Introduction to Machine Learning
Lecture 1 - Maximum Likelihood Estimator

1 Maximum Likelihood Estimator (MLE)

1.1 Introduction

1.1.1 Coin toss example

Consider an unfair coin X, that is:

$$X = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1-p \end{cases}$$

where $0 \leq p \leq 1$.
Suppose that p is unknown to us, but we have $N = 1,000$ realizations of X, that is,

$$\mathcal{D} = \{x_i\}_{i=1}^{N}$$

where x_i is the ith toss result ($x_i \in \{0, 1\}$).

Question How can we estimate p?

Solution Notice that $p = \mathbb{E}[X]$, namely, p is the expected value of the random variable X.
Hence, we can suggest the following estimator $\hat{p}$:

$$\hat{p} = \frac{1}{N} \sum_{i=1}^{N} x_i \triangleq \overline{x}$$

In words, $\hat{p}$ is the empirical mean of the set $\mathcal{D}$.
Indeed, the law of large number states that for $N \to \infty$ we have:

$$\hat{p} \underset{N \to \infty}{\longrightarrow} \mathbb{E}[X] = p$$

So $\hat{p} = \overline{x}$ is a reasonable estimator of p.

1.1.2 Die toss example

Consider an unfair die X, that is:

$$X = \begin{cases} 1 & \text{w.p. } p_1 \\ 2 & \text{w.p. } p_2 \\ 3 & \text{w.p. } p_3 \\ 4 & \text{w.p. } p_4 \\ 5 & \text{w.p. } p_5 \\ 6 & \text{w.p. } p_6 \end{cases}$$

where $p_i \geq 0$ and $\sum_i p_i = 1$.
$\{p_i\}$ are unknown, but we have $N = 1,000$ realizations of X, namely,

$$\mathcal{D} = \{x_i\}_{i=1}^{N}$$

where x_i is the ith die toss result ($x_i \in \{1, 2 \ldots, 6\}$).

Question How can we estimate p_3?

Solution Note that $p_3 \neq \mathbb{E}[X]$, so the previous method is not suitable.
Let us define the following (binary) random variable:

$$Y(X) = \begin{cases} 1 & X = 3 \\ 0 & X \neq 3 \end{cases}$$

Note that (Y is an indicator function):

$$\mathbb{E}[Y] = \Pr\{Y = 1\} = \Pr\{X = 3\} = p_3$$

by setting:

$$y_i = Y(x_i)$$

we now can write:

$$\hat{p}_3 = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{x_i = 3\}$$

In words, $\hat{p}_3$ is the ratio between the number of tosses results with $X = 3$ and the overall number of tosses.

Numeric examples

Case A Let $N = 60$ and:

	$X=1$	$X=2$	$X=3$	$X=4$	$X=5$	$X=6$
Number of occurrences	11	9	10	15	5	10

$$\Rightarrow \hat{p}_3 = \frac{1}{6}$$

Case B Let $N = 60$ and:

	$X=1$	$X=2$	$X=3$	$X=4$	$X=5$	$X=6$
Number of occurrences	5	25	6	4	13	7

$$\Rightarrow \hat{p}_3 = \frac{1}{10}$$

1.1.3 Discrete uniform random variable

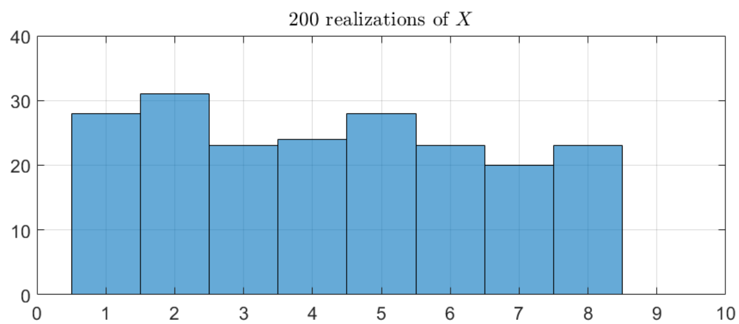
Consider the following random variable X:

$$P_X(k) \triangleq \Pr\{X = k\} = \begin{cases} \frac{1}{M} & 1 \leq k \leq M \\ 0 & \text{else} \end{cases}, \qquad M \in \mathbb{N}$$

In words, $X \in \{1, 2, 3, \ldots, M\}$ is a discrete uniform random variable.

Question Given a set of realizations $\{x_i\}_{i=1}^N$, find an estimator to M.

Solution Let us first think what will be the result for large enough N.
For example:



From this histogram, it is reasonable to deduce that $M = 8$.
Hence, in the general case we case suggest:

$$\hat{M} = \max_i \{x_i\}$$

1.2 Maximum likelihood - Definition

Instead of suggesting a specific estimator for each case, we can use a general formula.
Consider the parameter dependent probability of X:

$$P_X(k; \theta) = \Pr\{X = k; \theta\}$$

where $\theta \in \Theta$.
Given some set of realizations $\mathcal{D} = \{x_i\}$, we define the Maximum Likelihood Estimator (MLE) by:

$$\boxed{\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} P(\mathcal{D}; \theta) = \arg\max_{\theta \in \Theta} P(\{x_i\}; \theta)}$$

In words, $\hat{\theta}$ is the point which maximizes the probability to obtain the set $\mathcal{D}$.
We call $\mathcal{L}(\theta) \triangleq P(\{x_i\}; \theta)$ the **likelihood function**.

Let us revisit the previous cases and calculate the MLE for each one of them.

1.2.1 Coin toss

The probability of a single coin toss is given by:

$$P_X(x; p) = \begin{cases} p & x = 1 \\ 1 - p & x = 0 \end{cases} = p^x (1-p)^{1-x}, \qquad 0 < p < 1$$

To ignore the extreme cases, we will assume that $p \in (0, 1)$ (instead of $p \in [0, 1]$).
Given N i.i.d. (independent and identically distributed) realizations $\{x_i\}_{i=1}^N$, the likelihood function is:

$$\mathcal{L}(p) = P(\{x_i\}; p) = P(x_1, x_2, \ldots, x_N; p) = \prod_{i=1}^N P_X(x_i; p)$$

Hence, the Maximum Likelihood (ML) estimator for p is given by:

$$\begin{aligned}
\hat{p}_{ML} &= \arg\max_{0<p<1} P\left(\{x_i\};p\right) = \arg\max_{0<p<1} \prod_{i=1}^{N} P_X\left(x_i;p\right) \\
&= \arg\max_{0<p<1} \prod_{i=1}^{N} p^{x_i}\left(1-p\right)^{1-x_i} \\
&= \arg\max_{0<p<1} p^{\sum_i x_i}\left(1-p\right)^{\sum_i (1-x_i)} \\
&= \arg\max_{0<p<1} \underbrace{\log\left(p\right)\sum_i x_i + \log\left(1-p\right)\sum_i \left(1-x_i\right)}_{\triangleq \ell(p)}
\end{aligned}$$

ℓ is known as the **log-likelihood function**. We can find its maximum by comparing the derivative to zero:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}p}\ell\left(p\right) &= 0 \\
\frac{\sum_i x_i}{p} - \frac{\sum_i \left(1-x_i\right)}{1-p} &= 0, \qquad \overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \\
\frac{\overline{x}}{p} - \frac{1-\overline{x}}{1-p} &= 0 \\
p &= \overline{x}
\end{aligned}$$

$$\Rightarrow \boxed{\hat{p}_{ML} = \overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i}$$

1.2.2 Discrete uniform random variable

The probability of a single realization is given by:

$$P_X\left(x;M\right) \triangleq \begin{cases} \frac{1}{M} & 1 \le x \le M \\ 0 & \text{else} \end{cases}, \qquad M \in \mathbb{N}$$

Given N i.i.d. realizations $\{x_i\}_{i=1}^{N}$, the likelihood function is:

$$\mathcal{L}\left(M\right) = P\left(\{x_i\};M\right) = \prod_{i=1}^{N} P_X\left(x_i;M\right) = \prod_{i=1}^{N} \begin{cases} \frac{1}{M} & 1 \le x_i \le M \\ 0 & \text{else} \end{cases}$$

Hence, the Maximum Likelihood estimator for M is given by:

$$\begin{aligned}
\hat{M}_{ML} &= \arg\max_{M\in\mathbb{N}} \mathcal{L}\left(M\right) \\
&= \arg\max_{M\in\mathbb{N}} \prod_{i=1}^{N} \begin{cases} \frac{1}{M} & 1 \le x_i \le M \\ 0 & \text{else} \end{cases} \\
&= \arg\max_{M\in\mathbb{N}} \begin{cases} \left(\frac{1}{M}\right)^N & \forall i:\ 1 \le x_i \le M \\ 0 & \text{else} \end{cases}
\end{aligned}$$

Since $M \in \mathbb{N}$ is not continuous, we cannot compute the derivative with respect to M.
However, notice that the maximum value is obtained by the smallest M which satisfies: $M \ge x_i$ for all i.
Thus:

$$\Rightarrow \boxed{\hat{M}_{ML} = \max_i \{x_i\}}$$

2 Estimator Properties

Let $\hat{\theta}$ be an estimator of the parameter θ (not necessarily an MLE).
Generally, the estimator $\hat{\theta}$ is a function of the realizations, namely $\hat{\theta} = \hat{\theta}\left(\{x_i\}\right)$.
In other words, $\hat{\theta}$ can be considered as a random variable (a function of the random realizations $\{x_i\}$).

The following properties help to determine the quality of the estimator $\hat{\theta}$.

2.1 Bias

The bias of $\hat{\theta}$ is defined by:

$$b\left(\hat{\theta}\right) \triangleq \mathbb{E}\left[\hat{\theta}\right] - \theta$$

An estimator $\hat{\theta}$ with zero bias, $b\left(\hat{\theta}\right) = 0$ is called **unbiased**.
Usually, unbiased estimator are preferable.

2.2 Estimator Variance

The variance of an estimator $\hat{\theta}$ is given by:

$$V\left(\hat{\theta}\right) \triangleq \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right)^2\right]$$

2.3 Mean Squared Error

The Mean Squared Error (MSE) is given by:

$$\mathrm{MSE}\left(\hat{\theta}\right) \triangleq \mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right]$$

Question Show that:

$$\boxed{\mathrm{MSE}\left(\hat{\theta}\right) = V\left(\hat{\theta}\right) + b^2\left(\hat{\theta}\right)}$$

Solution

$$\begin{aligned}
\mathrm{MSE}\left(\hat{\theta}\right) &\triangleq \mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right] + \mathbb{E}\left[\hat{\theta}\right] - \theta\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right)^2 + 2\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right)\left(\mathbb{E}\left[\hat{\theta}\right] - \theta\right) + \left(\mathbb{E}\left[\hat{\theta}\right] - \theta\right)^2\right] \\
&= \underbrace{\mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right)^2\right]}_{=V\left(\hat{\theta}\right)} + 2\underbrace{\mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right)\right]}_{=0}\left(\mathbb{E}\left[\hat{\theta}\right] - \theta\right) + \underbrace{\left(\mathbb{E}\left[\hat{\theta}\right] - \theta\right)^2}_{b^2\left(\hat{\theta}\right)} \\
&= V\left(\hat{\theta}\right) + b^2\left(\hat{\theta}\right)
\end{aligned}$$

An estimator with low MSE is considered a good estimator.
Usually, trying to reduce the bias we will increase the variance and vice versa.
This is known as the bias-variance trade-off.

3 More MLE Examples (continuous and high-dimensional cases)

3.1 1D Gaussian (mean and covariance)

Assume:

$$X \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown. $\{x_i\}_{i=1}^{N}$ are N i.i.d realizations of X.

- Find the ML estimators of μ and σ^2.
- Are $\hat{\mu}_{ML}$ and $\hat{\sigma}^2_{ML}$ unbiased?

Solution:

$$p_X\left(x; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The likelihood function $\mathcal{L}$ is given by:

$$\mathcal{L}\left(\mu, \sigma^2\right) = P\left(\{x_i\}; \mu, \sigma^2\right) = \prod_{i=1}^{N} p_X\left(x_i; \mu, \sigma^2\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{2\pi}\right)^{\frac{N}{2}} \left(\frac{1}{\sigma^2}\right)^{\frac{N}{2}} e^{-\sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}}$$

The log-likelihood is given by:

$$\ell\left(\mu, \sigma^2\right) = \log\left(\mathcal{L}\left(\mu, \sigma^2\right)\right) = \underbrace{\log\left(\frac{1}{2\pi}\right)^{\frac{N}{2}}}_{\triangleq C} + \log\left(\left(\frac{1}{\sigma^2}\right)^{\frac{N}{2}} e^{-\sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}}\right) = C + \frac{N}{2}\log\frac{1}{\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(x_i - \mu\right)^2$$

The ML estimators is given by:

$$\hat{\mu}_{ML}, \hat{\sigma}^2_{ML} = \arg\max_{\mu, \sigma^2} \log\left(\mathcal{L}\left(\mu, \sigma^2\right)\right)$$

We can find the maximum value by comparing the derivatives (with respect to each parameter) to zero:

1. Mean:

$$\frac{\partial}{\partial \mu} \ell\left(\mu, \sigma^2\right) = 0$$

$$\frac{1}{\sigma^2} \sum_{i=1}^{N} \left(x_i - \mu\right) = 0$$

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$\Rightarrow \boxed{\hat{\mu}_{ML} = \frac{1}{N} \sum_{i=1}^{N} x_i = \overline{x}}$$

The bias of $\hat{\mu}_{ML}$ is given by:

$$\mathbb{E}\left[\hat{\mu}_{ML}\right] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[x_i\right] = \mu$$

$$\Rightarrow b\left(\hat{\mu}_{ML}\right) = \mathbb{E}\left[\hat{\mu}_{ML}\right] - \mu = 0$$

Therefore, $\hat{\mu}_{ML}$ is unbiased.

2. Variance:

$$\frac{\partial}{\partial \frac{1}{\sigma^2}} \ell\left(\mu, \sigma^2\right) = 0$$
$$\frac{N}{2}\sigma^2 - \frac{1}{2}\sum_{i=1}^{N}\left(x_i - \mu\right)^2 = 0$$
$$\Rightarrow \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \mu\right)^2$$

Since we don't know μ, we can use $\hat{\mu}_{ML}$ instead.

$$\Rightarrow \boxed{\hat{\sigma}_{ML}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \hat{\mu}_{ML}\right)^2}$$

The bias of $\hat{\sigma}_{ML}^2$:

$$\begin{aligned}
b\left(\hat{\sigma}_{ML}^2\right) &= \mathbb{E}\left[\hat{\sigma}_{ML}^2\right] - \sigma^2 \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \hat{\mu}_{ML}\right)^2\right] - \sigma^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left(x_i - \mu + \mu - \hat{\mu}_{ML}\right)^2\right] - \sigma^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left(x_i - \mu\right)^2 + 2\left(x_i - \mu\right)\left(\mu - \hat{\mu}_{ML}\right) + \left(\mu - \hat{\mu}_{ML}\right)^2\right] - \sigma^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\sigma^2 - 2\mathbb{E}\left[\left(x_i - \mu\right)\left(\frac{1}{N}\sum_{j=1}^{N}\left(x_j - \mu\right)\right)\right] + \mathbb{E}\left[\left(\frac{1}{N}\sum_{j=1}^{N}\left(x_j - \mu\right)\right)^2\right]\right) - \sigma^2 \\
&= \sigma^2 + \frac{1}{N}\sum_{i=1}^{N}\left(-2\frac{\sigma^2}{N} + \frac{1}{N^2}N\sigma^2\right) - \sigma^2 \\
&= -\frac{1}{N}\sum_{i=1}^{N}\frac{\sigma^2}{N} \\
&= -\frac{\sigma^2}{N} \\
&\neq 0
\end{aligned}$$

$\hat{\sigma}_{ML}^2$ is biased, but it is **asymptotically unbiased**: $b\left(\hat{\sigma}_{ML}^2\right) \underset{N\to\infty}{\longrightarrow} 0$

Note It is also common to use the following **unbiased** estimator:

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \hat{\mu}\right)^2$$

3.2 High dimensional Gaussian (mean estimation)

Consider the Gaussian random vector:

$$X \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{I}_{d \times d}\right), \qquad d \in \mathbb{N}$$

where $\boldsymbol{\mu} \in \mathbb{R}^d$ is unknown.
$\{\boldsymbol{x}_i\}_{i=1}^N$ are N i.i.d realizations of X.

- Find:

$$\hat{\boldsymbol{\mu}}_{ML} = ?$$

Solution:

$$p_X\left(\boldsymbol{x}; \boldsymbol{\mu}\right) = \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{\mu}\|_2^2}$$

The likelihood function $\mathcal{L}$ is given by:

$$\mathcal{L}\left(\boldsymbol{\mu}\right) = P\left(\{\boldsymbol{x}_i\}; \boldsymbol{\mu}\right) = \prod_{i=1}^N p_X\left(\boldsymbol{x}_i; \boldsymbol{\mu}\right) = \prod_{i=1}^N \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2}\|\boldsymbol{x}_i-\boldsymbol{\mu}\|_2^2} = \left(\frac{1}{(2\pi)^{\frac{d}{2}}}\right)^N e^{-\frac{1}{2}\sum_{i=1}^N \|\boldsymbol{x}_i-\boldsymbol{\mu}\|_2^2}$$

The log-likelihood is given by:

$$\ell\left(\boldsymbol{\mu}\right) = \log\left(\mathcal{L}\left(\boldsymbol{\mu}\right)\right) = \underbrace{\log\left(\frac{1}{(2\pi)^{\frac{d}{2}}}\right)^N}_{\triangleq C} + \log\left(e^{-\frac{1}{2}\sum_{i=1}^N \|\boldsymbol{x}_i-\boldsymbol{\mu}\|_2^2}\right) = C - \frac{1}{2}\sum_{i=1}^N \|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2^2$$

The ML estimator is given by:

$$\hat{\boldsymbol{\mu}}_{ML} = \arg\max_{\boldsymbol{\mu} \in \mathbb{R}^d} \log\left(L\left(\boldsymbol{\mu}\right)\right)$$

We can find the maximum value by comparing the gradient to zero $\left(\boldsymbol{0} \in \mathbb{R}^d\right)$:

$$\nabla_{\boldsymbol{\mu}} \log\left(L\left(\boldsymbol{\mu}\right)\right) = \boldsymbol{0}$$

$$\nabla_{\boldsymbol{\mu}}\left(C - \frac{1}{2}\sum_{i=1}^N \|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2^2\right) = \boldsymbol{0}$$

$$\sum_{i=1}^N \left(\boldsymbol{x}_i - \boldsymbol{\mu}\right) = \boldsymbol{0}$$

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^N \boldsymbol{x}_i$$

$$\Rightarrow \boxed{\hat{\boldsymbol{\mu}}_{ML} = \frac{1}{N}\sum_{i=1}^N \boldsymbol{x}_i \triangleq \overline{\boldsymbol{x}}}$$

This result is similar to the 1D case.

4 Classification (discrete estimation)

4.1 Two classes decision rule example

Consider the set of two possible classes:

$$\Omega = \{\omega_1, \omega_2\}$$

Given the class ω_i, the random variable X is given by:

$$X \sim \begin{cases} \mathcal{N}(0,1) & \omega = \omega_1 \\ U[0,1] & \omega = \omega_2 \end{cases}$$

That is:

$$P_X(x;\omega) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} & \omega = \omega_1 \\ u(x) & \omega = \omega_2 \end{cases}$$

where $u(x) \triangleq \begin{cases} 1 & 0 \le x \le 1 \\ 0 & \text{else} \end{cases}$.

4.1.1 Single realization

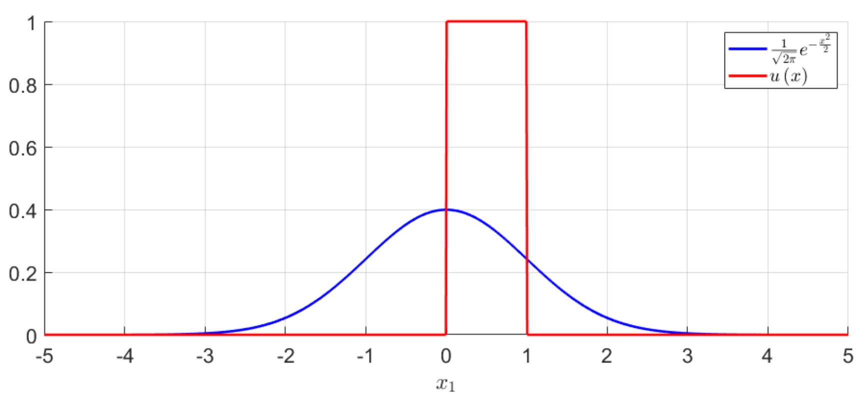
Given a single realization x_1 of X, find the MLE of ω:

$$\hat{\omega}_{ML}(x_1) = ?$$

Solution:

$$\begin{aligned} \hat{\omega}_{ML} &= \arg\max_{\omega \in \Omega} p_X(x_1;\omega) \\ &= \begin{cases} \omega_1 & p_X(x_1;\omega_1) > p_X(x_1;\omega_2) \\ \omega_2 & \text{else} \end{cases} \\ &= \begin{cases} \omega_1 & \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} > u(x) \\ \omega_2 & \text{else} \end{cases} \end{aligned}$$

A drawing can be helpful:



$$\Rightarrow \boxed{\hat{\omega}_{ML}(x_1) = \begin{cases} \omega_1 & x_1 < 0 \cup x_1 > 1 \\ \omega_2 & 0 \le x_1 \le 1 \end{cases}}$$

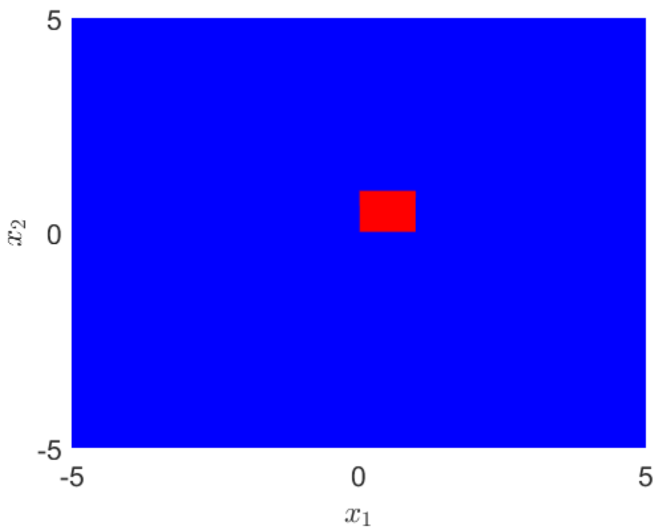
4.1.2 Two realizations

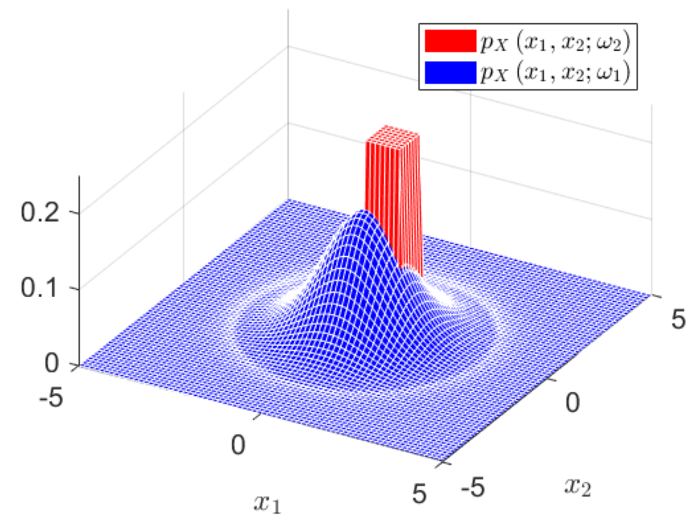
Given two i.i.d. realizations x_1 and x_2 of X.
Find the MLE of ω:

$$\hat{\omega}_{ML}\left(x_1, x_2\right) = ?$$

Solution:

$$\begin{aligned}\hat{\omega}_{ML} &= \arg\max_{\omega\in\Omega} P\left(x_1, x_2; \omega\right)\\ &= \arg\max_{\omega\in\Omega} p_X\left(x_1;\omega\right)\cdot p_X\left(x_2;\omega\right)\end{aligned}$$





$$\Rightarrow \boxed{\hat{\omega}_{ML}\left(x_1, x_2\right) = \begin{cases}\omega_2 & \left(x_1, x_2\right)\in[0,1]^2\\ \omega_1 & \text{else}\end{cases}}$$

Remark We can extend this method for any number of class $\Omega = \{\omega_1, \omega_2, \ldots, \omega_C\}$ and any number of observations $\{x_i\}_{i=1}^{N}$.

4.1.3 Two classes with a-priori probability (MAP introduction)

Consider the same setting as before:

$$X \sim \begin{cases} \mathcal{N}(0,1) & \omega = \omega_1 \\ U[0,1] & \omega = \omega_2 \end{cases}$$

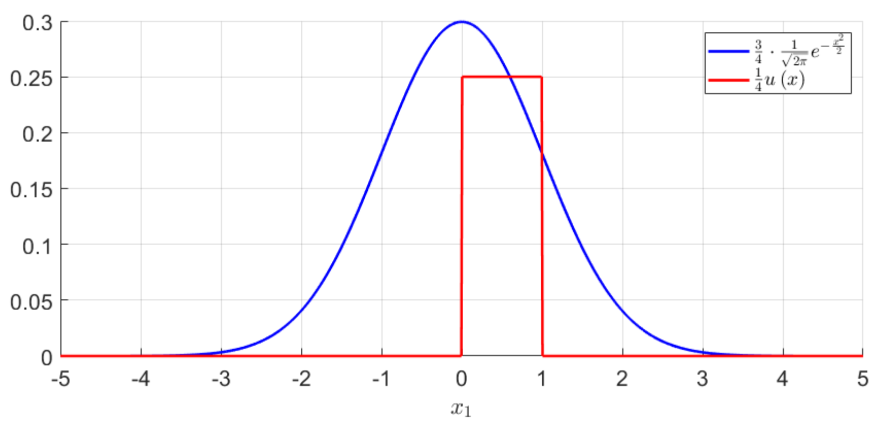
With the following a priori probability:

$$P_\Omega(\omega) = \begin{cases} \frac{3}{4} & \omega = \omega_1 \\ \frac{1}{4} & \omega = \omega_2 \end{cases}$$

Given a single realization x_1 of X, find an estimator of ω: $\hat{\omega} = ?$
Solution:
The MLE does not take into account the a-priori probability.
In that case the Maximum A Posteriori (MAP) estimator is more suitable:

$$\boxed{\hat{\omega}_{MAP} = \arg\max_{\omega\in\Omega} p_{X|\Omega}(x_1|\omega) P_\Omega(\omega)}$$

$$\begin{aligned} \hat{\omega}_{MAP}(x_1) &= \arg\max_{\omega\in\Omega} p(x_1|\omega) P_\Omega(\omega) \\ &= \begin{cases} \omega_1 & p(x_1;\omega_1) P_\Omega(\omega_1) > p(x_1;\omega_2) P_\Omega(\omega_2) \\ \omega_2 & \text{else} \end{cases} \\ &= \begin{cases} \omega_1 & \frac{3}{4}\cdot\frac{1}{\sqrt{2\pi}} e^{-\frac{x_1^2}{2}} > \frac{1}{4} u(x_1) \\ \omega_2 & \text{else} \end{cases} \end{aligned}$$



We can find the point where:

$$\frac{3}{4}\cdot\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = \frac{1}{4} u(x)$$

From the figure, we can assume $0 \le x \le 1$:

$$\begin{aligned} \frac{3}{4}\cdot\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} &= \frac{1}{4} \\ x^2 &= -2\log\left(\frac{\sqrt{2\pi}}{3}\right) \end{aligned}$$

Overall:

$$\Rightarrow \boxed{\hat{\omega}_{MAP}(x_1) = \begin{cases} \omega_1 & x_1 < \sqrt{2\log\left(\frac{3}{\sqrt{2\pi}}\right)} \cup x > 1 \\ \omega_2 & \text{else} \end{cases}}$$

5 Extra

5.1 Revisit the uniform discrete random variable

Question Is $\hat{M}$ is unbiased estimator?

Solution For simplicity, first consider $N = 1$ (therefore $\hat{M}_{ML} = x_1$, ...)

Full solution:

$$\begin{aligned}
\Pr\left\{\hat{M} \le k\right\} &= \Pr\left\{\max_i \{x_i\} \le k\right\}, \qquad k \ge 1 \\
&= \Pr\{\forall i :\, x_i \le k\} \\
&= \prod_{i=1}^{N} \Pr\{x_i \le k\} \\
&= \prod_{i=1}^{N} \min\left\{\frac{k}{M}, 1\right\} \\
&= \left(\frac{k}{M}\right)^N, \qquad \forall k \le M
\end{aligned}$$

Using the tail formula for expected value:

$$\begin{aligned}
\mathbb{E}\left[\hat{M}\right] &= \sum_{k=0}^{\infty} \Pr\left\{\hat{M} > k\right\} \\
&= 1 + \sum_{k=1}^{M-1} \Pr\left\{\hat{M} > k\right\} \\
&= 1 + \sum_{k=1}^{M-1} \left(1 - \Pr\left\{\hat{M} \le k\right\}\right) \\
&= M - \sum_{k=1}^{M-1} \left(\frac{k}{M}\right)^N \\
&\ne M
\end{aligned}$$

$\hat{M}$ is biased estimator but note that:

$$\mathbb{E}\left[\hat{M}\right] = M - \sum_{k=1}^{M-1} \left(\frac{k}{M}\right)^N \underset{N\to\infty}{\longrightarrow} M$$

$\hat{M}$ is asymptotically unbiased.

5.2 ML exercise - 2D Gaussian (covariance estimation, zero mean)

Let:

$$X \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ is a Symmetric Positive Definite (SPD) unknown matrix.
$\left\{\boldsymbol{x}_i\right\}_{i=1}^{N}$ are N i.i.d realizations of X.

- Find:

$$\hat{\boldsymbol{\Sigma}}_{ML} = ?$$

Use the following known gradients:

1.
$$\nabla_{\boldsymbol{\Sigma}^{-1}}\left(\log\left|\boldsymbol{\Sigma}\right|\right) = -\boldsymbol{\Sigma}$$

2.
$$\nabla_{\boldsymbol{\Sigma}^{-1}}\left(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right) = \boldsymbol{x}\boldsymbol{x}^T$$

Solution:

$$p_X\left(\boldsymbol{x};\boldsymbol{\Sigma}\right) = \frac{1}{2\pi}\left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} e^{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}$$

The likelihood function $\mathcal{L}$ is given by:

$$\mathcal{L}\left(\boldsymbol{\Sigma}\right) = p\left(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N; \boldsymbol{\Sigma}\right) = \prod_{i=1}^{N} p\left(\boldsymbol{x}_i;\boldsymbol{\Sigma}\right) = \prod_{i=1}^{N}\frac{1}{2\pi}\left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} e^{-\frac{1}{2}\boldsymbol{x}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_i} = \left(\frac{1}{2\pi}\right)^N \left|\boldsymbol{\Sigma}\right|^{-\frac{N}{2}} e^{-\sum_{i=1}^{N}\frac{1}{2}\boldsymbol{x}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_i}$$

The log-likelihood is given by:

$$\log\mathcal{L}\left(\boldsymbol{\Sigma}\right) = \underbrace{\log\left(\frac{1}{2\pi}\right)^N}_{\triangleq C} + \log\left(\left|\boldsymbol{\Sigma}\right|^{-\frac{N}{2}} e^{-\sum_{i=1}^{N}\frac{1}{2}\boldsymbol{x}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_i}\right) = C - \frac{N}{2}\log\left|\boldsymbol{\Sigma}\right| - \frac{1}{2}\sum_{i=1}^{N}\boldsymbol{x}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_i$$

The ML estimator is given by:

$$\hat{\boldsymbol{\Sigma}}_{ML} = \arg\max_{\boldsymbol{\Sigma}}\log\left(\mathcal{L}\left(\boldsymbol{\Sigma}\right)\right)$$

We can find the maximum value by comparing the gradient (with respect to $\boldsymbol{\Sigma}^{-1}$) to zero $\left(\mathbf{0} \in \mathbb{R}^{2\times 2}\right)$:

$$\nabla_{\boldsymbol{\Sigma}^{-1}}\log\left(\mathcal{L}\left(\boldsymbol{\Sigma}\right)\right) = \mathbf{0}$$

Using the known gradients, we have:

$$\nabla_{\boldsymbol{\Sigma}^{-1}}\log\left(\mathcal{L}\left(\boldsymbol{\Sigma}\right)\right) = \mathbf{0}$$

$$\Rightarrow \frac{N}{2}\boldsymbol{\Sigma} - \frac{1}{2}\sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i^T = \mathbf{0}$$

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i^T$$

$$\Rightarrow \hat{\boldsymbol{\Sigma}}_{ML} = \arg\max_{\boldsymbol{\Sigma}}\log\left(\mathcal{L}\left(\boldsymbol{\Sigma}\right)\right) = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i^T$$

In the general case $X \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ we have:

$$\boxed{\hat{\boldsymbol{\Sigma}}_{ML} = \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i - \boldsymbol{\mu}\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}\right)^T}$$