EE-411: Fundamentals of inference and learning **2022-2023**

Lecture 2: All of Statistics

Lecturer: Florent Krzakala

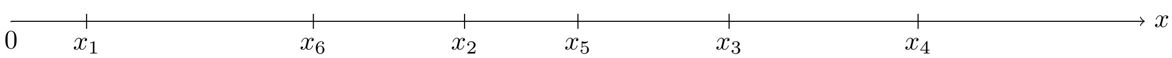
2.1 Intro

In this lecture we will learn “All of Statistics”. Let’s start with an example.

Imagine there is a source at point 0 that is emitting radioactive particles in the x-direction. Imagine that these particles are unstable and so, at some point, they are decaying and transforming to photons. When this happens a flash takes place that can be caught by a detector. We know that the probability for this to happen in an infinitesimal interval $[x, x + \mathrm{d}x]$ is $p_{\lambda^*}(x)\,\mathrm{d}x$, where the density distribution function is given by

$$p_{\lambda^*}(x) = \frac{1}{\lambda^*} e^{-x/\lambda^*}$$

where λ^* is the average distance at which the particule decays.



When the detector catches a flash its position is saved and so we collect the data $\{x_1, ..., x_6\}$ which begs the question : What is λ^* ? If we have enough points we can write

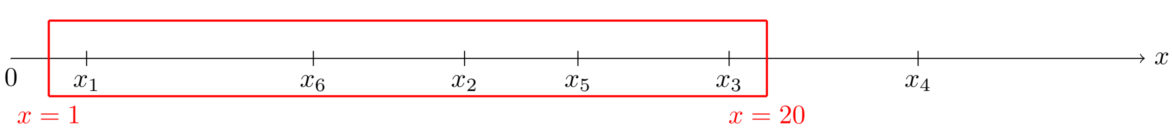
$$\hat{\lambda^*}\{x_1, ..., x_6\} = \frac{1}{6}\sum_{i=1}^{6} x_i$$

R Notation : In statistics a hat $\hat{}$ means that the variable is an estimator. An estimator is a function that takes data and then makes a guess on what the true variable is. Additionally,in this example we denote the true value of the variable with a star $*$.

By the law of large number if a lot of data is given $\{x_1, ..., x_n\}$ then we know that the estimator will be very good :

$$\hat{\lambda}_n \quad \xrightarrow[\text{n}\to\infty]{\text{LLN}} \quad \lambda^*$$

Now imagine that our detector is not good. It can only detect flashes between 1 and 20.



As a result, some flashes like x_4 will not be seen by the detector. Now how do we find an estimator $\hat{\lambda}$?

2.2 Bayesian approach

To continue with our source of radioactive particles example, we have to find the probability that we observe an element between 1 and 20 based on the true λ. Since we can only observe decays between $x = 1$ and $x = 20$ the probability is :

$$\tilde{p}_\lambda(x) = \begin{cases} \frac{1}{\mathcal{Z}(\lambda)} e^{-\frac{x}{\lambda}}, & \text{if } 1 \leq x \leq 20 \\ 0, & \text{otherwise} \end{cases}$$

And so

$$\begin{aligned} \int \tilde{p}_\lambda(x)\,\mathrm{d}x = 1 &= \int_1^{20} \frac{1}{\mathcal{Z}(\lambda)} e^{-\frac{x}{\lambda}}\,\mathrm{d}x = \frac{1}{\mathcal{Z}(\lambda)}[-\lambda e^{-\frac{x}{\lambda}}]_1^{20} \\ &= \frac{\lambda}{\mathcal{Z}(\lambda)}(e^{-\frac{1}{\lambda}} - e^{-\frac{20}{\lambda}}) \\ &\Rightarrow \mathcal{Z}(\lambda) = \lambda(e^{-\frac{1}{\lambda}} - e^{-\frac{20}{\lambda}}) \end{aligned}$$

Where $\mathcal{Z}(\lambda)$ is the normalization.

With the Bayesian approach, even though λ is a true variable that we just don't know, we treat it as a random variable and so :

$$\tilde{p}_\lambda(x) \rightarrow \tilde{p}(x|\lambda)$$

Definition 2.1 (Bayes Theorem) *Describes the probability of an event, based on prior knowledge of conditions that might be related to the event. Mathematically it is described as :*

$$\mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$$

Where A and B are 2 different events.

[B63][L74]

So back to our case, given what we see (x) we care about what is the probability that λ takes a given value. That probability based on Bayes' theorem is :

$$p(\lambda|x) = \frac{p(x|\lambda)p(\lambda)}{p(x)} = \frac{p(x|\lambda)p(\lambda)}{\mathcal{Z}}$$

R The notations for the previous statement are :

- $p(\lambda)$: Prior of $\lambda \to$ Probability that λ takes a certain value before doing the experiment.
- $p(x|\lambda)$: Likelihood of λ
- $p(\lambda|x)$: Posterior of $\lambda \to$ Probability of λ after doing the experiment and taking the data.

2.3 Theory of maximum likelihood

Definition 2.2 (Maximum likelihood) *is a method of estimating the parameters of an assumed probability distribution, given some observed data*

$$\hat{\lambda}_{ML}(\{x\}) = \underset{\lambda}{\operatorname{argmax}}\, p(\{x\}|\lambda) = \underset{\lambda}{\operatorname{argmax}}\, \{\log p(\{x\}|\lambda)\}$$

2.3.1 Consistency of maximum likelihood

Proving that maximum likelihood is consistent is equivalent to proving that the estimator converges to the truth with a certain probability (N being the number of data):

$$\hat{\lambda}_{ML}(\{x\}) \quad \xrightarrow[\text{n}\to\infty]{\text{LLN}} \quad \lambda^*$$

where λ^* is the ground truth for λ.
Restricting to the case of n i.i.d. random variables, we define the two following functions:

$$\ell(\lambda, x) = \log p_\lambda(x) = \log p(x|\lambda)$$
$$\mathcal{L}_n(\lambda, \{x\}) = \frac{1}{n}\sum_{i=1}^{n} \log p_\lambda(x_i)$$

Therefore, we can write

$$\hat{\lambda}_{ML} = \underset{\lambda}{\operatorname{argmax}}\, \mathcal{L}_n(\lambda|\{x\})$$

$$\begin{aligned}
\mathcal{L}_n(\lambda, \{x\}) - \mathcal{L}_n(\lambda^*, \{x\}) &= \frac{1}{n}\sum_{i=1}^{n} \log p_\lambda(x_i) - \frac{1}{n}\sum_{i=1}^{n} \log p_{\lambda^*}(x_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_\lambda(x_i)}{p_{\lambda^*}(x_i)}
\end{aligned}$$

Applying the law of large numbers:

$$\begin{aligned}
\mathcal{L}_n(\lambda, \{x\}) - \mathcal{L}_n(\lambda^*, \{x\}) \xrightarrow[\text{n}\to\infty]{\text{LLN}} \qquad E\left[\log \frac{p_\lambda(x)}{p_{\lambda^*}(x)}\right] &= \int \mathrm{d}x\, p_{\lambda^*}(x) \log \frac{p_\lambda(x)}{p_{\lambda^*}(x)} \\
&= -\int \mathrm{d}x\, p_{\lambda^*}(x) \log \frac{p_{\lambda^*}(x)}{p_\lambda(x)} \\
&= -D_{KL}\Big(p_{\lambda^*}(x) \| p_\lambda(x)\Big) \\
&\leq 0
\end{aligned}$$

where we have defined the Kullback-Leibler divergence D_{KL}.

Definition 2.3 (Kullback-Leibler divergence) *is a measure of how one probability distribution differs from another.*

$$D_{KL}(\mathbb{P}\|\mathbb{Q}) = \sum_i \mathbb{P}(x_i)\log\frac{\mathbb{P}(x_i)}{\mathbb{Q}(x_i)} \qquad \textit{for probabilities}$$

$$D_{KL}(p\|q) = \int_{-\infty}^{+\infty} dx p(x)\log\frac{p(x)}{q(x)} \qquad \textit{for probability distributions}$$

We note that the KL divergence is always positive : $D_{KL}(p||q) \geq 0$ *(Gibbs inequality see lecture notes 1)*
If it is zero the two probabilities/distributions are equal : $D_{KL}(p||q) = 0 \implies p(x) = q(x) \quad \forall x$

Here we see that the expected log-likelihood $\mathbb{E}\mathcal{L}(\lambda, x)$ is maximized for $\lambda = \lambda^*$. Given that the empirical log-likelihood $\mathcal{L}_n(\lambda, \{x\})$ converges point-wise to the expected log-likelihood, for large enough n, we thus expect[1] that the empirical log-likelihood $\mathcal{L}_n(\lambda, \{x\})$ to be also maximized at λ^*. This implies that maximum likelihood is consistent:

$$\hat{\lambda}_{ML} \to \lambda^*, \mathrm{as\, n} \to \infty \tag{2.1}$$

2.4 Fundamental limits of learning

2.4.1 What are the limits for finite n?

Definition 2.4 (Mean Squared Error (MSE))

$$\begin{aligned}
MSE = \underset{x}{E}\left[\left(\hat{\lambda}(\{x\}) - \lambda^*\right)^2\right] &= \mathbb{E}\left[\hat{\lambda}(\{x\})^2\right] - 2\mathbb{E}\left[\hat{\lambda}(\{x\})\lambda^*\right] + \mathbb{E}\left[(\lambda^*)^2\right] \\
&= \mathbb{E}\left[\hat{\lambda}(\{x\})^2\right] - \mathbb{E}\left[\hat{\lambda}(\{x\})\right]^2 + \mathbb{E}\left[\hat{\lambda}(\{x\})\right]^2 - 2\mathbb{E}\left[\hat{\lambda}(\{x\})\right]\lambda^* + (\lambda^*)^2 \\
&= \left(\mathbb{E}\left[\hat{\lambda}(\{x\})^2\right] - \mathbb{E}\left[\hat{\lambda}(\{x\})\right]^2\right) + \left(\mathbb{E}\left[\hat{\lambda}(\{x\})\right] - \lambda^*\right)^2 \\
&= Var[\hat{\lambda}\{x\}] + b(\hat{\lambda}\{x\}, \lambda^*)^2
\end{aligned}$$

We defined two quantities: The bias $b(\hat{\lambda}\{x\}, \lambda^*) = \left(\mathbb{E}\left[\hat{\lambda}\{x\} - \lambda^*\right]\right)$ *and the variance* $Var[\hat{\lambda}\{x\}] = \mathbb{E}\left[\hat{\lambda}\{x\}^2\right] - \mathbb{E}\left[\hat{\lambda}\{x\}\right]^2$

Definition 2.5 (Cramer-Rao bound) *is a lower bound on the variance of unbiased estimators.*

$$MSE \geq b(\hat{\lambda}\{x\}, \lambda^*)^2 + \frac{1}{nI(\lambda^*)}\left(1 + \partial_{\lambda^*} b(\hat{\lambda}\{x\}, \lambda^*)\right)^2$$

If $\hat{\lambda}$ *is unbiased* $\left(\mathbb{E}[\hat{\lambda}] = \lambda^*\right)$ *then :* $MSE \geq \frac{1}{nI(\lambda^*)}$

To prove the Cramer-Rao bound, we first introduce the Fisher score:

[1] Note that this argument is not rigorous: we should require a uniform convergence instead of the point-wise one. Indeed the maximum of the limit is not assured to be the limit of of the maximum unless the convergence is uniform in λ.

Definition 2.6 (Fisher score and information) *is the derivative of the log-likelihood.*

$$S_\lambda(x) = \frac{\partial}{\partial\lambda}(log\,(P_\lambda(x))\,) \qquad \textit{Fisher score}$$
$$S_{n,\lambda}(\{x\}) = \sum_{i=1}^{n} \frac{\partial}{\partial\lambda}(log\,(P_\lambda(x_i))\,) \qquad \textit{Total Fisher score}$$

We derive the following properties of the Fisher score:

$$\begin{aligned} E\left[S_{\lambda^*}(x)\right] &= 0 \\ E\left[S^2_{\lambda^*}(x)\right] &= -E\left[\frac{\partial}{\partial\lambda}S_\lambda(x)\Big|_{\lambda^*}\right] \qquad \textit{Fisher information} \\ &= I(\lambda^*) \\ E\left[S^2_{n,\lambda^*}(\{x\})\right] &= nI(\lambda^*) \qquad \textit{Fisher information additivity} \end{aligned}$$

The first property can be deduced as follows:

$$\begin{aligned} \int \mathrm{d}x p_{\lambda^*}(x)\partial_\lambda \log p_\lambda(x)\Big|_{\lambda^*} &= \int \mathrm{d}x p_{\lambda^*}(x)\frac{\partial_\lambda p_\lambda(x)}{p_{\lambda^*}(x)}\Big|_{\lambda^*} \\ &= \partial_\lambda \underbrace{\int \mathrm{d}x p_{\lambda^*}(x)}_{1} = 0 \end{aligned}$$

and the second from

$$\begin{aligned} \int \mathrm{d}x p_{\lambda^*}(x)\partial^2_\lambda \log p_\lambda(x)\Big|_{\lambda^*} &= \int \mathrm{d}x p_{\lambda^*}(x)\frac{\partial^2_\lambda p_\lambda(x) p_\lambda(x) - (\partial_\lambda p_\lambda(x))^2}{p^2_{\lambda^*}(x)}\Big|_{\lambda^*} \\ &= \partial^2_\lambda \underbrace{\int \mathrm{d}x p_\lambda(x)}_{1} - \int \mathrm{d}x p_{\lambda^*}(x)(\partial_\lambda \log p_\lambda(x))^2\Big|_{\lambda^*} \end{aligned}$$

Using the Cauchy–Schwarz inequality, can now tackle the proof of the Cramer-Rao bound:

$$\begin{aligned} \mathrm{Cov}^2\left(\hat\lambda(\{x\}), S_{n,\lambda^*}(\{x\})\right) &\leq \mathrm{Var}(\hat\lambda(\{x\}))\cdot \underbrace{\mathrm{Var}\left(S_{n,\lambda^*}\right)}_{nI(\lambda^*)} \\ \mathrm{Var}\left(\hat\lambda\right) &\geq \frac{\mathrm{Cov}^2\left(\hat\lambda, S_{n,\lambda^*}\right)}{nI(\lambda^*)} \\ &= \frac{\left(\mathbb{E}\left[\hat\lambda S_{n,\lambda^*}\right] - E\left[\hat\lambda\right]\overbrace{\mathbb{E}\left[S_{n,\lambda^*}\right]}^{0}\right)^2}{nI(\lambda^*)} \\ &= \frac{\mathbb{E}\left[\hat\lambda S_{n,\lambda^*}\right]^2}{nI(\lambda^*)} \end{aligned}$$

$$
\begin{aligned}
1+\frac{\partial b(\hat{\lambda}\{x\},\lambda^*)}{\partial\lambda^*} &= \partial_{\lambda^*}\mathbb{E}(\hat{\lambda}) \\
&= \partial_{\lambda^*}\int \mathrm{d}x p_{\lambda^*}(x)\hat{\lambda}(x) \\
&= \int \mathrm{d}x(\partial_{\lambda^*}p_{\lambda^*}(x))\hat{\lambda}(x) \\
&= \int \mathrm{d}x p_{\lambda^*}(x)[\frac{\partial_\lambda p_{\lambda^*}(x)}{p_{\lambda^*}(x)}]\hat{\lambda}(x) \\
&= \int \mathrm{d}x p_{\lambda^*}(x)\frac{\partial(\log p_{\lambda^*}(x))}{\partial\lambda^*}\hat{\lambda}(x) \\
&= \mathbb{E}[S_{n,\lambda^*}\hat{\lambda}]
\end{aligned}
$$

2.4.2 How things change when n is large

What we want to show in this section, is that when we have a lot of data (when n is large) the maximum likelihood estimator is saturating the Cramer-Rao bound.

We assume that $\hat{\lambda}$ is asymptotically unbiased, therefore there is no bias since n is large and $MSE(\hat{\lambda}) \geq \frac{1}{nI(\lambda^*)}$

First let's start by taking the log-likelihood:

$$
\begin{aligned}
\mathcal{L}_n(\{x\},\lambda) &= \frac{1}{n}\sum_{i=1}^{n}\log p_\lambda(x|\lambda) \\
=> \hat{\lambda}_{ML} &= \operatorname*{argmax}_\lambda \mathcal{L}_n(\{x_i\},\lambda)
\end{aligned}
$$

By the definition of $\hat{\lambda}$ maximum likelihood we know that the first derivative of the log-likelihood evaluated at $\hat{\lambda}_{ML}$ is equal to zero. Therefore:

$$\dot{\mathcal{L}}_n(\{x\},\hat{\lambda}_{ML}) = 0$$

When n is large we consider that $\hat{\lambda}_{ML}$ is not very far from λ^*, we can therefore perform the Taylor expansion with respect to $\hat{\lambda}_{ML}$:

$$\dot{\mathcal{L}}_n(\{x\},\hat{\lambda}_{ML}) = \dot{\mathcal{L}}_n(\{x\},\lambda^*) + (\hat{\lambda}_{ML}-\lambda^*)\ddot{\mathcal{L}}_n(\{x\},\lambda^*) + \mathcal{O}(\hat{\lambda}_{ML}-\lambda^*)^2$$

By re-arranging the two previous equations, we finally obtain that:

$$
\begin{aligned}
\hat{\lambda}_{ML}-\lambda^* &= -\frac{\dot{\mathcal{L}}_n(\{x\},\lambda^*)}{\ddot{\mathcal{L}}_n(\{x\},\lambda^*)} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}\partial_\lambda \log p_\lambda(\{x_i\}|\lambda)|_{\lambda^*}}{\frac{1}{n}\sum_{i=1}^{n}(\partial_\lambda^2 \log p_\lambda(\{x_i\}|\lambda))|_{\lambda^*}}
\end{aligned}
$$

When applying the Law of Large Numbers (LLN) we see that there is a convergence in probability to deterministic quantities:

- Nominator: $\mathbb{E}[\partial\lambda \log p_\lambda(\{x_i\}|\lambda)]|_{\lambda^*} \to 0$
- Denominator: $\mathbb{E}[\partial^2\lambda \log p_\lambda(\{x_i\}|\lambda)]|_{\lambda^*} \to I(\lambda^*)$

This shows, non surprisingly, that $\hat{\lambda}_{\mathrm{ML}} \to \lambda^*$ as $n \to \infty$.

To go beyond this results, we should "zoom in" and look at the fluctuations. We thus multiply both sides of the equation by $\sqrt{nI}$ and finds that at large n:

$$\sqrt{nI}(\hat{\lambda}_{ML} - \lambda^*) = \frac{\overbrace{\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\partial_\lambda \log p_\lambda(x_i|\lambda)}^{S_n}}{\sqrt{I(\lambda^*)}} = \frac{S_n}{\sqrt{I}}$$

Now, as $n \to \infty$, we can apply the Central Limit Theorem and observe a convergence in law, so that

$$\sqrt{nI}(\hat{\lambda}_{ML} - \lambda^*) \rightsquigarrow \mathcal{N}(0,1) \tag{2.2}$$

Equivalently, that means that, for large numbers, the maximum likelihood estimate is asymptotically Gaussian

$$\hat{\lambda}_{ML} \sim \mathcal{N}(\lambda^*, \frac{1}{nI})$$
$$Variance = \frac{1}{nI(\lambda^*)}$$

To summarize, we can now recap all the reason why is Maximum Likelihood (ML) good, at large n: It has the following properties:

- ML is consistent, $\hat{\lambda}_{ML}^{(*)} \to \lambda^*)$
- ML is efficient, as it achieve the Cramers Rao bound: $MSE(\hat{\lambda}_{ML}) \to \frac{1}{nI(\lambda^*)}$
- ML is asymptotically Gaussian, so that asymptotically, it is easy to estimate confidence intervals.

Multi-variate case — All these considerations can be extended beyond the scalar case to the multi-variate case when we want to estimate a vector d-dimensional vector $\boldsymbol{\lambda}$. In this case, the only difference is that the Fisher information becomes a $d \times d$ matrix:

Definition 2.7 (Fisher information matrix)

$$I_{ij} = \mathbb{E}\left[\frac{\partial \log p(x|\overrightarrow{\lambda})}{\partial\lambda_i} \cdot \frac{\partial \log p(x|\overrightarrow{\lambda})}{\partial\lambda_j}\right] = -\mathbb{E}\left[\frac{\partial}{\partial\lambda_i\partial\lambda_j}\log p(x|\overrightarrow{\lambda})\right]$$

In the multivariate case, the Cramers-Rao bound for unbiased estimator now gives a strict bound on the covariance of any estimator $\hat{\boldsymbol{\lambda}}$ using the matrix inverse Fisher Information Matrix :

$$\mathrm{Cov}[\hat{\boldsymbol{\lambda}}] \geq \frac{1}{n}I^{-1}$$