IFT 6269: Probabilistic Graphical Models **Fall 2017**

Lecture 4 — September 15

Lecturer: Simon Lacoste-Julien *Scribe: Philippe Brouillard & Tristan Deleu*

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

$$\mathcal{L}(\theta) \triangleq p(x;\theta) \tag{4.1}$$

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 θ, with the observation x fixed.

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

$$\hat{\theta}_{\mathrm{ML}}(x) \triangleq \underset{\theta\in\Theta}{\operatorname{argmax}}\, p(x;\theta) \tag{4.2}$$

This means $\hat{\theta}_{\mathrm{ML}}(x)$ is the value of the parameter that maximizes the probability of observation $p(x;\cdot)$ (as a function of θ). 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

$$\mathcal{L}(\theta) = p(x_1, x_2, \ldots, x_n;\theta) = \prod_{i=1}^{n} p(x_i;\theta) \tag{4.3}$$

4.1.1 Example: Binomial model

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

$$X \sim \mathrm{Bin}(n,\theta) \qquad \text{with} \qquad \Omega_X = \{0, 1, \ldots, n\}$$

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

$$p(x;\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \tag{4.4}$$

Our goal is to maximize the likelihood function $\mathcal{L}(\theta) = p(x;\theta)$, even though it is a highly non-linear function of θ. 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 (ie. $0 < a < b \Leftrightarrow \log a < \log b$), one common choice is to maximize the *log likelihood function* $\ell(\theta) \triangleq \log p(x;\theta)$. This leads to the same value of the MLE

$$\hat{\theta}_{\mathrm{ML}}(x) = \operatorname*{argmax}_{\theta\in\Theta} p(x;\theta) = \operatorname*{argmax}_{\theta\in\Theta} \log p(x;\theta) \tag{4.5}$$

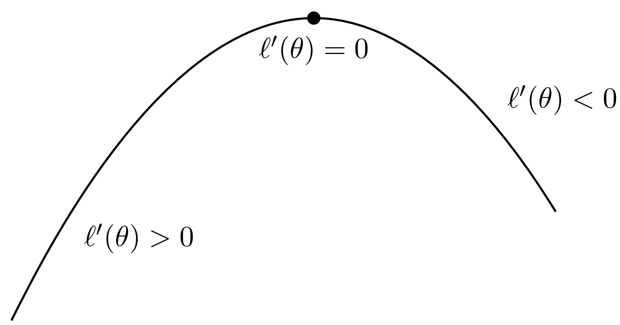
Using the log likelihood function could be problematic when $p(x;\theta) = 0$ for some parameter θ. In that case, assigning $\ell(\theta) = -\infty$ for this value of θ has no effect on the maximization later on. Here, for the Binomial model, we have

$$\begin{aligned} \ell(\theta) &= \log p(x;\theta) \\ &= \underbrace{\log \binom{n}{x}}_{\text{constant in } \theta} + x \log\theta + (n-x)\log(1-\theta) \end{aligned} \tag{4.6}$$

Now that we know the form of $\ell(\theta)$, how do we maximize it? We can first search for *stationary points* of the log likelihood, that is values of θ such that

$$\nabla_\theta\, \ell(\theta) = 0 \tag{4.7}$$

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



The stationary points of the log likelihood are given by

$$\frac{\partial \ell}{\partial \theta} = \frac{x}{\theta} - \frac{n-x}{1-\theta} = 0 \qquad \Rightarrow \qquad x - \theta x - (n-x)\theta = 0 \qquad \Rightarrow \qquad \theta^\star = \frac{x}{n} \tag{4.8}$$

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

$$\boxed{\hat{\theta}_{\mathrm{ML}} = \frac{x}{n}} \tag{4.9}$$

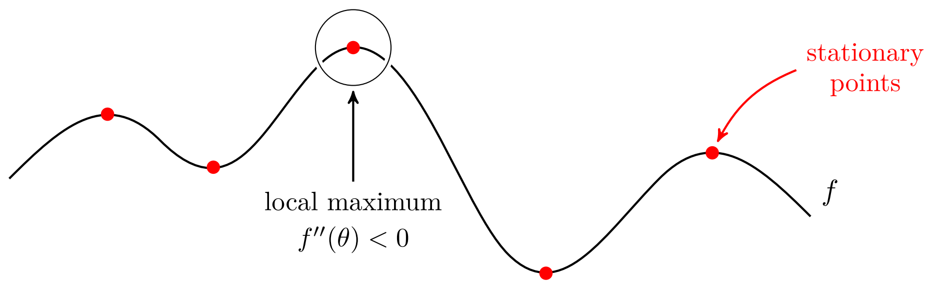
The MLE of the Binomial model is the relative frequency of the observation x, which follows the intuition. Furthermore, even though it is not a general property of the MLE, this estimator is unbiased

$$X \sim \mathrm{Bin}(n,\theta) \qquad \Rightarrow \qquad \mathbb{E}_X\Big[\hat{\theta}_{\mathrm{ML}}\Big] = \mathbb{E}_X\Big[\frac{X}{n}\Big] = \frac{n\theta}{n} = \theta \tag{4.10}$$

Note that we maximized $\ell(\theta)$ without specifying any constraint on θ, 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 ℓ on Θ to conclude that the maxima are on the boundaries of Θ (resp. 0 and 1).

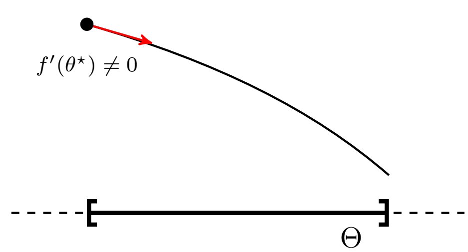
4.1.2 Comments on optimization

- In general, being a stationary point (ie. $f'(\theta) = 0$ in 1D) is a necessary condition for θ to be a *local maximum* when θ is in the interior of the parameter space Θ. 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.



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

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



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

$$\text{Hessian}(f)(\theta^\star) \prec 0 \qquad \text{where} \qquad \text{Hessian}(f)(\theta^\star)_{i,j} = \frac{\partial f(\theta^\star)}{\partial \theta_i \partial \theta_j} \tag{4.11}$$

  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 $\nabla f(\theta^\star) = 0$ is also a sufficient condition for $\theta^\star$ 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 Θ (the multivariate equivalent of $\forall \theta \in \Theta$, $f''(\theta) \leq 0$ in 1D).

$$\forall \theta \in \Theta,\ \text{Hessian}(f)(\theta) \preceq 0 \qquad \Leftrightarrow \qquad f \text{ is concave} \tag{4.12}$$

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 $\hat{\theta}_{\text{ML}} \in \text{boundary}(\Theta)$ but Θ is an open set.
- The MLE is not necessarily unique; the likelihood function could have multiple maxima.
- The MLE is not admissible in general

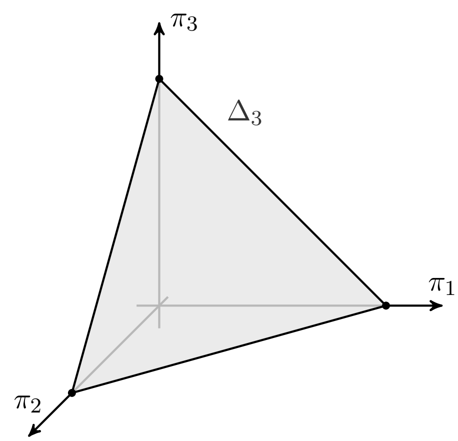
4.1.4 Example: Multinomial model

Suppose that X_i is a discrete random variable over K choices. We could choose the domain of this random variable as $\Omega_{X_i} = \{1, 2, \dots, K\}$. 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.

$$\Omega_{X_i} = \{e_1, e_2, \dots, e_K\} \qquad \text{where } e_j = \begin{pmatrix} 0 & \dots & \underset{\substack{\uparrow \\ j^{\text{th}} \text{ coordinate}}}{1} & \dots & 0 \end{pmatrix}^T \in \mathbb{R}^K$$

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

$$\Delta_K \triangleq \left\{ \pi \in \mathbb{R}^K \ ; \ \forall j \ \pi_j \geq 0 \text{ and } \sum_{j=1}^{K} \pi_j = 1 \right\} \tag{4.13}$$



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 Δ_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 π, and is denoted $X_i \sim \text{Mult}(\pi)$. Its pmf is

$$p(x_i; \pi) = \prod_{j=1}^{K} \pi_j^{x_{i,j}} \qquad \text{where } x_{i,j} \in \{0, 1\} \text{ is the } j^{\text{th}} \text{ component of } x_i \in \Omega_{X_i} \tag{4.14}$$

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 $X_1, X_2, \dots, X_n \overset{\text{iid}}{\sim} \text{Mult}(\pi)$, then we can define the random vector X as

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

The distribution of X is called a *Multinomial distribution* with parameters n and π, 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 π that best explains this observation with the maximum likelihood principle. The likelihood function is

$$\begin{aligned}
\mathcal{L}(\pi) = p(x;\pi) &= \frac{1}{Z}\prod_{i=1}^{n} p(x_i;\pi) \\
&= \frac{1}{Z}\prod_{i=1}^{n}\left[\prod_{j=1}^{K}\pi_j^{x_{i,j}}\right] = \frac{1}{Z}\prod_{j=1}^{K}\left[\prod_{i=1}^{n}\pi_j^{x_{i,j}}\right] \qquad \text{Where } Z \text{ is a normalization constant } \frac{1}{Z} = \binom{n}{n_1, n_2, \ldots, n_K} = \frac{n!}{n_1!\cdot n_2!\ldots n_K!} \\
&= \frac{1}{Z}\prod_{j=1}^{K}\pi_j^{\sum_{i=1}^{n} x_{i,j}} \qquad (4.15)
\end{aligned}$$

Where $n_j = \sum_{i=1}^{n} x_{i,j}$ is the number of times we observe the value j (or $e_j \in \Omega_{X_i}$). Note that n_j remains a function of the observation $n_j(x)$, although this explicit dependence on x is omitted here. Equivalently, we could have looked for the MLE of a Multinoulli model (with parameter π) 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

$$\ell(\pi) = \log p(x;\pi) = \sum_{j=1}^{n} n_j \log \pi_j - \underbrace{\log Z}_{\text{constant in } \pi} \qquad (4.16)$$

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

$$\left\{\begin{array}{rl} \max\limits_{\pi} & \ell(\pi) \\ \text{subject to} & \pi \in \Delta_K \end{array}\right. \quad \Leftrightarrow \quad \left\{\begin{array}{rl} \max\limits_{\pi} & \sum\limits_{j=1}^{K} n_j \log \pi_j \\ \text{s.t.} & \pi_j \geq 0 \\ & \sum\limits_{j=1}^{K} \pi_j = 1 \end{array}\right. \qquad (4.17)$$

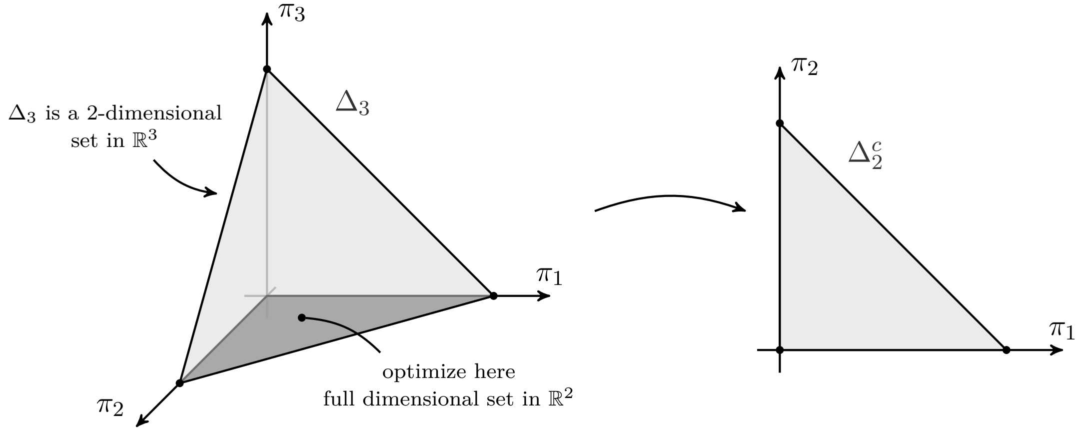
To solve this optimization problem, we have 2 options:

- 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

$$\ell(\pi_1, \pi_2, \ldots, \pi_{K-1}) = \sum_{j=1}^{K-1} n_j \log \pi_j + n_K \log\left(1 - \pi_1 - \pi_2 - \ldots - \pi_{K-1}\right) \qquad (4.18)$$

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)

$$\Delta_{K-1}^c = \left\{(\pi_1, \pi_2, \ldots, \pi_{K-1}) \in \mathbb{R}^{K-1} \ ; \ \forall j \ \pi_j \geq 0 \text{ and } \sum_{j=1}^{K-1}\pi_j \leq 1\right\} \qquad (4.19)$$



- 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

$$\begin{cases} \max_{\pi} & f(\pi) \\ \text{s.t.} & g(\pi) = 0 \end{cases}$$

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

$$\sum_{j=1}^{K} \pi_j = 1 \qquad \Leftrightarrow \qquad \underbrace{1 - \sum_{j=1}^{K} \pi_j}_{=g(\pi)} = 0 \tag{4.20}$$

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 $g(\pi)$.

$$\mathcal{J}(\pi, \lambda) = \sum_{j=1}^{K} n_j \log \pi_j + \lambda \left(1 - \sum_{j=1}^{K} \pi_j \right) \tag{4.21}$$

Where λ 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 (π, λ) 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

$$\frac{\partial \mathcal{J}}{\partial \pi_j} = \frac{n_j}{\pi_j} - \lambda = 0 \qquad \Rightarrow \qquad \pi_j^\star = \frac{n_j}{\lambda} \tag{4.22}$$

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

$$\sum_{j=1}^{K} \pi_j^\star = 1 \qquad \Rightarrow \qquad \lambda = \sum_{j=1}^{K} n_j = n$$

Once again, in order to check that $\pi^\star$ is indeed a local maximum, we would also have to verify that the Hessian of the log likelihood at $\pi^\star$ is negative definite. However here, ℓ is a concave function ($\forall \pi,\ \text{Hessian}(\ell)(\pi) \preceq 0$). This means, according to Section 4.1.2, that $\pi^\star$ being a stationary point is a sufficient condition for it to be a global maximum.

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

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

$$\begin{cases} \max\limits_{\pi} & f(\pi) \\ \text{s.t.} & g(\pi) = 0 \end{cases} \tag{4.24}$$

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

$$\nabla_x \mathcal{J}(x, \lambda) = 0 \qquad \Rightarrow \qquad \nabla f(x) = -\lambda \nabla g(x) \tag{4.25}$$

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