \right)$$

**Exercise.** Prove this.

## Example: KL Divergence between Gaussians

- For instance, consider the equal variance case  $\sigma_p^2 = \sigma_q^2 =: \sigma^2$ .
- Then, the KL divergence simplifies to

$$\begin{aligned} KL(p\|q) &= \frac{1}{2} \left( \frac{(\mu_p - \mu_q)^2}{\sigma^2} + \log \left( \frac{\sigma^2}{\sigma^2} \right) + \frac{\sigma^2}{\sigma^2} - 1 \right) \\ &= \frac{(\mu_p - \mu_q)^2}{2\sigma^2}. \end{aligned}$$

- We can make the following observations:
  - As difference between the means  $\mu_p$  and  $\mu_q$  approaches 0, the KL divergence converges to 0.

$$KL(p\|q) \rightarrow 0 \quad \text{as} \quad (\mu_p - \mu_q)^2 \rightarrow 0.$$

- As the variance  $\sigma^2$  increases, the KL divergence converges to 0

$$KL(p\|q) \rightarrow 0 \quad \text{as} \quad \sigma^2 \rightarrow \infty$$

# MLE as KL Divergence Minimization

- We now look at a connection between MLE and KL divergence.
- The estimate  $\hat{\theta}$  of MLE can be obtained as

$$\begin{aligned}\hat{\theta} &\in \arg \max_{\theta \in \Theta} \log \ell_n(\theta) = \arg \max_{\theta \in \Theta} \log \prod_{i=1}^n p_{\theta}(X_i) \\ &= \arg \max_{\theta \in \Theta} \sum_{i=1}^n \log p_{\theta}(X_i) = \arg \max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \log p_{\theta}(X_i).\end{aligned}$$

- The objective function in the last expression is the **empirical average** of the log density  $\log p_{\theta}(x)$  with the **i.i.d. data**  $X_1, \dots, X_n \sim p$ :

$$\frac{1}{n} \sum_{i=1}^n \log p_{\theta}(X_i).$$

## MLE as KL Divergence Minimization

- Thus, we can interpret the objective function of MLE as an **empirical approximation** to the **expected log density**:

$$\frac{1}{n} \sum_{i=1}^n \log p_{\theta}(X_i) \approx \mathbb{E}_{X \sim p}[\log p_{\theta}(X)] = \int (\log p_{\theta}(x)) p(x) dx.$$

where the expectation is with respect to the **true unknown density**,  $X \sim p$ .

- Thus, under an appropriate **identifiability condition** (introduced later), we may expect that

$$\hat{\theta}_n \approx \theta^* \in \arg \max_{\theta \in \Theta} \int p(x) \log p_{\theta}(x) dx.$$

where  $\theta^* \in \Theta$  is a maximizer of the expected log density.

- (We use the notation  $\theta^*$  as for the “true parameter” intentionally, for a reason that will be clear later).

# MLE as KL Divergence Minimization

- We show that this maximizer  $\theta^*$  is the **minimizer** of the KL divergence between the **true density**  $p$  and the **model density**  $p_\theta$ :

$$\theta^* \in \arg \max_{\theta \in \Theta} \int p(x) \log p_\theta(x) dx$$

$$= \arg \min_{\theta \in \Theta} - \int p(x) \log p_\theta(x) dx$$

$$= \arg \min_{\theta \in \Theta} - \int p(x) \log p_\theta(x) dx + \int p(x) \log p(x) dx$$

$$= \arg \min_{\theta \in \Theta} \int p(x) (-\log p_\theta(x) + \log p(x)) dx$$

$$= \arg \min_{\theta \in \Theta} \int p(x) \log \frac{p(x)}{p_\theta(x)} dx$$

$$= \arg \min_{\theta \in \Theta} KL(p \| p_\theta).$$

adding  
for sample

# MLE as KL Divergence Minimization

- Thus, we have:

$$\theta^* \in \arg \max_{\theta \in \Theta} \int p(x) \log p_{\theta}(x) dx = \arg \min_{\theta \in \Theta} KL(p \| p_{\theta}).$$

- Therefore, the estimate  $\hat{\theta}_n$  of MLE

$$\hat{\theta}_n \in \arg \max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \log p_{\theta}(X_i)$$

can be seen as an approximation to the **minimizer of the KL divergence**:

$$\theta^* \in \arg \min_{\theta \in \Theta} KL(p \| p_{\theta})$$

- We will look at closely the **conditions** required for this interpretation to be valid.
- These are conditions required for MLE to “succeed”, thus providing a guideline for the use of MLE in practice.



# Outline

- 1 Estimation in Parametric Models
- 2 Maximum Likelihood Estimation
- 3 MLE as Kullback-Leibler Divergence Minimization
- 4 Consistency of MLE**
- 5 Conclusions and Further Readings

# Consistency of MLE

- We saw that MLE may be interpreted as an **estimator** of the **optimal parameter**  $\theta^*$  given by

$$\theta^* \in \arg \max_{\theta \in \Theta} \int p(x) \log p_{\theta}(x) dx = \arg \min_{\theta \in \Theta} KL(p \| p_{\theta}).$$

- We'll investigate the consistency of the estimate  $\hat{\theta}_n$  in estimating such  $\theta^*$  in a large sample limit  $n \rightarrow \infty$ .
  - This is based on [White 82]; see this paper for details.
- The purpose is to clarify **conditions** under which MLE “works well.”
- To this end, we'll introduce several **assumptions** (= conditions).

# Assumptions on the Data Distribution

## Assumption 1 (Data and the True Density)

The data  $X_1, \dots, X_n \in \mathcal{X} \subset \mathbb{R}^d$  are i.i.d. with a distribution  $P$  with a density function  $p$ .

# Assumptions on the Parametric Model

## Assumption 2 (Model)

- The parameter set  $\Theta \subset \mathbb{R}^q$  is compact.  
— i.e.,  $\Theta$  is a **bounded** and **closed** subset.
- For every  $x \in \mathcal{X}$ , the mapping

$$\theta \rightarrow p_{\theta}(x)$$

is a **continuous function** of  $\theta \in \Theta$ .

## Consequence of the Continuity Assumption

- The **likelihood function**  $\ell_n(\theta) = \prod_{i=1}^n p_{\theta}(X_i)$  is a **continuous function** of  $\theta \in \Theta$ , because the mapping

$$\theta \rightarrow p_{\theta}(X_i)$$

is **continuous** for all  $i = 1, \dots, n$ .

# Assumptions on the Parametric Model

- Assumption 2 guarantees that the **maximum** of the likelihood function is **bounded**: i.e.,

$$\max_{\theta \in \Theta} \ell_n(\theta) = \max_{\theta \in \Theta} \prod_{i=1}^n p_{\theta}(X_i) < \infty.$$

This follows from

- 1 The likelihood function  $\ell_n(\theta)$  is a continuous function of  $\theta \in \Theta$ ;
- 2  $\Theta$  is compact;
- 3 **Extreme value theorem** (a general fact): a **continuous function** on a **compact domain** is **bounded**.

# Assumptions on the Parametric Model

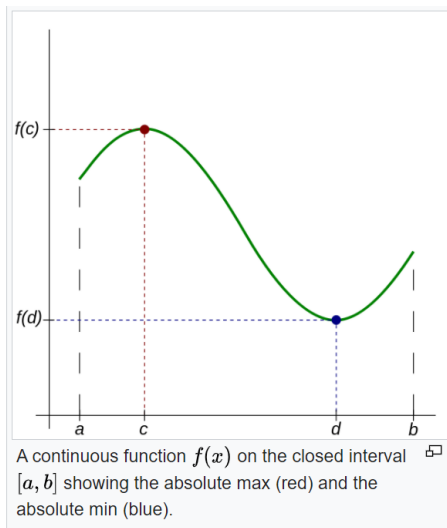


Figure 3: From Wikipedia “Extreme value theorem”

# Assumptions on the Parametric Model

- Thus, Assumption 2 guarantees that MLE

$$\hat{\theta}_n \in \arg \max_{\theta \in \Theta} \ell_n(\theta)$$

is well-defined.

- If Assumption 2 is not satisfied, then we may have

$$\max_{\theta \in \Theta} \ell_n(x) = \infty$$

- In this case, MLE  $\hat{\theta}_n \in \arg \max_{\theta \in \Theta} \ell_n(x)$  is not well-defined.

## Example where MLE is not Well-Defined

- Consider a 2-component Gaussian mixture model;

$$p_{\theta}(x) = \frac{1}{2}p_{\text{gauss}}(x; \theta_1, \theta_2) + \frac{1}{2}p_{\text{gauss}}(x; \theta_3, \theta_4),$$

with

$$\theta := (\theta_1, \theta_2, \theta_3, \theta_4) \in \Theta \subset \mathbb{R} \times (0, \infty) \times \mathbb{R} \times (0, \infty).$$

- Define the parameter set  $\Theta$  as

$$\Theta := [-a, a] \times (0, c] \times [-a, a] \times (0, c]$$

for constants  $a, c > 0$ .

- In this case,  $\Theta$  is **not closed** and thus **not compact**.
- Therefore Assumption 2 is **not satisfied**.



## Example where MLE is not Well-Defined

- We'll show that in this case the **maximum** of the likelihood function is **unbounded**:

$$\max_{\theta \in \Theta} \ell_n(\theta) = \infty,$$

and thus **MLE is not well-defined**.

## Example where MLE is not Well-Defined

- Define  $\theta_1 := X_k$  for with  $k \in \{1, \dots, n\}$  arbitrary, and fix  $\theta_3$  and  $\theta_4$ .

$$\begin{aligned} p_{\theta}(X_k) &= \frac{1}{2} p_{\text{gauss}}(X_k; \theta_1, \theta_2) + \frac{1}{2} p_{\text{gauss}}(X_k; \theta_3, \theta_4) \\ &= \frac{1}{2} \frac{1}{\sqrt{2\pi\theta_2^2}} \exp\left(-\frac{(X_k - \theta_1)^2}{2\theta_2^2}\right) + \frac{1}{2} p_{\text{gauss}}(X_k; \theta_3, \theta_4) \\ &= \frac{1}{2} \frac{1}{\sqrt{2\pi\theta_2^2}} + \frac{1}{2} p_{\text{gauss}}(X_k; \theta_3, \theta_4). \end{aligned}$$

- Taking the limit  $\theta_2 \rightarrow +0$  (i.e., the variance  $\theta_2$  going to 0), we have

$$\lim_{\theta_2 \rightarrow +0} p_{\theta}(X_k) = \lim_{\theta_2 \rightarrow +0} \left( \frac{1}{2} \frac{1}{\sqrt{2\pi\theta_2^2}} + \frac{1}{2} p_{\text{gauss}}(X_k; \theta_3, \theta_4) \right) = \infty.$$

- The limit  $\theta_2 \rightarrow +0$  can be taken, because  $\theta_2 \in (0, c]$ .

## Example where MLE is not Well-Defined

- On the other hand, for all  $i \neq k$  we have

$$\begin{aligned} p_{\theta}(X_i) &= \frac{1}{2} p_{\text{gauss}}(X_i; \theta_1, \theta_2) + \frac{1}{2} p_{\text{gauss}}(X_i; \theta_3, \theta_4) \\ &\geq \frac{1}{2} p_{\text{gauss}}(X_i; \theta_3, \theta_4). \end{aligned}$$

- Therefore,

$$\begin{aligned} \lim_{\theta_2 \rightarrow +0} \ell_n(\theta) &= \lim_{\theta_2 \rightarrow +0} \prod_{i=1}^n p_{\theta}(X_i) = \lim_{\theta_2 \rightarrow +0} p_{\theta}(X_k) \prod_{i \neq k}^n p_{\theta}(X_i) \\ &\geq \left( \lim_{\theta_2 \rightarrow +0} p_{\theta}(X_k) \right) \prod_{i \neq k} \frac{1}{2} p_{\text{gauss}}(X_i; \theta_3, \theta_4) = \infty. \end{aligned}$$

This implies that

$$\max_{\theta \in \Theta} \ell_n(\theta) \geq \lim_{\theta_2 \rightarrow +0} \ell_n(\theta) = \infty.$$

## Example where MLE is not Well-Defined

- This example shows that MLE is **not always well-defined**.
- We need to be careful about **how the parameter set  $\Theta$  is defined**.

### Exercise

Construct other examples where MLE is not well-defined.

# Assumptions for the KL Divergence to be Well-Defined

## Assumption 3 (The existence of the KL divergence)

- The true density  $p(x)$  satisfies

$$-\infty < \int p(x) \log p(x) dx < \infty.$$

- For the model  $p_\theta(x)$ , there exists a function  $g : \mathcal{X} \rightarrow [0, \infty)$  such that

$$|\log p_\theta(x)| \leq g(x) \quad \text{for all } x \in \mathcal{X} \text{ and } \theta \in \Theta$$

and

$$\int g(x)p(x)dx < \infty.$$

# Assumptions for the KL Divergence to be Well-Defined

- The latter condition implies that

$$\begin{aligned} \left| \int p(x) \log p_{\theta}(x) dx \right| &< \int p(x) |\log p_{\theta}(x)| dx \\ &\leq \int p(x) g(x) dx < \infty. \end{aligned}$$

- Therefore, the above conditions imply that the KL divergence

$$\begin{aligned} KL(p \| p_{\theta}) &= \int p(x) \log \frac{p(x)}{p_{\theta}(x)} dx \\ &= \int p(x) \log p(x) dx - \int p(x) \log p_{\theta}(x) dx \end{aligned}$$

is finite and thus well-defined.

## Exercise

- Construct examples of  $p$  and  $p_{\theta}$  for which the KL divergence **cannot be defined**.

# Assumption for the Identifiability

## Assumption 4 (Identifiability)

- Expected log density  $\int p(x) \log p_{\theta}(x) dx$  has a **unique** maximizer  $\theta^* \in \Theta$ : i.e.,

$$\int p(x) \log p_{\theta^*}(x) dx > \int p(x) \log p_{\theta}(x) dx \quad \text{for all } \theta \in \Theta \text{ with } \theta \neq \theta^*.$$

- In other words,  $\theta^*$  is the **unique** minimizer of the KL-divergence:

$$\begin{aligned} KL(p \| p_{\theta^*}) &= \int p(x) \log p(x) dx - \int p(x) \log p_{\theta^*}(x) dx \\ &< \int p(x) \log p(x) dx - \int p(x) \log p_{\theta}(x) dx = KL(p \| p_{\theta}) \\ &\quad \text{for all } \theta \in \Theta \text{ with } \theta \neq \theta^*. \end{aligned}$$

- In this case, we call the model  $P_{\Theta} = \{p_{\theta} \mid \theta \in \Theta\}$  is **identifiable** with respect to  $p$ .

## Assumption for the Identifiability

- If Assumption 4 (identifiability) is true, the notation

$$\theta^* = \arg \max_{\theta \in \Theta} \int p(x) \log p_{\theta}(x) dx = \arg \min_{\theta \in \Theta} KL(p \| p_{\theta})$$

is justified (because the “argmax” only consists of one element,  $\theta^*$ ).

- Assumption 4 enables us to **define  $\theta^*$**  as the **quantity of interest** (or the **estimand**) in statistical estimation.
- Thus, we can discuss the “consistency” of the MLE  $\hat{\theta}_n \rightarrow \theta^*$  as  $n \rightarrow \infty$ .
- This will be important in particular
  - when we are interested in the optimal parameter  $\theta^*$  **itself**; and
  - when we want to perform **hypothesis testing** regarding  $\theta^*$ .



## Interpretation of the Optimal Parameter $\theta^*$

- Let's consider what the optimal parameter  $\theta^*$  is.
- Assume that the KL divergence between the true unknown density  $p$  and the optimal model density  $p_{\theta^*}$  is zero:

$$KL(p \| p_{\theta^*}) = 0.$$

- In this case,
  - We have  $p = p_{\theta^*}$ , because  $KL(p \| p_{\theta^*}) = 0$  if and only if  $p = p_{\theta^*}$ .
  - Therefore,  $p \in \mathcal{P}_{\Theta} = \{p_{\theta} \mid \theta \in \Theta\}$  i.e., the model  $\mathcal{P}_{\Theta}$  is correctly specified.
- Thus, we can interpret  $\theta^*$  as the true parameter in this case.
- The convergence of MLE  $\hat{\theta}_n \rightarrow \theta^*$  implies that the MLE is consistent in estimating the true parameter  $\theta^*$ .

# Interpretation of the Optimal Parameter $\theta^*$

## Summary

- $KL(p||p_{\theta^*}) = 0$  corresponds to the correctly specified case  $p \in \mathcal{P}_{\Theta}$ .
- Since  $p = p_{\theta^*}$ , the optimal parameter  $\theta^*$  is interpreted as the true parameter.

# Interpretation of the Optimal Parameter $\theta^*$

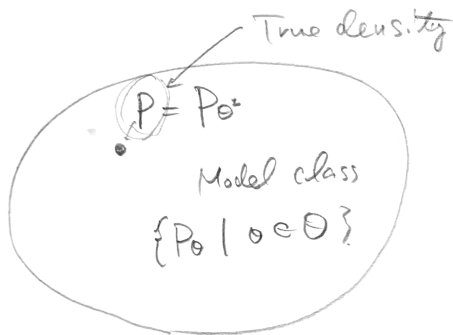


Figure 4: When  $KL(p \| p_{\theta^*}) = 0$  (correctly specified case)

## Interpretation of the Optimal Parameter $\theta^*$

- Assume the KL divergence between the true density  $p$  and the optimal model density  $p_{\theta^*}$  is **larger than zero**:

$$KL(p \| p_{\theta^*}) = \min_{\theta \in \Theta} KL(p \| p_{\theta}) > 0,$$

- In this case,
  - we have  $p \neq p_{\theta^*}$ , i.e., the optimal model density  $p_{\theta^*}$  **does not match** the true density  $p$ ;
  - thus  $p \notin \mathcal{P}_{\Theta} = \{p_{\theta} \mid \theta \in \Theta\}$ , i.e., the **model  $\mathcal{P}_{\Theta}$  is misspecified**.
- In this case, we can interpret  $p_{\theta^*}$  as the **best approximation** to the true density  $p$  as measured by the KL divergence.
- Thus, we can interpret  $\theta^*$  as the **parameter that gives the best approximation** of the model  $\mathcal{P}_{\Theta}$  to the true  $p$ .

# Interpretation of the Optimal Parameter $\theta^*$

## Summary

- $KL(p\|p_{\theta^*}) > 0$  corresponds to the misspecified case  $p \notin \mathcal{P}_\Theta$ .
- Since  $KL(p\|p_{\theta^*}) = \min_{\theta \in \Theta} KL(p\|p_\theta)$ , the optimal parameter  $\theta^*$  is interpreted as the parameter that gives the best approximation  $p_{\theta^*}$  to the true density  $p$  under the KL divergence.

# Interpretation of the Optimal Parameter $\theta^*$

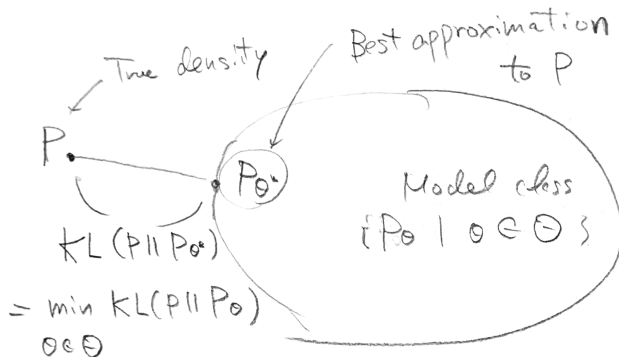


Figure 5: When  $KL(p \parallel p_{\theta^*}) > 0$  (model misspecification).

## Example where the Model is not Identifiable

- Consider a 2-component Gaussian mixture model;

$$p_{\theta}(x) = \frac{1}{2}p_{\text{gauss}}(x; \theta_1, \theta_2) + \frac{1}{2}p_{\text{gauss}}(x; \theta_3, \theta_4)$$

with

$$\theta = (\theta_1, \theta_2, \theta_3, \theta_4) \in \Theta \subset \mathbb{R} \times (0, \infty) \times \mathbb{R} \times (0, \infty).$$

- Define the parameter set  $\Theta$  by

$$\Theta := [-a, a] \times [b, c] \times [-a, a] \times [b, c]$$

for constants  $a, b, c > 0$ .

- The model is **not identifiable**, because **switching**  $(\theta_1, \theta_2)$  and  $(\theta_3, \theta_4)$  produces the **same density function**.

## Example where the Model is not Identifiable

- To show this, let

$$(\mu_1, \sigma_1^2) \in [-a, a] \times [b, c], \quad (\mu_2, \sigma_2^2) \in [-a, a] \times [b, c]$$

be arbitrary constants such that  $\sigma_1^2 \neq \sigma_2^2$ .

- Then, for  $\theta^* := (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ , we have

$$p_{\theta^*}(x) = \frac{1}{2}p_{\text{gauss}}(x; \mu_1, \sigma_1^2) + \frac{1}{2}p_{\text{gauss}}(x; \mu_2, \sigma_2^2)$$

- For  $\tilde{\theta}^* := (\mu_2, \sigma_2^2, \mu_1, \sigma_1^2)$

$$p_{\tilde{\theta}^*}(x) = \frac{1}{2}p_{\text{gauss}}(x; \mu_2, \sigma_2^2) + \frac{1}{2}p_{\text{gauss}}(x; \mu_1, \sigma_1^2)$$

- Thus, we have

$$p_{\theta^*} = p_{\tilde{\theta}^*} \quad \text{while} \quad \theta^* \neq \tilde{\theta}^*.$$

- Therefore the mixture model with this parameter set  $\Theta$  is **not identifiable**.



## Example where the Model is not Identifiable

- A simple trick to make this model identifiable is to restrict the parameter set  $\Theta$ .
- For instance, if we define the parameter set as

$$\Theta := \{(\theta_1, \theta_2, \theta_3, \theta_4) \in [-a, a] \times [b, c] \times [-a, a] \times [b, c] \mid \theta_2 < \theta_4\}$$

then the mixture model becomes identifiable.

- This corresponds to assuming that one mixture component has a smaller variance than the other.

### Exercise

Construct other examples where the model is not identifiable.

# MLE Consistency Theorem

Theorem: Consistency of MLE (Theorem 2.2 of [White, 1982])

- Suppose that Assumptions 1, 2, 3 and 4 are satisfied.
- Let

$$\hat{\theta}_n \in \arg \max_{\theta \in \Theta} \ell_n(\theta) = \arg \max_{\theta \in \Theta} \prod_{i=1}^n p_{\theta}(X_i)$$

be the MLE with i.i.d. data  $X_1, \dots, X_n \sim p$ .

- Let  $\theta^* \in \Theta$  be the optimal parameter

$$\theta^* = \arg \max_{\theta \in \Theta} \int p(x) \log p_{\theta}(x) dx = \arg \min_{\theta \in \Theta} KL(p \| p_{\theta})$$

- Then  $\hat{\theta}_n$  converges to  $\theta^*$  almost surely: i.e.,

$$\Pr(\lim_{n \rightarrow \infty} \hat{\theta}_n = \theta^*) = 1.$$

# MLE Consistency Theorem

The proof idea is that

- 1 First show that

$$\frac{1}{n} \sum_{i=1}^n \log p_{\theta}(X_i) \rightarrow \int p(x) \log p_{\theta}(x) dx \quad \text{as } n \rightarrow \infty$$

uniformly for all  $\theta \in \Theta$ .

- 2 Then conclude that

$$\hat{\theta}_n \in \arg \max_{\theta \in \Theta} \sum_{i=1}^n \log p_{\theta}(X_i) \rightarrow \theta^* = \arg \max_{\theta \in \Theta^*} \int p(x) \log p_{\theta}(x) dx.$$

as  $n \rightarrow \infty$ .

# Outline

- 1 Estimation in Parametric Models
- 2 Maximum Likelihood Estimation
- 3 MLE as Kullback-Leibler Divergence Minimization
- 4 Consistency of MLE
- 5 Conclusions and Further Readings

# Conclusions

- MLE can be understood as searching for a model density that **best approximates** the true density in terms of the **KL divergence**.
- MLE makes sense also in the **misspecified case** where the true density does not belong to the model class.
- MLE is **not always consistent**; we need **conditions = assumptions**.
- These conditions provide a **guideline** for designing your parametric model.

# Conclusions

More generic takeaways:

- A role of convergence analysis is to understand **conditions** under which **the method of interest works well**.
- Even the MLE - one of the simplest approaches - requires several conditions.
- So please always try to understand conditions under which your favorite statistical/ML method should work!

## Further Readings

- [Fisher, 1922, Section 6].
- [White, 1982]
- [Efron and Hastie, 2016, Chapter 4]



Akaike, H. (1998).

Information theory and an extension of the maximum likelihood principle.

In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer.



Berger, J. O. (1985).

*Statistical Decision Theory and Bayesian Analysis*.

Springer Science & Business Media.



Efron, B. and Hastie, T. (2016).

*Computer Age Statistical Inference*.

Cambridge University Press.



Fisher, R. A. (1922).

On the mathematical foundations of theoretical statistics.

*Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*,  
222(594-604):309–368.



Gray, R. M. (2011).



*Entropy and Information Theory.*  
Springer Science & Business Media.



Silverman, B. W. (1986).  
*Density Estimation for Statistics and Data Analysis.*  
Chapman and Hall.



Van der Vaart, A. W. (1998).  
*Asymptotic statistics.*  
Cambridge University Press.



White, H. (1982).  
Maximum likelihood estimation of misspecified models.  
*Econometrica*, pages 1–25.