4.1 Maximum Likelihood principle

Given a parametric family \( p(\cdot; \theta) \) for \( \theta \in \Theta \), we define the likelihood function for some observation \( x \), denoted \( \mathcal{L}(\theta) \), as

\[
\mathcal{L}(\theta) \triangleq p(x; \theta)
\]

Depending on the nature of the corresponding random variable \( X \), \( p(\cdot; \theta) \) here is either the probability mass function (pmf) if \( X \) is discrete or the probability density function (pdf) if \( X \) is continuous. The likelihood is a function of the parameter \( \theta \), with the observation \( x \) fixed.

We want to find (estimate) the value of the parameter \( \theta \) that best explains the observation \( x \). This estimate is called the Maximum Likelihood Estimator (MLE), and is given by

\[
\hat{\theta}_{\text{ML}}(x) \triangleq \arg\max_{\theta \in \Theta} p(x; \theta)
\]

This means \( \hat{\theta}_{\text{ML}}(x) \) is the value of the parameter that maximizes the probability of observation \( p(x; \cdot) \) (as a function of \( \theta \)). Usually though, we are not only given a single observation \( x \), but iid samples \( x_1, x_2, \ldots, x_n \) of some distribution with pmf (or pdf) \( p(\cdot; \theta) \). In that case, the likelihood function is

\[
\mathcal{L}(\theta) = p(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} p(x_i; \theta)
\]

4.1.1 Example: Binomial model

Consider the family of Binomial distributions with parameters \( n \) and \( \theta \in [0, 1] \).

\[
X \sim \text{Bin}(n, \theta) \quad \text{with} \quad \Omega_X = \{0, 1, \ldots, n\}
\]

Given some observation \( x \in \Omega_X \) of the random variable \( X \), we want to estimate the parameter \( \theta \) that best explains this observation with the maximum likelihood principle. Recall that the pmf of a Binomial distribution is

\[
p(x; \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}
\]

Our goal is to maximize the likelihood function \( \mathcal{L}(\theta) = p(x; \theta) \), even though it is a highly non-linear function of \( \theta \). To make things easier, instead of maximizing the likelihood function \( \mathcal{L}(\theta) \) directly, we can maximize any strictly increasing function of \( \mathcal{L}(\theta) \).
Since \( \log \) is a strictly increasing function (i.e. \( 0 < a < b \iff \log a < \log b \)), one common choice is to maximize the \emph{log likelihood function} \( \ell(\theta) \triangleq \log p(x; \theta) \). This leads to the same value of the MLE
\[
\hat{\theta}_{\text{ML}}(x) = \arg\max_{\theta \in \Theta} p(x; \theta) = \arg\max_{\theta \in \Theta} \log p(x; \theta) \tag{4.5}
\]

Using the log likelihood function could be problematic when \( p(x; \theta) = 0 \) for some parameter \( \theta \). In that case, assigning \( \ell(\theta) = -\infty \) for this value of \( \theta \) has no effect on the maximization later on. Here, for the Binomial model, we have
\[
\ell(\theta) = \log p(x; \theta) = \log \left( \frac{n}{x} \right) + x \log \theta + (n - x) \log(1 - \theta) \tag{4.6}
\]

Now that we know the form of \( \ell(\theta) \), how do we maximize it? We can first search for \emph{stationary points} of the log likelihood, that is values of \( \theta \) such that
\[
\nabla_{\theta} \ell(\theta) = 0 \tag{4.7}
\]

Or, in 1D, \( \ell'(\theta) = 0 \). This is a necessary condition for \( \hat{\theta} \) to be a maximum (see Section 4.1.2).

The stationary points of the log likelihood are given by
\[
\frac{\partial \ell}{\partial \theta} = \frac{x}{\theta} - \frac{n - x}{1 - \theta} = 0 \quad \Rightarrow \quad x - \theta x - (n - x)\theta = 0 \quad \Rightarrow \quad \theta^* = \frac{x}{n} \tag{4.8}
\]

The log likelihood function of the Binomial model is also strictly concave (i.e. \( \ell''(\theta) < 0 \)), thus \( \theta^* \) being a stationary point of \( \ell(\theta) \) is also a sufficient condition for it to be a global maximum (see Section 4.1.2).

The MLE of the Binomial model is the relative frequency of the observation \( x \), which follows our intuition. Furthermore, even though it is not a general property of the MLE, this estimator is unbiased
\[
X \sim \text{Bin}(n, \theta) \quad \Rightarrow \quad \mathbb{E}_X[\hat{\theta}_{\text{ML}}] = \mathbb{E}_X\left[\frac{X}{n}\right] = \frac{n\theta}{n} = \theta \tag{4.10}
\]

Note that we maximized \( \ell(\theta) \) without specifying any constraint on \( \theta \), even though it is required that \( \theta \in [0, 1] \). However, here this extra condition has little effect on the optimization since the stationary point (4.8) is already in the interior of the parameter space \( \Theta = [0, 1] \) if \( x \neq 0 \) or \( n \). In two latter cases, we can exploit the monotonicity of \( \ell \) on \( \Theta \) to conclude that the maxima are on the boundaries of \( \Theta \) (resp. 0 and 1).
4.1.2 Comments on optimization

- In general, being a stationary point (i.e. $f'(\theta) = 0$ in 1D) is a necessary condition for $\theta$ to be a local maximum when $\theta$ is in the interior of the parameter space $\Theta$. However, it is not sufficient. A stationary point can be either a local maximum or a local minimum in 1D (or a saddle point in the multivariate case). We also need to check the second derivative $f''(\theta) < 0$ for it to be a local maximum.

![Diagram showing local maximum with $f''(\theta) < 0$]

- The previous point only gives us a local result. To guarantee that $\theta^* \in \Theta$ is a global maximum, we need to know global properties about the function $f$. For example, if $\forall \theta \in \Theta, f''(\theta) \leq 0$ (i.e. the function $f$ is concave, the negative of a convex function), then $f'(\theta^*) = 0$ is a sufficient condition for $\theta^*$ to be a global maximum.

- We need to be careful though with cases where the maximum is on the boundary of the parameter space $\Theta$ ($\theta^* \in \text{boundary}(\Theta)$). In that case, $\theta^*$ may not necessarily be a stationary point, meaning that $\nabla_\theta f(\theta^*)$ may be non-zero.

- Similarly for the multivariate case, $\nabla f(\theta^*) = 0$ is in general a necessary condition for $\theta^*$ to be a local maximum if it belongs to the interior of $\Theta$. To make sure it is a local maximum, we need to check if the Hessian matrix of $f$ is negative definite at $\theta^*$ (this is the multivariate equivalent of $f''(\theta) < 0$ in 1D)

\[
\text{Hessian}(f)(\theta^*) < 0 \quad \text{where} \quad \text{Hessian}(f)(\theta^*)_{i,j} = \frac{\partial^2 f(\theta^*)}{\partial \theta_i \partial \theta_j} \quad (4.11)
\]

We also get similar results in the multivariate case if we know global properties on the function $f$. For example, if the function $f$ is concave, then $\nabla f(\theta^*) = 0$ is also a sufficient condition for $\theta^*$ to be a global maximum. To verify that a multivariate function is concave, we have to check if the Hessian matrix is negative semi-definite on the whole parameter space $\Theta$ (the multivariate equivalent of $\forall \theta \in \Theta, f''(\theta) \leq 0$ in 1D).

\[
\forall \theta \in \Theta, \text{Hessian}(f)(\theta) \preceq 0 \quad \iff \quad \text{f is concave} \quad (4.12)
\]
4.1.3 Properties of the MLE

- The MLE does not always exist. For example, if the estimate is on the boundary of the parameter space \( \hat{\theta}_{ML} \in \text{boundary}(\Theta) \) but \( \Theta \) is an open set.
- The MLE is not necessarily unique; the likelihood function could have multiple maxima.
- The MLE is not admissible in general; for some estimation problems, there exists some other estimates dominating the MLE for the frequentist risk (see James-Stein estimator, Lecture 5).

4.1.4 Example: Multinomial model

Suppose that \( X_i \) is a discrete random variable over \( K \) choices. We could choose the domain of this random variable as \( \Omega_{X_i} = \{1, 2, \ldots, K\} \). Instead, it is convenient to encode \( X_i \) as a random vector, taking values in the unit bases in \( \mathbb{R}^K \). This encoding is called the one-hot encoding, and is widely used in the neural networks literature.

\[
\Omega_{X_i} = \{e_1, e_2, \ldots, e_K\}, \quad \text{where } e_j = \begin{pmatrix} 0 \ldots 1 \ldots 0 \end{pmatrix}^T \in \mathbb{R}^K \text{ (jth coordinate)}
\]

To get the pmf of this discrete random vector, we can define a family of probability distributions with parameter \( \pi \in \Delta_K \). The parameter space \( \Theta = \Delta_K \) is called the probability simplex on \( K \) choices, and is given by

\[
\Delta_K \triangleq \left\{ \pi \in \mathbb{R}^K; \forall j \pi_j \geq 0 \text{ and } \sum_{j=1}^{K} \pi_j = 1 \right\} \quad (4.13)
\]

The probability simplex is a \((K - 1)\)-dimensional object in \( \mathbb{R}^K \) because of the constraint \( \sum_{j=1}^{K} \pi_j = 1 \). For example, here \( \Delta_3 \) is a 2-dimensional set. This makes optimization over the parameter space more difficult.

The distribution of the random vector \( X_i \) is called a Multinoulli distribution with parameter \( \pi \), and is denoted \( X_i \sim \text{Mult}(\pi) \). Its pmf is

\[
p(x_i; \pi) = \prod_{j=1}^{K} \pi_{x_i,j}^{x_{i,j}} \quad \text{where } x_{i,j} \in \{0, 1\} \text{ is the } j^{th} \text{ component of } x_i \in \Omega_{X_i} \quad (4.14)
\]

The Multinoulli distribution can be seen as the equivalent of the Bernoulli distribution over \( K \) choices (instead of 2). If we consider \( n \) iid Multinoulli random vectors \( X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Mult}(\pi) \), then we can define the random vector \( X \) as

\[
X = \sum_{i=1}^{n} X_i \sim \text{Mult}(n, \pi) \quad \text{with} \quad \Omega_X = \left\{ (n_1, n_2, \ldots, n_K); \forall j n_j \in \mathbb{N} \text{ and } \sum_{j=1}^{K} n_j = n \right\}
\]

The distribution of \( X \) is called a Multinomial distribution with parameters \( n \) and \( \pi \), and is the analogue of the Binomial distribution over \( K \) choices (similar to Multinoulli/Bernoulli). Given
some observation \( x \in \Omega_X \), we want to estimate the parameter \( \pi \) that best explains this observation with the maximum likelihood principle. The likelihood function is

\[
L(\pi) = p(x; \pi) = \frac{1}{Z} \prod_{i=1}^{n} p(x_i; \pi)
\]

\[
= \frac{1}{Z} \prod_{j=1}^{K} \left[ \prod_{i=1}^{n} \pi_{x_i,j} \right] = \frac{1}{Z} \prod_{j=1}^{K} \left[ \prod_{i=1}^{n} \pi_{x_i,j} \right] = \frac{1}{Z} \prod_{j=1}^{K} \sum_{i=1}^{n} x_{i,j}
\]

Where \( n_j = \sum_{i=1}^{n} x_{i,j} \) is the number of times we observe the value \( j \) (or \( e_j \in \Omega_X \)). Note that \( n_j \) remains a function of the observation (thus we could write \( n_j(x) \)), although this explicit dependence on \( x \) is omitted here. Equivalently, we could have looked for the MLE of a Multinoulli model (with parameter \( \pi \)) with \( n \) observations \( x_1, x_2, \ldots, x_n \) instead of the MLE of a Multinomial model with a single observation \( x \); the only effect here would be the lack of normalization constant \( Z \) in the likelihood function. Like in Section 4.1.1, we take the log likelihood function to make the optimization simpler

\[
\ell(\pi) = \log p(x; \pi) = \sum_{j=1}^{K} n_j \log \pi_j - \log Z
\]

We want to maximize \( \ell(\pi) \) such that \( \pi \) still is a valid element of \( \Delta_K \). Given the constraints (4.13) induced by the probability simplex \( \Delta_K \), this involves solving the following constrained optimization problem

\[
\max_{\pi} \ell(\pi) \quad \text{subject to} \quad \pi \in \Delta_K \quad \Leftrightarrow \quad \max_{\pi} \quad \frac{1}{Z} \sum_{j=1}^{K} n_j \log \pi_j \quad \text{s.t.} \quad \forall j, \pi_j \geq 0 \quad \sum_{j=1}^{K} \pi_j = 1
\]

To solve this optimization problem, we have 2 options:

- We could reparametrize (4.17) with \( \pi_1, \pi_2, \ldots, \pi_{K-1} \geq 0 \) with the constraint \( \sum_{j=1}^{K-1} \pi_j \leq 1 \) and set \( \pi_K = 1 - \sum_{j=1}^{K-1} \pi_j \). The log likelihood function to maximize would become

\[
\ell(\pi_1, \pi_2, \ldots, \pi_{K-1}) = \sum_{j=1}^{K-1} n_j \log \pi_j + n_K \log (1 - \pi_1 - \pi_2 - \ldots - \pi_{K-1})
\]

The advantage here would be that the parameter space would be a full dimensional object \( \Delta_{K-1}^c \subset \mathbb{R}^{K-1} \), sometimes called the corner of the cube, which is a more suitable setup for optimization (in particular, we could apply the techniques from Section 4.1.2)
• We choose to use the Lagrange multipliers approach. The Lagrange multipliers method can be used to solve constrained optimization problems with equality constraints (and, more generally, with inequality constraints as well) of the form

\[
\begin{align*}
\text{max} & \quad f(\pi) \\
\text{s.t.} & \quad g(\pi) = 0
\end{align*}
\]

Here, we can apply it to the optimization problem (4.17); i.e. the maximization of \( \ell(\pi) \), under the equality constraint

\[
\sum_{j=1}^{K} \pi_j = 1 \iff 1 - \sum_{j=1}^{K} \pi_j = 0
\]

(4.20)

The fundamental part of the Lagrange multipliers method is an auxiliary function \( \mathcal{J}(\pi, \lambda) \) called the Lagrangian function. This is a combination of the function to maximize (here \( \ell(\pi) \)) and the equality constraint function \( g(\pi) \).

\[
\mathcal{J}(\pi, \lambda) = \sum_{j=1}^{K} n_j \log \pi_j + \lambda \left( 1 - \sum_{j=1}^{K} \pi_j \right)
\]

(4.21)

Where \( \lambda \) is called a Lagrange multiplier. We dropped the constant \( Z \) since it has no effect on the optimization. We can search the stationary points of the Lagrangian, i.e pairs \( (\pi, \lambda) \) satisfying \( \nabla_\pi \mathcal{J}(\pi, \lambda) = 0 \) and \( \nabla_\lambda \mathcal{J}(\pi, \lambda) = 0 \). Note that the second equality is equivalent to the equality constraint in our optimization problem \( g(\pi) = 0 \). The first equality leads to

\[
\frac{\partial \mathcal{J}}{\partial \pi_j} = \frac{n_j}{\pi_j} - \lambda = 0 \quad \Rightarrow \quad \pi_j^* = \frac{n_j}{\lambda}
\]

(4.22)

Here, the Lagrange multiplier \( \lambda \) acts as a scaling constant. As \( \pi^* \) is required to satisfy the constraint \( g(\pi^*) = 0 \), we can evaluate this scaling factor

\[
\sum_{j=1}^{K} \pi_j^* = 1 \quad \Rightarrow \quad \lambda = \sum_{j=1}^{K} n_j = n
\]
Once again, in order to check that \( \pi^* \) is indeed a local maximum, we would also have to verify that the Hessian of the log likelihood at \( \pi^* \) is negative definite. However here, \( \ell \) is a concave function \((\forall \pi, \text{Hessian}(\ell)(\pi) \preceq 0)\). This means, according to Section 4.1.2, that \( \pi^* \) being a stationary point is a sufficient condition for it to be a global maximum.

\[
\hat{\pi}^{(j)}_{\text{ML}} = \frac{n_j}{n} \tag{4.23}
\]

The MLE of the Multinomial model, similar to the Binomial model from Section 4.1.1, is the relative frequency of the observation vector \( x = (n_1, n_2, \ldots, n_K) \), and again follows our intuition. Note that \( \pi^*_j \geq 0 \), which was also one of the constraints of \( \Delta_K \).

### 4.1.5 Geometric interpretation of the Lagrange multipliers method

The Lagrange multipliers method is applied to solve constrained optimization problems of the form

\[
\begin{aligned}
\max_{\pi} \quad & f(\pi) \\
\text{s.t.} \quad & g(\pi) = 0
\end{aligned} \tag{4.24}
\]

With this generic formulation, the Lagrangian is \( J(x, \lambda) = f(x) + \lambda g(x) \), with \( \lambda \) the Lagrange multiplier. In order to find an optimum of (4.24), we can search for the stationary points of the Lagrangian, i.e. pairs \( (x, \lambda) \) such that \( \nabla_x J(x, \lambda) = 0 \) and \( \nabla_\lambda J(x, \lambda) = 0 \). The latter equality is always equivalent to the constraint \( g(x) = 0 \), whereas the former can be rewritten as

\[
\nabla_x J(x, \lambda) = 0 \quad \Rightarrow \quad \nabla f(x) = -\lambda \nabla g(x) \tag{4.25}
\]

At a stationary point, the Lagrange multiplier \( \lambda \) is a scaling factor between the gradient vectors \( \nabla f(x) \) and \( \nabla g(x) \). Geometrically, this means that these two vectors are parallel.