M2 MVA - Mathématiques, Vision, Apprentissage

PROBABILISTIC GRAPHICAL MODELS

Lecturers:

→ Pierre Latouche — Université Paris Descartes, CNRS
pierre.latouche@math.cnrs.fr
→ Nicolas Chopin — École Nationale de la Statistique et de l'Administration Économique
nicolas.chopin@ensae.fr

Lecture notes:

→ Antoine Barrier — École Normale Supérieure de Lyon
antoine.barrier@ens-lyon.fr
https://perso.ens-lyon.fr/antoine.barrier/fr/

2019/2020 (first semester) – Last update: October 10, 2020

In these lecture notes, I develop the students's notes of Francis Bach's Probabilistic Graphical Models course. I add some points of view developed by Pierre Latouche and Nicolas Chopin who are now giving the course.

This document is work in progress.

The scribes of Francis Bach's courses are: Jean-Baptiste Alayrac, Mathieu Andreux, Ismael Belghiti, Thomas Belhalfaoui, Robin Benesse, Lilian Besson, Lénaic Chizat, Basile Clément, Lê- Huu D. Khuê, Marie d'Autume, Nathan de Lara, Gauthier Gidel, Pauline Luc, Maryan Morel, Lucas Plaetevoet, Aymeric Reshef, Jaime Roquero, Khalife Sammy, Claire Vernade, JieYing Wu.

Summary

Introduction: complex data modelization

Problem To model complex data, one is confronted with two main questions:

- how to manage the complexity of the data to be processed?
- how to infer global properties from local models?

These questions lead to 3 types of problems:

- the representation of data: how to obtain a global model from a local model?
- the inference of the distributions: how to use the model?
- the learning of the model: what are the parameters of the model?

For instance we present some models associated to classical problems.

Image Consider a $n \times n$ (pixels) monochromatic image. If each pixel is modelled by a discrete random variable (so there are n^2 of them), then the image can be modelled using a grid.

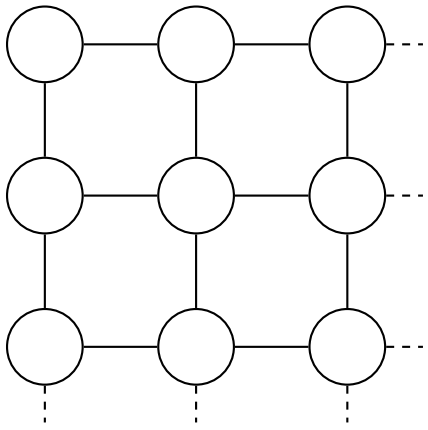


Figure 1: Grid modelling the image

Bioinformatics Consider a long sequence of n DNA bases. If each base of this sequence is modelled by a discrete random variable (that, in general, can take values in $\{A, C, G, T\}$), then the sequence can be modelled by a MARKOV chain:

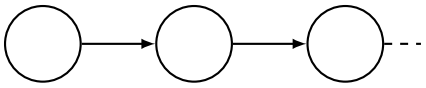


Figure 2: Graph of a MARKOV chain

Finance Consider the evolution of stock prices in discrete time, where we have values at time t. It is reasonable to postulate that the change of price of a stock at time $t+1$ only depends on its price or the price of all stocks at time t.

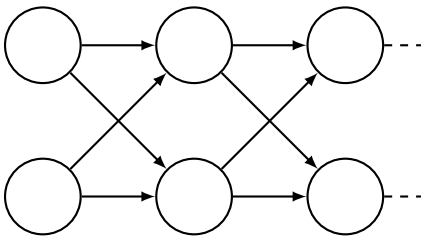


Figure 3: Possible graph for 2 stocks

Speech processing Consider the syllables of a word and the way they are interpreted by a human ear or by a computer. Each syllable can be represented by a random sound. The objective is then to retrieve the word from the sequence of sounds heard or recorded. In this case, we can use a hidden MARKOV model:

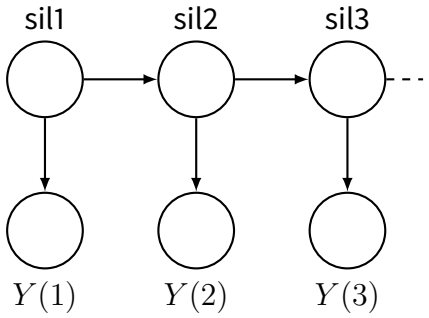


Figure 4: Graph for speech processing

Text Consider a text with 1000000 words. The text is modelled by a vector such that each of its components equals to the number of times each keyword appears. This is usually called the "bag of words" model. This model seems to be weak, as it does not take the order of the words into account. However, it works quite well in practice. A so-called *naive* BAYES *classifier* can be used for classification (for instance spam vs non spam).

It is clear that models which ignore the dependence among variables are too simple for real-world problems. On the other hand, models in which every random variable is dependent all or too many other ones are doomed both for statistical (lack of data) and computational reasons. Therefore, in practice, one has to make suitable assumptions to design models with the right level of complexity,

so that the models obtained are able to *generalize* well from a statistical point of view and lead to tractable computations from an algorithmic perspective.

[todo]

General issues in this class [todo]

1. Representation $\rightarrow$ DGM, UGM / parameterization $\rightarrow$ exponential family
2. Inference (computing $p(x_A \mid x_B)$) $\rightarrow$ sum-product algorithm
3. Statistical estimation $\rightarrow$ maximum likelihood, maximum entropy

CHAPTER 1

Maximum likelihood estimation

In Annex I. you can find a review of basic probabilities.

I. Statistical models

Definition I. .1. [Statistical model]
A (parametric) statistical model $\mathcal{P}_\Theta$ is a collection of probability distributions (or a collection of probability density functions[a]) defined on the same space and parameterized by parameters θ belonging to a set $\Theta \subset \mathbb{R}^p$. Formally:

$$\mathcal{P}_\Theta = \{p_\theta \,|\, \theta \in \Theta\}$$

[a] in which case they are all defined with respect to the same base measure, such as the Lebesgue measure in $\mathbb{R}^d$

Bernoulli model Consider a binary random variable X that can take the value 0 or 1. If $p(X=1)$ is parametrized by $\theta \in [0,1]$, we have:

$$p(X=1) = \theta \qquad \text{and} \qquad p(X=0) = 1-\theta$$

that we can summarized by $p(X=x) = \theta^x(1-\theta)^{1-x}$ and we write $X \sim \mathcal{B}(\theta)$.

The Bernoulli model is the collection of these distributions for $\theta \in \Theta = [0,1]$:

$$\mathcal{P}_{\text{Bernoulli}} = \{\mathcal{B}(\theta) \,|\, \theta \in [0,1]\}$$

Binomial model A binomial random variable $\mathcal{B}(\theta, n)$ is defined as the value of the sum of n i.i.d. Bernoulli random variables with parameter $\theta \in \Theta = [0,1]$. The distribution of a random variable $N \sim \mathcal{B}(\theta, n)$ is given by

$$\forall k \in [\![0, N]\!], \quad p(N=k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

The binomial model is then

$$\mathcal{P}_{\text{binomial}} = \{\mathcal{B}(\theta, n) \,|\, \theta \in [0,1]\}$$

Note that in many cases n is known and thus only θ is the parameter, but sometimes we can have both θ and n as parameters.

Multinomial model Consider a discrete random variable C that can take values in $[\![1, K]\!]$. The random variable C can be represented by a K-dimensional random variable $X = (\mathbb{1}_{C=1}, \mathbb{1}_{C=2}, \dots, \mathbb{1}_{C=K})$ and we have the following event correspondance: $\{C = k\} = \{X_k = 1\}$.

If we parametrize $p(C = k)$ by a parameter $\pi_k \in [0, 1]$, then by definition we also have

$$\forall k \in [\![1, K]\!], \quad p(X_k = 1) = \pi_k$$

and we know that $\sum_{k=1}^{K} \pi_k = 1$. The probability distribution can be written as:

$$\forall x \in \mathcal{X}_K, \quad p(x) = \prod_{k=1}^{K} \pi_k^{x_k} \qquad \text{where } \mathcal{X}_K = \left\{ x \in \{0,1\}^K \mid \sum_{k=1}^{K} x_k = 1 \right\}$$

We will denote $\mathcal{M}(1, \pi_1, \dots, \pi_K)$ such a discrete distribution[1]. The multinomial model is:

$$\mathcal{P}_{\text{multinomial}} = \left\{ \mathcal{M}(1, \pi) \,|\, \pi \in \mathbb{R}_+^K, \sum_{k=1}^{K} \pi_k = 1 \right\}$$

Consider now $C_1, \dots, C_n$ i.i.d. random variables of distribution $\mathcal{M}(1, \pi)$, and denote by N_k the number of variables equal to k, then the joint distribution of $(N_1, N_2, \dots, N_K)$ is called a multinomial distribution of parameters n and π, denoted by $\mathcal{M}(n, \pi)$. With the second representation, we have that $\mathcal{M}(n, \pi)$ is the law of $\sum_{i=1}^{n} X_i$ where $(X_i)_{1 \le i \le n} \overset{\text{i.i.d.}}{\sim} \mathcal{M}(1, \pi)$. It takes the form:

$$p(n_1, n_2, \dots, n_K) = \frac{n!}{\prod_{k=1}^{K} n_k!} \prod_{k=1}^{K} \pi_k^{n_k}$$

The multinomial $\mathcal{M}(n, \pi)$ is to the $\mathcal{M}(1, \pi)$ distribution as the binomial distribution is to the BERNOULLI distribution. In the rest of this course, when we will talk about multinomial distributions, we will always refer to a $\mathcal{M}(1, \pi)$ distribution.

Gaussian models The Gaussian distribution is also known as the normal distribution. $\mathcal{N}(\mu, \sigma^2)$ the normal distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ can be written in the form

$$\forall x \in \mathbb{R}, \quad p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(x-\mu)^2}{2\sigma^2} \Big)$$

The model is then:

$$\mathcal{P}_{\text{Gaussian}} = \Big\{ \mathcal{N}(\mu, \sigma^2) \,|\, \mu \in \mathbb{R}, \sigma > 0 \Big\}$$

The multivariate Gaussian distribution of a d-dimensional vector with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathcal{M}_{d\,d}(\mathbb{R})$ symmetric positive definite matrix takes the form

$$\forall x \in \mathbb{R}^d, \quad p(x) = \frac{1}{(2\pi)^{d/2}} \frac{1}{\sqrt{\det(\Sigma)}} \exp\Big(- \frac{1}{2} (x-\mu)^\top \Sigma^{-1} (x-\mu) \Big)$$

and is denoted by $\mathcal{N}(\mu, \Sigma)$. Tt is a well-known property that μ is equal to the expectation of the law and that Σ is the covariance matrix of the law. The model is then:

$$\mathcal{P}_{\text{multivariate Gaussian}} = \Big\{ \mathcal{N}(\mu, \Sigma^2) \,|\, \mu \in \mathbb{R}^d, \Sigma \in \mathcal{S}_d(\mathbb{R}), \Sigma \succ 0 \Big\}$$

[1] note that this stands for C and X: we choose the representation we want, depending on the context

II. Parameter estimation by maximum likelihood

II. A. Definition

Maximum likelihood estimation is a method of estimating the parameters of a statistical model.

Let $\mathcal{P}_\Theta$ be a statistical model and $x_1, \dots, x_n$ be i.i.d. observations from a distribution p_{θ^*} for a fixed unknown $\theta^* \in \Theta$. As the name suggests, the maximum likelihood estimator is the parameter $\hat{\theta}_{\text{MLE}}$ under which the data are most likely.

Definition II. .1. [likelihood]
The likelihood of $x_1, \dots, x_n$ is defined as the function:

$$\begin{array}{rccc} \mathcal{L}: & \Theta & \longrightarrow & [0,1] \\ & \theta & \longmapsto & p_\theta(x_1, \dots, x_n) = \prod_{i=1}^n p_\theta(x_i) \end{array}$$

We can also consider[a] the $\log$-likelihood:

$$\ell : \theta \in \Theta \longmapsto \log \mathcal{L}(\theta) = \sum_{i=1}^n \log p_\theta(x_i)$$

[a] in practice it is often more convenient to work with the $\log$-likelihood function

Definition II. .2. [maximum likelihood estimator]
The maximum likelihood estimator of θ^* is defined as:

$$\hat{\theta}_{\text{MLE}} = \text{argmax}_{\theta \in \Theta}\, \mathcal{L}(\theta) = \text{argmax}_{\theta \in \Theta}\, \ell(\theta)$$

Next, we will apply this method for the models previously presented.

II. B. MLE for the Bernoulli model

Consider $x_1, x_2, \dots, x_n$ i.i.d. observations of $\mathcal{B}(\theta)$. We have

$$\ell(\theta) = \sum_{i=1}^n \log p_\theta(x_i) = \sum_{i=1}^n \log \theta^{x_i}(1-\theta)^{1-x_i} = n_1 \log(\theta) + (n - n_1)\log(1-\theta)$$

where $n_1 = \sum_{i=1}^n x_i = \sum_{i=1}^n \mathbb{1}_{x_i=1}$ is the number of success in our sample.
As $\ell(\theta)$ is strictly concave, it has a unique maximizer, and since the function is in addition differentiable, its maximizer $\hat{\theta}_{\text{MLE}}$ is the zero of its gradient. One can compute:

$$\nabla \ell(\theta) = \frac{\partial}{\partial \theta}\ell(\theta) = \frac{n_1}{\theta} - \frac{n - n_1}{1-\theta}$$

and the zero of the gradient is $\frac{n_1}{n}$. Therefore we have

$$\hat{\theta}_{\text{MLE}} = \frac{n_1}{n} = \frac{1}{n}\sum_{i=1}^n x_i$$

II. C. MLE for the multinomial model

Consider $x_1, x_2, \dots, x_n \in \mathcal{X}_K$ i.i.d. observations of $\mathcal{M}(1, \pi)$. We have

$$\ell(\pi) = \sum_{i=1}^{n} \log p(x_i) = \sum_{i=1}^{n} \log \Big(\prod_{k=1}^{K} \pi_k^{x_{ik}} \Big) = \sum_{i=1}^{n} \sum_{k=1}^{K} x_{ik} \log \pi_k = \sum_{k=1}^{K} n_k \log \pi_k$$

where $n_k = \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} \mathbb{1}_{x_{ik}=1}$ is the number of observations of the value k.

We need to maximize this quantity subject to the constraint $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k \geq 0$ for all $k \in [\![1, K]\!]$.

We forget the inequality constraint and we try to minimize $f(\pi) = -\ell(\pi) = -\sum_{k=1}^{K} n_k \log \pi_k$ subject to the constraint $\mathbf{1}^\top \pi = 1$. We introduce the Lagrangian of this problem (see Annex II. for more details):

$$\mathcal{L}(\pi, \lambda) = -\sum_{k=1}^{K} n_k \log \pi_k + \lambda \Big(\sum_{k=1}^{K} \pi_k - 1 \Big)$$

Clearly, as all $(n_k)_{1 \leq k \leq K}$ are nonnegative, f is convex and this problem is a convex optimization problem. Moreover, it is trivial that there exist a strictly feasible point[2], so by SLATER's constraint qualification, the problem satisfies strong duality. Therefore, we have

$$\min_{\pi} f(\pi) = \max_{\lambda} \min_{\pi} L(\pi, \lambda)$$

As $\mathcal{L}(., \lambda)$ is convex, it suffices to find a zero of the gradient of $\mathcal{L}$ w.r.t. π to find $\min_\pi \mathcal{L}(\pi, \lambda)$. This yields

$$\forall k \in [\![1, K]\!], \quad \frac{\partial \mathcal{L}}{\partial \pi_k} = -\frac{n_k}{\pi_k} + \lambda = 0 \quad \Longleftrightarrow \quad \pi_k = \frac{n_k}{\lambda}$$

Substituting these into the constraint $\sum_{k=1}^{K} \pi_k = 1$ we get $\lambda = \sum_{k=1}^{K} n_k = n$. Finally we get the MLE of π:

$$\hat{\pi}_{\mathsf{MLE}} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

II. D. MLE for the univariate Gaussian model

Consider $x_1, x_2, \dots, x_n$ i.i.d. observations a $\mathcal{N}(\mu, \sigma^2)$. We have

$$\begin{aligned} \ell(\mu, \sigma^2) &= \sum_{i=1}^{n} \log p_{\mu,\sigma^2}(x_i) = \sum_{i=1}^{n} \log \Big[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x_i - \mu)^2}{2\sigma^2} \Big) \Big] \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2} \end{aligned}$$

We need to maximize this quantity with respect to μ and σ^2. By taking derivative w.r.t. μ and then σ^2, it is easy to obtain that the pair $(\hat{\mu}_{\mathsf{MLE}}, \hat{\sigma}^2_{\mathsf{MLE}})$, defined by

$$\hat{\mu}_{\mathsf{MLE}} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \text{and} \qquad \hat{\sigma}^2_{\mathsf{MLE}} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

[2] that is to say a vector π such that $\pi_1, \pi_2, \dots, \pi_K$ are positive and $\sum_{k=1}^{K} \pi_k = 1$

is the only stationary point of the likelihood. One can actually check (for example computing the Hessian w.r.t. (μ, σ^2)) that this is actually a maximum. We will have a confirmation of this in the chapter on exponential families (see Chapter 7).

II. E. MLE for the multivariate Gaussian model

Let $x_1, \dots, x_n$ be an i.i.d. sample of $\mathcal{N}(\mu, \Sigma)$ with $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathcal{S}_d^{++}(\mathbb{R})$ definite positive. The $\log$-likelihood is given by:

$$\ell(\mu, \Sigma) = \sum_{i=1}^{n} \log p_{\mu,\Sigma}(x_i) = -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log(\det \Sigma) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^\top \Sigma^{-1}(x_i - \mu)$$

In this case, one should be careful that these $\log$-likelihoods are not concave w.r.t. the pair of parameters (μ, Σ). They are concave w.r.t. μ when Σ is fixed but they are not even concave w.r.t. Σ when μ is fixed.

Let us first differentiate ℓ w.r.t. μ. We need to differentiate for a fixed x:

$$\mu \longmapsto (x - \mu)^\top \Sigma^{-1}(x - \mu)$$

which is equal to $f \circ g$ where:

$$f : \begin{array}{ccc} \mathbb{R}^d & \longrightarrow & \mathbb{R} \\ y & \longmapsto & y^\top \Sigma^{-1} y \end{array} \qquad \text{and} \qquad g : \begin{array}{ccc} \mathbb{R}^d & \longrightarrow & \mathbb{R}^d \\ \mu & \longmapsto & x - \mu \end{array}$$

Using the example of Annex III. , we know that

$$\nabla f_y = \Sigma^{-1} y \qquad \text{and} \qquad \nabla g_\mu = -\mathbf{1}$$

as Σ^{-1} is symmetric. By the differentiation of a composition we obtain:

$$\nabla f \circ g_\mu(h) = \Sigma^{-1}(\mu - x)$$

Thus we have:

$$\nabla_\mu \ell(\mu, \Sigma^{-1}) = \sum_{i=1}^{n} \Sigma^{-1}(\mu - x_i) = \Sigma^{-1}\Big(n\mu - \sum_{i=1}^{n} x_i\Big) = \Sigma^{-1}\Big(n\mu - n\overline{x}\Big)$$

where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. One can check that there is a unique zero of this gradient, which give the MLE of μ:

$$\hat{\mu}_{\mathsf{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Let us now differentiate ℓ w.r.t. $\Lambda = \Sigma^{-1}$. We have:

$$\ell(\mu, \Sigma) = -\frac{nd}{2}\log(2\pi) + \frac{n}{2}\log(\det \Lambda) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^\top \Lambda(x_i - \mu)$$

All terms of the sum are real number, so are equal to their trace. Thus:

$$\begin{aligned}\ell(\mu,\Sigma) &= -\frac{nd}{2}\log(2\pi) + \frac{n}{2}\log(\det\Lambda) - \frac{1}{2}\sum_{i=1}^{n}\operatorname{Tr}\left((x_i-\mu)^\top\Lambda(x_i-\mu)\right)\\ &= -\frac{nd}{2}\log(2\pi) + \frac{n}{2}\log(\det\Lambda) - \frac{n}{2}\operatorname{Tr}(\Lambda\tilde{\Sigma})\end{aligned}$$

where $\tilde{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^\top$ is the empirical covariance matrix.

We need to differentiate $\lambda \longmapsto \log\det(\lambda)$ and $\lambda \longmapsto \operatorname{Tr}(\Lambda\tilde{\Sigma})$. One can obtain (see Annex III. for details):

$$\nabla\log\det(\lambda) = \lambda^{-1} = \Sigma \qquad \text{and} \qquad \nabla\operatorname{Tr}(\Lambda\tilde{\Sigma}) = \tilde{\Sigma}$$

And the gradient of ℓ w.r.t. Λ is:

$$\nabla_\Lambda\ell = \frac{n}{2}(\tilde{\Sigma}-\Sigma)$$

which is equal to zero if and only if $\Sigma = \tilde{\Sigma}$.

Finally we have shown that the pair

$$\hat{\mu} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n}x_i \qquad \text{and} \qquad \hat{\Sigma}_{\mathsf{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(x_i-\overline{x})^\top$$

is the only stationary point of the likelihood. One can actually check (for example computing the Hessian w.r.t. (μ,Σ) that this is actually a maximum. We will also have a confirmation of this in the lecture on exponential families.

> REMARK II. .1. Note that we assumed that Λ was invertible, which is an implicit condition when writing $\log\det\Lambda$. This implies that in a rigorous sense the maximum likelihood estimator is undefined when $\tilde{\Sigma}$ is not invertible. In practice, the maximum likelihood estimator is extended by continuity to the rank deficient case.

CHAPTER 2

Regression

In the last chapter, we considered a model with one node, i.e. with a unique variable and thus distribution. In this lecture, we work with two nodes: one corresponding to an input X, and the other corresponding to an output Y.

Recall that when dealing with two random variables X and Y, one can use a generative model, i.e. which models the joint distribution $p(X, Y)$, or one can use instead a conditional model[1], which models the conditional probability of the output, given the input $p(Y \mid X)$. The two following models, linear regression and logistic regression, are conditional models.

I. Linear regression

We consider the following model: we assume that $Y \in \mathbb{R}$ depends linearly on $X \in \mathbb{R}^p$: there exists a $\mathsf{w} \in \mathbb{R}^p$ called weighting vector and $\sigma^2 > 0$ such that

$$Y \mid X \sim \mathcal{N}(\mathsf{w}^\top X, \sigma^2)$$

which can be rewritten as

$$Y = \mathsf{w}^\top X + \varepsilon \qquad \text{where } \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

REMARK I. .1. It is possible to add an offset $\mathsf{w}_0 \in \mathbb{R}$, that is, if the model is $Y = \mathsf{w}^\top X + \mathsf{w}_0 + \varepsilon$, we can redefine a weighting vector $\tilde{\mathsf{w}} = (\mathsf{w}, \mathsf{w}_0) \in \mathbb{R}^{p+1}$ such that

$$Y = \tilde{\mathsf{w}}^\top \begin{pmatrix} X \\ 1 \end{pmatrix} + \varepsilon$$

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be i.i.d. observations. Each y_i is a label (a decision) on the observation x_i. We consider the conditional distribution of all outputs given all inputs:

$$p_{\mathsf{w},\sigma^2}(y \mid x) = \prod_{i=1}^{n} p_{\mathsf{w},\sigma^2}(y_i \mid x_i)$$

[1] often considered equivalent to the slightly different concept of discriminative model

The associated $\log$-likelihood has the following expression:

$$\ell(\mathsf{w}, \sigma^2) = \sum_{i=1}^{n} \log p_{\mathsf{w},\sigma^2}(y_i \mid x_i) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \mathsf{w}^\top x_i)^2}{\sigma^2}$$

The minimization problem with respect to w can now be reformulated as:

$$\min_{\mathsf{w}} \quad \frac{1}{2n}\sum_{i=1}^{n}(y_i - \mathsf{w}^\top x_i)^2$$

DEFINITION I. .1. [DESIGN MATRIX]
The design matrix $\mathbf{X} \in \mathcal{M}_{n\,p}(\mathbb{R})$ is defined as:

$$\mathbf{X} = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix}$$

The minimization problem over w can be rewritten in a more compact way as:

$$\min_{\mathsf{w}} \quad \frac{1}{2n}\|y - \mathbf{X}\mathsf{w}\|_2^2$$

Introduce now the following function:

$$f : \mathsf{w} \longmapsto \frac{1}{2n}\|y - \mathbf{X}w\|_2^2 = \frac{1}{2n}(y^\top y - 2\mathsf{w}^\top\mathbf{X}^\top y + \mathsf{w}^\top\mathbf{X}^\top\mathbf{X}\mathsf{w})$$

f is strictly convex if and only if its Hessian matrix is nonsingular.

REMARK I. .2. This is never the case when $n < p$ (and we say that we deal with underdetermined problems). Most of the time, the Hessian matrix is nonsingular when $n \geq p$. When this is not the case, we often use the TYCHONOV regularization, which adds a penalization of the ℓ_2-norm of w by minimizing $f(w) + \lambda\|\mathsf{w}\|_2^2$ with some hyperparameter $\lambda > 0$.

The gradient of f is

$$\nabla f(\mathsf{w}) = \frac{1}{n}\mathbf{X}^\top(\mathbf{X}w - y)$$

which is equal to zero if and only if $\mathbf{X}^\top\mathbf{X}w = \mathbf{X}^\top y$. This equation is known as the normal equation. If $\mathbf{X}^\top\mathbf{X}$ is nonsingular, then the optimal weighting vector is

$$\hat{\mathsf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top y = \mathbf{X}^\dagger y$$

where $\mathbf{X}^\dagger = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is the MOORE-PENROSE pseudo-inverse of X.

REMARK I. .3. If $\mathbf{X}^\top\mathbf{X}$ is singular, the solution is not unique anymore, and for any $h \in \ker(\mathbf{X})$, $\hat{\mathsf{w}} = (\mathbf{X}^\top\mathbf{X})^\dagger\mathbf{X}^\top y + h$ is an admissible solution. In that case however it would be necessary to use regularization.

REMARK I. .4. The computational cost to evaluate the optimal weighting vector from $\mathbf{X}$ and y is $\mathcal{O}(p^3)$ (we use a CHOLESKY decomposition of $\mathbf{X}^\top\mathbf{X}$ and solve two triangular systems).

Now, let us differentiate ℓ w.r.t. σ^2: we have

$$\nabla_{\sigma^2}\ell(\mathsf{w},\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \mathsf{w}^\top x_i)^2$$

Setting $\nabla_{\sigma^2}\ell(\mathsf{w},\sigma^2)$ to zero gives the MLE of σ^2:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathsf{w}^\top x_i)^2$$

REMARK I. .5. In practice, whenever we use a data matrix $\mathbf{X}$ in machine learning, we first preprocess it to try and avoid that it would be too badly conditioned, so to avoid numerical issues. Two main operations are applied columnwise: first, a centering (remove the mean of the coefficients) and a normalization (divide coefficients from a column by the standard deviation of the column vector). Note that this preprocessing does not guarantee that the matrix we obtain is well-conditioned: in particular, it can be low rank ...

II. Logistic regression

DEFINITION II. .1. [SIGMOID FUNCTION]
The sigmoid function is defined as:

$$\begin{array}{rccc} \sigma: & \mathbb{R} & \longrightarrow & [0,1] \\ & z & \longmapsto & \frac{1}{1+\mathrm{e}^{-z}} \end{array}$$

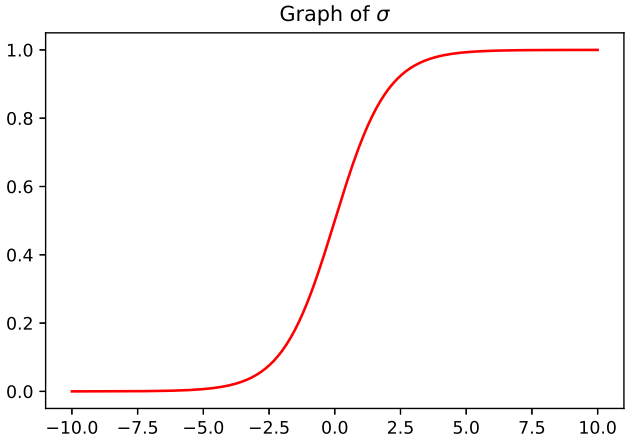


Figure 2.1: The sigmoid function

PROPOSITION II. .1. *σ satisfies the following properties:*

$$\forall z \in \mathbb{R}, \quad \sigma(-z) = 1 - \sigma(z) \quad \textit{and} \quad \sigma'(z) = \sigma(z)(1-\sigma(z)) = \sigma(z)\sigma(-z)$$

Consider now another model where $Y \in \{0,1\}$ and $X \in \mathbb{R}^p$. We assume that Y follows a BERNOULLI distribution with parameter $\theta = \sigma(\mathsf{w}^\top x)$ where $\mathsf{w} \in \mathbb{R}^p$ is a fixed weighting vector. The problem is to estimate θ.

REMARK II. .1. There again we can add an offset.

The conditional distribution is given by

$$p_\theta(Y = y \,|\, X = x) = \theta^y (1-\theta)^{1-y} = \sigma(\mathsf{w}^\top x)^y \sigma(-\mathsf{w}^\top x)^{1-y}$$

Given an i.i.d. training set $(x_1, y_1), \dots, (x_n, y_n)$, we can compute the $\log$-likelihood:

$$\ell(\mathsf{w}) = \sum_{i=1}^n y_i \log \sigma(\mathsf{w}^\top x_i) + (1-y_i) \log \sigma(-\mathsf{w}^\top x_i)$$

In order to minimize the $\log$-likelihood, since $z \longmapsto \log(1+\mathrm{e}^{-z})$ is a convex function and $\mathsf{w} \longmapsto \mathsf{w}^\top x_i$ is linear, we calculate its gradient. With $\eta_i = \sigma(\mathsf{w}^\top x_i)$:

$$\nabla \ell(\mathsf{w}) = \sum_{i=1}^n y_i x_i \frac{\sigma(\mathsf{w}^\top x_i)\sigma(-\mathsf{w}^\top x_i)}{\sigma(\mathsf{w}^\top x_i)} - (1-y_i) x_i \frac{\sigma(\mathsf{w}^\top x_i)\sigma(-\mathsf{w}^\top x_i)}{\sigma(-\mathsf{w}^\top x_i)} = \sum_{i=1}^n x_i (y_i - \eta_i)$$

Thus the gradient vanishes if and only if $\sum_{i=1}^n x_i(y_i - \eta_i) = 0$. This equation is nonlinear and we need an iterative optimization method to solve it (see Annex IV. for more details). For this purpose, we derive the Hessian matrix of ℓ:

$$H\ell(\mathsf{w}) = \sum_{i=1}^n x_i (0 - \sigma'(\mathsf{w}^\top x_i)\sigma'(-\mathsf{w}^\top x_i) x_i^\top) = -\sum_{i=1}^n \eta_i(1-\eta_i) x_i x_i^\top = -\mathbf{X}^\top \operatorname{diag}(\eta(1-\eta))\mathbf{X}$$

We focus on the NEWTON's algorithm and try to apply it for logistic regression.

The second-order TAYLOR-expansion of the loss function leads to

$$\ell(\mathsf{w}) = \ell(\mathsf{w}^t) + (\mathsf{w} - \mathsf{w}^t)^\top \nabla \ell(\mathsf{w}^t) + \frac{1}{2}(\mathsf{w} - \mathsf{w}^t)^\top H\ell(\mathsf{w}^t)(\mathsf{w} - \mathsf{w}^t) + o\Big(\big\| \mathsf{w} - \mathsf{w}^t \big\|^2 \Big)$$

With $h = \mathsf{w} - \mathsf{w}^t$ and the previous expressions of ℓ, $\nabla\ell$ and $H\ell$, the minimization problem becomes:

$$\min_h h^\top \mathbf{X}^\top (y - \eta) - \frac{1}{2} h^\top \mathbf{X}^\top \operatorname{diag}(\eta(1-\eta))\mathbf{X} h$$

This leads, according to the method, to set $\mathsf{w}^{t+1} = \mathsf{w}^t + H\ell(\mathsf{w}^t)^{-1}\nabla\ell(\mathsf{w})$. The minimization problem above can be seen as some weighted linear regression over h of some function of the form $\sum_{i=1}^n \frac{(\tilde{y}_i - x_i^\top h)^2}{\sigma_i^2}$, where $\tilde{y}_i = y_i - \eta_i$ and $\sigma_i^2 = [\eta_i(1-\eta_i)]^{-1}$. Thus, this method is often refered as the iterative reweighted least squares algorithm.

III. Generative models [todo]

This part briefly presents the FISHER linear discriminant also known as the linear discriminant analysis. Suppose that we have $X \in \mathbb{R}^p$ and $Y \in \{0, 1\}$. Then by the BAYES formula:

$$p(Y = 1 \,|\, X = x) = \frac{p(X = x \,|\, Y = 1)p(Y = 1)}{p(X = x \,|\, Y = 1)p(Y = 1) + p(X = x \,|\, Y = 0)p(Y = 0)}$$

The assumption then consists in considering $p(X = x \,|\, Y = 0) \sim \mathcal{N}(x, \mu_0, \Sigma_0)$ and $p(X = x \,|\, Y = 1) \sim \mathcal{N}(x, \mu_1, \Sigma_1)$. FISHER's assumption is the assumption that $\Sigma_1 = \Sigma_0 = \Sigma$.

CHAPTER 3

Unsupervised classification

In this chapter we run into a classification problem with more than two classes. We assume that $Y \in [\![1, K]\!]$ for a fixed $K \geq 2$.

To talk about estimation of "hidden" parameters, French speaking people and English speaking people use different terms which can lead to some confusions. Within a supervised framework, English people would prefer to use the term "classification" whereas the French use the term "discrimination". Within an unsupervised context, English people would rather use the term "clustering", whereas French people would use "classification" or "classification non supervisée". In the following we will only use the English terms.

Unsupervised learning consists in finding a label prediction function based on unlabeled training data only. In the case where the learning problem is a classification problem, and under the assumption that the classes form clusters in input space, the problem reduces to a clustering problem, which consists in finding groups of points that form denser clusters.

When the clusters are assumed to be isotropic the formulation of the K-means algorithm is appropriate.

I. K-means

K- means clustering is a method of vector quantization. It is an algorithm of alternate minimization that aims at partitioning n observations into K clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype to the cluster (see Figure 3.1).

I. A. The K-means algorithm

We will use the following notations:

- $x_1, \ldots, x_n \in \mathbb{R}^p$ are the observations we want to partition into K clusters,
- $\mu_1, \ldots, \mu_K \in \mathbb{R}^p$ are the means: μ_k is the center of cluster k. We will denote $\mu = (\mu_1, \ldots, \mu_K)$.
- To each x_i we associate the indicator variable $z_i = (\mathbb{1}_{i \in C_1}, \ldots, \mathbb{1}_{i \in C_K})$ where C_k are the indices of points belonging to cluster k. We set $z = (z_1, \ldots, z_n)$.

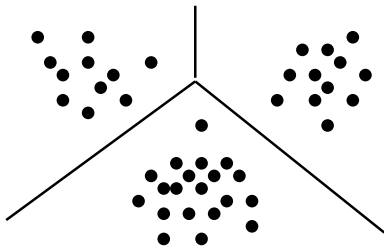


Figure 3.1: Clustering on a 2D point data set with 3 clusters **[todo]**

We also define the distortion as the function J defined by:

$$J(\mu, z) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \|x_i - \mu_k\|^2$$

The aim of the algorithm is to minimize J. To do so we proceed with an alternating minimization:

Algorithm 1: K-means

Input : $x_1, \dots, x_n, K, \mu$
Output: μ, z

while *no convergence* **do**
  $z = \text{argmin}_z \, J(\mu, z)$
  $\mu = \text{argmin}_\mu \, J(\mu, z)$
end

During the minimization w.r.t. z, we set $i \in C_k$ thus $z_{ik} = 1$ if $k \in \text{argmin}_{k'} \|x_i - \mu_{k'}\|_2^2$. In other words we associate to each x_i the cluster with nearest center μ_k.

During the minimization w.r.t. μ, one can show[1] that the new μ is defined by

$$\forall k \in [\![1, K]\!], \quad \mu_k = \frac{\sum_{i=1}^n z_{ik} x_i}{\sum_{i=1}^n z_{ik}} = \frac{\sum_{i \in C_k} x_i}{|C_k|}$$

that is to say each cluster's center is the average of the points in the cluster.

REMARK I. .1. The step of minimization with respect to z is equivalent to allocating the x_i in the VORONOÏ cells which centers are the $(\mu_k)_{1 \le k \le K}$.

I. B. Convergence and initialization

We can show that this algorithm converges in a finite number of iterations. Therefore the convergence could be local, thus it introduces the problem of initialization.

Random restarts A classic method consists in using random restarts. By choosing several random vectors μ, we can compute the algorithm for each case and finally keep the partition which

[1] by setting to zeros the gradient of J with respect of μ, as $\nabla_{\mu_k} J = -2 \sum_{i=1}^n z_{ik}(x_i - \mu_k)$

minimizes the distortion. Thus we hope that at least one of the local minimum is close enough to a global minimum.

K-means++ One other well known method is the K-means++ algorithm, which aims at correcting a major theoretic shortcomings of the K-means algorithm: the approximation found can be arbitrarily bad with respect to the objective function compared to the optimal clustering. The K-means++ algorithm addresses this obstacles by specifying a procedure to initialize the cluster centers before proceeding with the standard K-means optimization iterations. With the K-means++ initialization, the algorithm is guaranteed to find a solution that is $\mathcal{O}(\log K)$ competitive to the optimal K-means solution.

The intuition behind this approach is that it is a clever thing to well spread out the K initial cluster centers. At each iteration of the algorithm we will build a new center. We will repeat the algorithm until we have K centers. Here are the steps of the algorithm :

Algorithm 2: Initialization of K-means++

Input : $x_1, \dots, x_n, K$
Output: μ

Choose μ_1 uniformly among $x_1, \dots, x_n$
for $k \in [\![2, K]\!]$ **do**
  Set $D_i = \min_{k'<k} d(x_i, \mu_{k'})$ for $i \in [\![1, n]\!]$
  Choose μ_k as x_i with probability $D_i^2 / \sum_{i=1}^n D_i^2$.
end

We see that we have now built K vectors with respect to our first intuition which was to well-spread out the centers (because we used a well chosen weighted probability). We can now use those vectors as the initialization of our standard K-means algorithm.

I. C. Choice of K

The parameter K is an hyperparameter that we need to specify to the algorihm.

It is important to point out that the choice of K is not universal. Indeed, we see that if we increase K, the distortion J decreases, until it reaches 0 when $K = n$, that is to say when each data point is the center of its own center. To address this issue one solution could be to add to J a penalty term over K. Usually it takes the following form:

$$J(\mu, z, K) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \|x_i - \mu_k\|_2^2 + \lambda K$$

for an hyperparameter $\lambda > 0$ which is again arbitrary.

I. D. Other problems

We can also point out that K-means will work pretty well when the width of the different clusters are similar, for example if we deal with spheres. But clustering by K-means could also be disap-

pointing in some cases such as the example given in Figure 3.2.

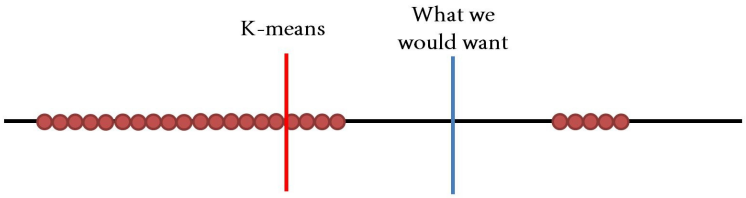


Figure 3.2: Example where K-means does not provide a satisfactory clustering result

Using Gaussian mixtures provides a way to avoid this problem (see next part).

II. Expectation Maximization algorithm

The Expectation Maximization algorithm (EM) is an iterative method for finding maximum likelihood estimates of parameters in statistical models, where the models depend on unobserved latent or hidden variables Z. Latent variables are variables that are not directly observed but are rather inferred from other variables that are observed.

Previous algorithms aimed at estimating the parameter θ that maximized the likelihood of $p_\theta(x)$, where x is the vector of observed variables.

In this section we proceed differently, by assuming an observation x of a r.v. X (our data) depends on a second random variable Z with observation z unknown (cluster center for example). Our model is then the joint density $p_\theta(X, Z)$ depending on a parameter $\theta \in \Theta$, and the goal is to maximize $p_\theta(x) = \sum_z p_\theta(x, z)$.

We can already infer that, because of the sum, the problem should be slightly more difficult than before. Indeed, taking the $\log$ of our probability would not lead to a simple convex optimization problem. In the following we will see that EM is a method to solve those kinds of problems.

II. A. An example

Let us present a simple example to illustrate what we just said. The probability density represented on Figure 3.3 is akin to an average of two Gaussians. Thus, it is natural to use a mixture model and to introduce a hidden variable z, following a BERNOULLI distribution defining which Gaussian the point is sampled from.

In this example we have $z \in \{1, 2\}$ and $x \mid z = i \sim \mathcal{N}(\mu_i, \Sigma_i)$. The density $p(x)$ is a convex combination of normal densities:

$$p(x) = p(x, z = 1) + p(x, z = 2) = p(x \mid z = 1)p(z = 1) + p(x \mid z = 2)p(z = 2)$$

This is a mixture model. It represents a simple way to model complicated phenomena.

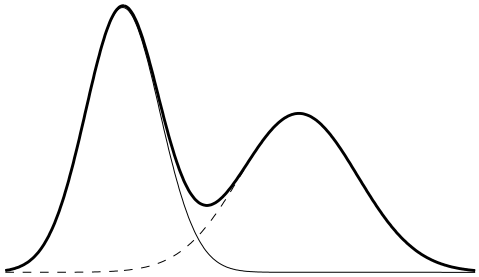


Figure 3.3: Average of two Gaussian distributions

II. B. Our objective: maximum likelihood

Let $\mathcal{D} = \{(x_1, z_1), \dots, (x_n, z_n)\}$ are n i.i.d. observations of the random variable (X, Z). The aim is to maximize the *incomplete* likelihood or $\log$-likelihood, where $x = (x_1, \dots, x_n)$ and $z = (z_1, \dots, z_n)$:

$$\mathcal{L}(x) = \sum_z p_\theta(x, z) = \prod_{i=1}^n \sum_{z_i} p_\theta(x_i, z_i) \qquad \ell(x) = \log\Big(\sum_z p_\theta(x, z)\Big) = \sum_{i=1}^n \log\Big(\sum_{z_i} p_\theta(x_i, z_i)\Big)$$

A direct way to solve this problem is for example to do a gradient ascent. EM algorithm will be another way to do it.

II. C. The EM algorithm

We recall the JENSEN's inequality:

PROPOSITION II. .1. [JENSEN'S INEQUALITY]
Let $f : \mathbb{R} \longrightarrow \mathbb{R}$ be a convex function and X is an integrable random variable. Then

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

In addition, if f is strictly convex, then we have equality if and only if X is constant a.s..

Let us introduce a nonnegative function $q(z)$ such that $\sum_z q(z) = 1$. Using the concavity of $\log$ and JENSEN's inequality, one has:

$$\begin{aligned}\ell(x) &= \log\Big(\sum_z p_\theta(x, z)\Big) = \log\Big(\sum_z \Big(\frac{p_\theta(x, z)}{q(z)}\Big) q(z)\Big) \\ &\geq \sum_z q(z) \log\Big(\frac{p_\theta(x, z)}{q(z)}\Big) = \sum_z q(z) \log p_\theta(x, z) - \sum_z q(z) \log q(z) := \mathcal{L}(q, \theta)\end{aligned}$$

with equality if and only if

$$\forall z, \quad q(z) = \frac{p_\theta(x, z)}{\sum_{z'} p_\theta(x, z')} = p_\theta(z \,|\, x)$$

by strict concavity of $\log$.

We have just proved:

Proposition II. .2. *$\forall \theta \in \Theta$ and q, we have:*

$$\log p_\theta(x) \geq \mathcal{L}(q, \theta)$$

with equality if and only if $q(z) = p_\theta(z \mid x)$ for all z.

Thus we have introduced an auxiliary function $\mathcal{L}(q, \theta)$ that is always below the function $\log(p_\theta(x))$.

As with K-means, EM algorithm consists in an alternating minimization:

Algorithm 3: EM

Input : $x_1, \ldots, x_n, \theta$
Output: θ, z

while *no convergence* **do**
  $q = \operatorname{argmax}_q \mathcal{L}(q, \theta)$ `// E-step`
  $\theta = \operatorname{argmax}_\theta \mathcal{L}(q, \theta)$ `// M-step`
end
$z = \operatorname{argmax}_z p_\theta(z \mid x)$

Algorithm properties

- EM is an ascent algorithm, indeed it goes up in term of likelihood (compare to before where we were descending along the distortion).
- The sequence of $\log$-likelihoods converges to a local maximum because we are dealing here with a non-convex problem (see the illustration in Figure 3.4).
  As it was already the case for K-means, we can reiterate the result in order to be more confident, keeping the result with highest likelihood.

Initialization Because EM gives a local maximum, it is clever to choose an initial θ relatively close to the final solution. For Gaussian mixtures, it is quite usual to initiate EM by the output of K-means, which gives a good initialization in practice, but with a large variance.

In practice: the EM recipe In practice we do at each iteration of the algorithm:

(i) Compute the probability of $z \mid x$, $p_\theta(z \mid x)$, which corresponds to the new $q(z)$.
(ii) Write the *complete* $\log$-likelihood $\ell_c(x, z) = \log(p_\theta(x, z))$.
(iii) E-Step: calculate $\mathbb{E}_{Z \mid X}(\ell_c(x, z))$ the expected value of the complete $\log$-likelihood function, with respect to the conditional distribution of $Z \mid X$ under the current estimate of the parameter θ.
(iv) M-Step: find θ by maximizing $\mathcal{L}(q, \theta)$ with respect to θ.

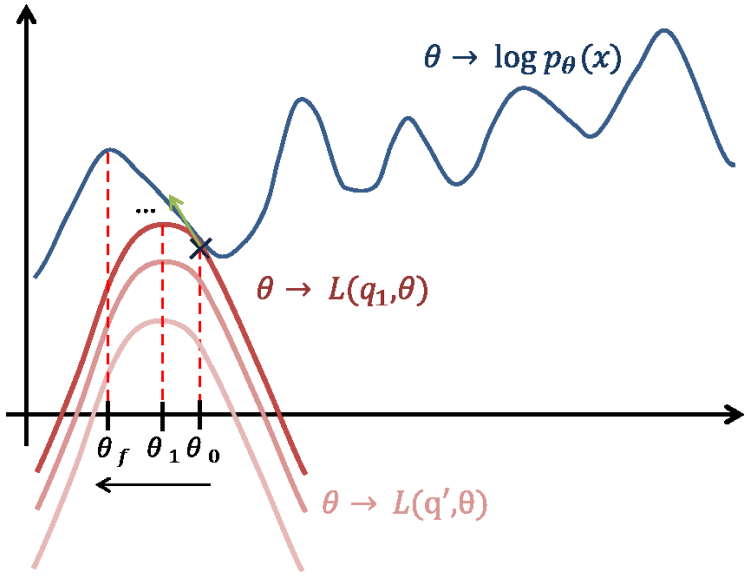


Figure 3.4: Convergence of the EM algorithm to a local maximum

II. D. Gaussian mixture

In this section we assume (X, Z) is such that $X \in \mathbb{R}^d$ and $Z \in [\![1, K]\!]$ with $Z \sim \mathcal{M}(1, \pi_1, \dots, \pi_K)$ and $X \mid Z = k \sim \mathcal{N}(\mu_k, \Sigma_k)$.

We set $\theta = (\pi, \mu, \Sigma)$ and we want to apply the EM algorithm to this model:

(i) To compute $p_\theta(z \mid x)$, we use a BAYES formula:

$$\tau_{i,k}(\theta) := p_\theta(z_i = k \mid x_i) = \frac{p_\theta(x_i \mid z_i = k) p_\theta(z_i = k)}{p_\theta(x_i)} = \frac{\pi_k \varphi(x_i, \mu_k, \Sigma_k)}{\sum_{k'} \pi_{k'} \varphi(x_i, \mu_{k'}, \Sigma_{k'})}$$

where $\varphi(x, \mu, \Sigma)$ is the density function of $\mathcal{N}(\mu, \Sigma)$ at x:

$$\varphi(x, \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma)}} \exp\Big(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\Big)$$

Suppose we are at iteration t:

(ii) We write the complete $\log$-likelihood of the problem:

$$\begin{aligned}\ell_c^{(t)}(x, z) = \log p_{\theta^{(t)}}(x, z) &= \sum_{i=1}^n \log p_{\theta^{(t)}}(x_i, z_i) = \sum_{i=1}^n \log(p_{\theta^{(t)}}(z_i) p_{\theta^{(t)}}(x_i \mid z_i)) \\ &= \sum_{i=1}^n \log p_{\theta^{(t)}}(z_i) + \log p_{\theta^{(t)}}(x_i \mid z_i) = \sum_{i=1}^n \sum_{k=1}^K \Big[\log(\pi_k^{(t)}) + \log(\varphi(x_i, \mu_k^{(t)}, \Sigma_k^{(t)}))\Big] \mathbb{1}_{z_i = k}\end{aligned}$$

(iii) E-step: we can now write the expectation of the previous quantity with respect to the conditional distribution of $Z \mid X$. In fact it is equivalent to replace $\mathbb{1}_{z_i=k}$ by $p_{\theta^{(t)}}(z = k \mid x_i) = \tau_{i,k}(\theta^{(t)}) = \tau_{i,k}^{(t)}$, as the other terms of the sum are constant from the point of view of the

conditional probability of $Z \mid X$, and we finally obtain:

$$f(\theta^{(t)}) = \mathbb{E}_{Z \mid X}[\ell_c^{(t)}(x,z)] = \sum_{i=1}^{n}\sum_{k=1}^{K}\Big[\log(\pi_k^{(t)}) + \log(\varphi(x_i,\mu_k^{(t)},\Sigma_k^{(t)}))\Big]\tau_{i,k}^{(t)}$$

(iv) M-step: we need to maximize f w.r.t. θ.

- First we maximize w.r.t. π.
  Maximizing f w.r.t. p corresponds to maximize $\sum_{i=1}^{n}\sum_{k=1}^{K}\log(\pi_k^{(t)})\tau_{i,k}^{(t)}$ under the constraints $\sum_{k=1}^{K}\pi_k = 1$ and $\pi_k \geq 0$ for $k \in [\![1,K]\!]$.
  We forget the inequality constraint in a first time and consider the following Lagrangian:

$$\mathcal{L}(\pi,\lambda) = \sum_{i=1}^{n}\sum_{k=1}^{K}\log(\pi_k)\tau_{i,k}^{(t)} + \lambda\Big(1 - \sum_{k=1}^{K}\pi_k\Big)$$

  One has for all $k \in [\![1,K]\!]$:

$$\frac{\partial\mathcal{L}(\pi,\lambda)}{\partial\pi_k} = \frac{1}{\pi_k}\sum_{i=1}^{n}\tau_{i,k}^{(t)} - \lambda$$

  Thus $\nabla_\pi\mathcal{L}(\pi,\lambda)$ vanishes when $\pi_k = \frac{1}{\lambda}\sum_{i=1}^{n}\tau_{i,k}^{(t)}$ for all $k \in [\![1,K]\!]$, which implies:

$$\lambda = \lambda\sum_{k=1}^{K}\pi_k = \sum_{k=1}^{K}\sum_{i=1}^{n}\tau_{i,k}^{(t)} = \sum_{i=1}^{n}\underbrace{\sum_{k=1}^{K}\tau_{i,k}^{(t)}}_{=1} = n$$

  and we deduce that the maximizer $\pi^{(t+1)}$ is defined by (note that it is a non-negative vector):

$$\forall k \in [\![1,K]\!], \quad p_k^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\tau_{i,k}^{(t)}$$

- Then we try to maximize w.r.t. to μ. This corresponds to maximize:

$$\mathcal{L}(\mu,\Sigma) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{i,k}^{(t)}(x_i-\mu_k)^\top\Sigma_k^{-1}(x_i-\mu_k)$$

  Fixing $k \in [\![1,K]\!]$, one has using the results of Annex III. :

$$\nabla_{\mu_k}\mathcal{L}(\mu,\Sigma) = \sum_{i=1}^{n}\tau_{i,k}^{(t)}\Sigma_k^{-1}(x_i-\mu_k)$$

  which vanishes for $\mu_k = \frac{\sum_{i=1}^{n}\tau_{i,k}^{(t)}x_i}{\sum_{i=1}^{n}\tau_{i,k}^{(t)}}$. Thus we have the following expression of $\mu^{(t+1)}$:

$$\forall k \in [\![1,K]\!], \quad \mu_k^{(t+1)} = \frac{\sum_{i=1}^{n}\tau_{i,k}^{(t)}x_i}{\sum_{i=1}^{n}\tau_{i,k}^{(t)}}$$

- Finally we can maximize w.r.t. Σ. We need to maximize if $\Lambda_k = \Sigma_k^{-1}$:

$$\mathcal{L}(\mu, \Lambda) = \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{i,k}^{(t)} \left[\frac{1}{2}\log(\det(\Lambda_k)) - \frac{1}{2}(x_i - \mu_k)^\top \Lambda_k (x_i - \mu_k)\right]$$

Fixing $k \in [\![1, K]\!]$, one has using the results of Annex III. :

$$\nabla_{\Lambda_k}\mathcal{L}(\mu, \Lambda) = \sum_{i=1}^{n} \tau_{i,k}^{(t)} \left[\frac{1}{2}\Lambda_k^{-1} - \frac{1}{2}(x_i - \mu_k)(x_i - \mu_k)^\top\right]$$

that will be equal to 0 if

$$\Sigma_k = \frac{\sum_{i=1}^{n} \tau_{i,k}^{(t)}(x_i - \mu_k)(x_i - \mu_k)^\top}{\sum_{i=1}^{n} \tau_{i,k}^{(t)}}$$

Thus we can define $\Sigma^{(t+1)}$:

$$\forall k \in [\![1, K]\!], \quad \Sigma_k^{(t+1)} = \frac{\sum_{i=1}^{n} \tau_{i,k}^{(t)}(x_i - \mu_k^{(t+1)})(x_i - \mu_k^{(t+1)})^\top}{\sum_{i=1}^{n} \tau_{i,k}^{(t)}}$$

REMARK II. .1. The M-step corresponds to the estimation of means in K-means.

Possible forms for Σ We can add some constraints on the form of each $(\Sigma_k)_{1\leq k\leq K}$, depending of our model assumptions. The most frequent ones are the following:

- isotropic: $\Sigma_k = \sigma_k^2 I_d$: the cluster is a sphere and there is only one parameter,
- diagonal: Σ_k is a diagonal matrix: the cluster is an ellipse oriented along the axis, there are d parameters.
- general: we make no assumptions on Σ_k: the cluster is an ellipse, there are $\frac{d(d+1)}{2}$.

CHAPTER 4

Bayesian method

[todo]

I. Introduction

Vocabulary:

- *a priori* or prior: $p(\theta)$
- *likelihood*: $p(x \mid \theta)$
- *marginal* likelihood: $\int p(x \mid \theta)\, p(\theta)\, d\theta$
- *a posteriori* or posterior: $p(\theta \mid x)$

Caricature Bayesian vs Frequentist:

1. the *Bayesian* is "optimistic": he thinks that he can come up with good models and obtain a method by "pulling the Bayesian crank" (basically a high dimensional integral),
2. the *frequentist* is more "pessimistic" and uses analysis tools.

The Bayesian formulation enables us to introduce the a priori information in the process of estimation. For instance , let's imagine that we play heads or tails. The Bayesian model is:

$$X_i \in \{0,1\}, \qquad X_i \mid \theta \sim Ber(\theta), \quad p(x_i \mid \theta) = \theta^{x_i} (1-\theta)^{1-x_i}$$

the graphical model associated is represented on Figure 4.1.

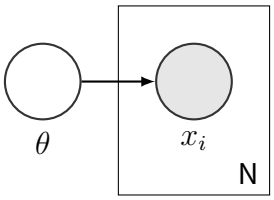


Figure 4.1: Graphical model of the biased coin game

Now we can compute the posterior:

$$p(\theta \mid x_{1:n}) \propto p(x_{1:n} \mid \theta) p(\theta)$$

then

$$p(\theta \mid x_{1:n}) = \theta^{n_1} (1-\theta)^{n-n_1} \mathbf{1}_{[0,1]}(\theta) = Beta(\alpha, \beta)$$

where $n_1 = \sum_{i=1}^{n} x_i$ is the number of 1, $\beta = n - n_1 + 1$ and $\alpha = n_1 + 1$.

Question: what is the probability of head on the next flip?

- Frequensist: $\hat{\theta}_{ML} = n_1/n$ by a maximum likelihood approach.
- Bayesian: $p(x_{n+1} \mid x_{1:n}) = \int p(x_{n+1} \mid \theta) p(\theta \mid x_{1:n}) d\theta$, where $p(\theta \mid x_{1:n}) d\theta$ is the posterior distribution. Then,

  $$\hat{\theta}_B = \frac{\alpha}{\alpha + \beta} = \frac{n_1 + 1}{n + 2}$$

  hence,

  $$\hat{\theta}_B = \frac{n_1}{n} \left[\frac{n}{n+2} \right] + \frac{1}{2} \left[\frac{2}{n+2} \right] = \rho_n \hat{\theta}_{ML} + (1 - \rho_n) \hat{\theta}_{prior}$$

  is a convex combination of $\hat{\theta}_{ML}$ and $\hat{\theta}_{prior}$. Then we can notice that for $n = 0$, the quantity $\hat{\theta}_B = \frac{1}{2}$ whereas $\hat{\theta}_{ML}$ is not defined. It underlines the importance of the prior distibution:
  - with an "unknown" coin, we've got the information a priori : we'll use the uniform law for $p(\theta)$.
  - with a "normal" coin , we'll use a distribution with an important concentration of mass around 0,5 for $p(\theta)$.

  For a Bayesian, offering a "limited" estimator, as the maximum likelihood estimator, which gives a unique value for θ, is not enough because the estimator itself do not translate the inherent uncertainty of the learning process. Thus, its estimator will be the density a posteriori, obtained from the Bayes rule, which is written in continuous notations as:

  $$p(\theta \mid x) = \frac{p(x \mid \theta) p(\theta)}{\int p(x \mid \theta) p(\theta) d\theta}$$

  The Bayesian specifies the uncertainty with distributions that form its estimator, rather than combining an estimator with confidence intervals.
  If the Bayesian is forced to produce a limited estimator, he uses the expectation of the underlying quantity under the a posteriori distribution; for instance for θ:

  $$\mu_{post} = \mathbb{E}[\theta \mid D] = \mathbb{E}[\theta \mid x_1, x_2, \ldots, x_n] = \int \theta p(\theta \mid x_1, x_2, \ldots, x_n) d\theta$$

  For more details about Bayesians see subsection IV. and IV. A. in annex.
  We then need to show that $\hat{\theta}_{ML} \to \theta^*$. Its variance is the variance of a Beta law

  $$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \left(\frac{n_1}{n}\right)\left(1 - \frac{n_1}{n}\right) \cdot O\left(\frac{1}{n}\right) = \hat{\theta}_{ML}\left(1 - \hat{\theta}_{ML}\right) O\left(\frac{1}{n}\right)$$

  then the posterior covariance vanishes and

  $$\hat{\theta}_B \overset{a.s.}{\to} \hat{\theta}_{ML} \overset{a.s.}{\to} \theta^*$$

  where θ^* is the "true" parameter of the model.

II. Bernstein von Mises Theorem

It says that if prior puts non-zero mass around the true model θ^*, then posterior asymptotically concentrate around θ^* as a Gaussian.

i. Revisiting example Consider repeating several times the experiment above: T coins picked randomly each flipped n times. (Figure 13.1)

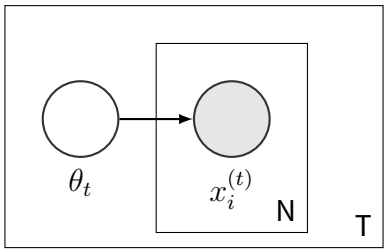


Figure 4.2: Graphical model of the biased coin game repeated T times

As a frequentist, empirical distribution on $x_{1:n}$ will converge (as $T \to \infty$) to

$$p\left(x_1, \ldots, x_n\right) = \int_\theta \left(\prod_{i=1}^{n} p(x_i \mid \theta) \right) p(\theta) d\theta$$

where $p(\theta)$ is the distribution of coins of parameter θ in the jar and $\prod_{i=1}^{n} p\left(x_i \mid \theta\right)$ is the mixture distribution. Note that $X_1, \ldots, X_n$ are **NOT** independent.

On the other hand, for all $\pi \in \mathcal{S}_n$

$$p\left(x_1, \ldots, x_n\right) = p\left(x_{\pi(1)}, \ldots, x_{\pi(n)}\right)$$

III. Exchangeable situations

a. Exchangeablility

The random variables $X_1, X_2, \ldots, X_n$ are exchangeable if they have the same distribution as $X_{\pi(1)}, X_{\pi(2)}$, $\ldots, X_{\pi(n)}$ for any permutation of indices $\pi \in \mathcal{S}_n$.

b. Infinite Exchangeablility

The definition naturally generalizes to infinite families (indexed by $\mathbb{N}$). The random variables $X_1, X_2, \ldots$ are exchangeable if every finite subfamily $X_{i_1}, \ldots, X_{i_n}$ is exchangeable.

c. de Finetti's theorem

$X_1, X_2, \ldots$ are infinitely exchangeable, if and only if $\exists!\ p(\theta)$ (on some space Θ) such that

$$\forall n \in \mathbb{N},\ p\left(x_1, x_2, \ldots, x_n\right) = \int \left(\prod_{i=1}^{n} p\left(x_i \,|\, \theta\right) \right) p(\theta) d\theta$$

d. Why do we care about exchangeable situations?

The i.i.d. variables are a particular case of the situation of exchangeable variables, that we see in practice. However when the i.i.d. data are combined with non scalar observations, the different components are no longer independent. In some cases, those components are nonetheless exchangeable. For instance in a text, words are shown as sequences that are not exchangeable because of the syntax. But if we forget the order of the words as in the "bag of word" model, then the components are exchangeable. It's the basic principle used in the LDA model.

e. Multinomial example

Let $X \,|\, \theta \sim Mult(\theta, 1)$ where $\theta \in \Delta_k$ i.e.

$$p(X = l \,|\, \theta) = \theta_l \quad \text{and} \quad \sum_{l=1}^{k} \theta_l = 1,\ 0 \leq \theta_l \leq 1.$$

for that distribution we have,

$$\hat{\theta}_l^{ML} = \frac{n_l}{n}$$

hence if $k \geq n$ there exists a l such that $\hat{\theta}_l^{ML} = 0$.

In that case this frequentist model overfits. In the Bayesian model one puts a prior on $\Delta_k = \Theta$, but which one? A convenient property of prior families is "conjugacy", introduced below:

i. Conjugacy Consider a family of distribution

$$F = \left\{ p(\theta \,|\, \alpha)\ :\ \alpha \in \mathcal{A} \right\}.$$

One says that F is a "conjugate family" for the observation model $p(x \,|\, \theta)$ if the posterior

$$p(\theta \,|\, x, \alpha) = \frac{p(x \,|\, \theta) p(\theta \,|\, \alpha)}{p(x \,|\, \alpha)}$$

belongs to the same family F than the prior, i.e.

$$\exists \alpha' \in \mathcal{A} \quad s.t \quad p(\theta \,|\, x, \alpha) = p(\theta \,|\, \alpha')$$

For the multinomial distribution it gives us

$$p(x_{1:n} \,|\, \theta) = \prod_{l=1}^{n} p(x_l \,|\, \theta) = \prod_{l=1}^{n} \theta_l^{n_l}$$

so if $p(\theta) \propto \prod_{l=1}^{n} \theta_l^{\alpha_l}$, then $p(x_{1:n} \,|\, \theta) \propto \prod_{l=1}^{n} \theta_l^{\beta_l}$.

f. Dirichlet Distribution

The Dirichlet distribution is the conjugate of the Multinomial law (see on Wikipédia for more details).

$$p(\theta_1, \theta_2, \ldots, \theta_K) = \frac{\Gamma(\alpha_1 + \alpha_2 + \ldots + \alpha_K)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_K)} \theta_1^{\alpha_1 - 1}\theta_2^{\alpha_2 - 1}\ldots\theta_K^{\alpha_K - 1} d\mu(\theta)$$

Where μ stands for the uniform measure on $\Delta_K = \left\{ s \in \mathbb{R}^K \mid \sum_i s_i = 1; \forall i, s_i \geq 0 \right\}$ (K-dim simplex).

- $\mathbb{E}[\theta_l \mid \alpha_1, \ldots, \alpha_K]$,
- $\mathbb{V}(\theta_l) \equiv O\left(\frac{1}{\sum_{j=1}^K \alpha_j}\right)$,
- If $\alpha_l = 1$ for all l then one gets an uniform distribution,
- if $k = 2$ one gets the Beta distribution,
- if there exists l such that $\alpha_l < 1$ one gets a $\cup$ shape distribution,
- if $\alpha_l \geq 1$ for all l, one gets a $\cap$ (unimodal bump).

For the multinomial model, if the we assume that the prior is

$$p(\theta) = Dir(\theta \mid \alpha)$$

then the posterior is

$$p(\theta \mid x_{1:n}) \propto \prod_{l=1}^K \theta_l^{n_l + \alpha_l - 1}$$

and the posterior mean is

$$\mathbb{E}[\theta_l \mid x_{1:n}] = \frac{n_l + \alpha_l}{n + \sum\limits_{j=1}^K \alpha_j}$$

for instance with $\alpha_l = 1$ for all l it adds 1, "smoothing" the maximum likelihood estimator.

$$\mathbb{E}[\theta_l \mid x_{1:n}] = \frac{n_l + 1}{n + K}$$

i. NB One can consider that posterior can be used for prior of next observation. This is the *sequential approach*.

IV. Bayesian linear regression

Let us assume that

$$y = \omega^\top x + \epsilon \tag{4.1}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Then the observation issue

$$p(y \mid x) = \mathcal{N}\left(y \mid \omega^\top x, \sigma^2\right)$$

Then if we also choose a Gaussian prior on ω.

$$p(\omega) = \mathcal{N}\left(\omega; 0, \frac{I_n}{\lambda}\right)$$

then the posterior is also a Gaussian with the following parameters

- covariance: $\hat{\Sigma}_n = \lambda I_n + \frac{X^\top X}{\sigma^2}$
- mean: $\hat{\mu}_n = \hat{\Sigma}_n^{-1}\left(X^\top \vec{y}/\sigma^2\right)$

where

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad \text{and} \quad \vec{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

the covariance and the mean are the same as the ones for the *ridge regression* with $\tilde{\lambda} = \lambda\sigma^2$.

As a Bayesian: compute predictive distribution

$$\begin{aligned} p(y_{new} \mid x_{new}, x_{1:n}, y_{1:n}) &= \int_\omega p(y_{new} \mid x_{new}, \omega) p(\omega \mid data) d\omega \\ &= \mathcal{N}\left(y_{new} \mid \hat{\mu}_n^\top x_{new}, \sigma^2_{predictive}\right) \end{aligned}$$

where

$$\sigma^2_{predictive}(x_{new}) = \sigma^2 + x_{new}^\top \hat{\Sigma}_n x_{new},$$

the real number σ comes from the noise model and the second quantity of the right hand side comes from the posterior covariance.

CHAPTER 5

Directed and undirected graphical models

You can find a review on probabilities, including independence and conditional independence in Annex I. , and a review on graphs in Annex V. .

In this lecture, all random variables are assumed to be discrete, in order to keep notations as simple as possible. All the theory presented generalizes immediately to continuous random variables that have a density by replacing:

- the discrete probability distributions considered in this lecture by densities,
- summations by integration w.r.t. a reference measure (most of the time the LEBESGUE measure).

Graphical models combine probability and graph theory into an efficient data structure. We want to be able to handle probabilistic models of hundreds of variables. For example, assume we are trying to model the probability of diseases given the symptoms, as shown below:

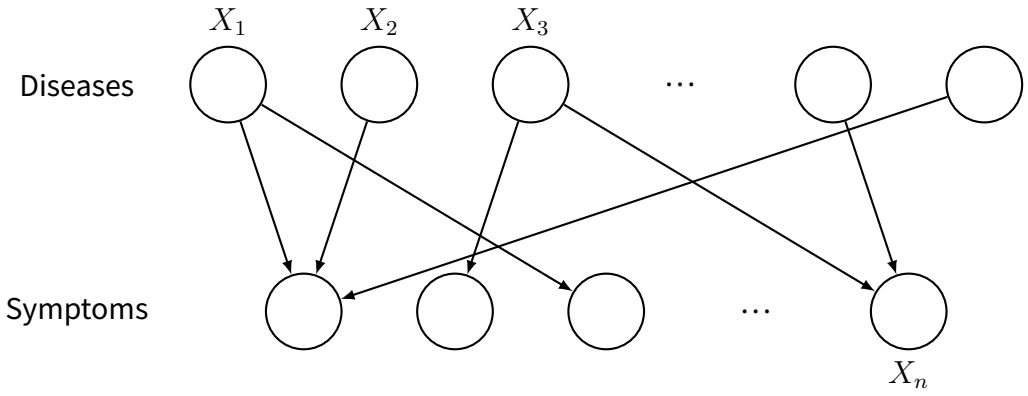


Figure 5.1: Graph representing binary variables which indicate the presence or not of a disease or symptom

In this example we consider n nodes, each associated to a binary variable $X_i \in \{0, 1\}$, indicating the presence or absence of a disease or a symptom. The number of joint probability terms would grow exponentially. For 100 diseases and symptoms, we would need a table of size 2^{100} to store all the possible states. This is clearly intractable. Instead, we will use graphical models to represent the relationships between nodes.

I. Directed Graphical Model

Let $G = (V, E)$ be a graph. A directed graphical model, also historically called a "Bayesian network" when the variables are discrete, represents a *family of distributions* denoted $\mathcal{L}(G)$:

$$\mathcal{L}(G) := \left\{ p \mid \exists (f_i)_{1 \leq i \leq n} \text{ s.t. } \forall x, p(x) = \prod_{i=1}^{n} f_i(x_i, x_{\pi_i}) \right\}$$

where the $(f_i)_{1 \leq i \leq n}$, called *legal factors*, satisfy $f_i \geq 0$ and $\sum_{x_i} f_i(x_i, x_{\pi_i}) = 1$ for all $i \in [\![1, n]\!]$ and x_{π_i}, and we recall that π_i stands for the set of parents of the vertex i in G.

I. A. First definitions and properties

Let $X_1, \dots, X_n$ be n random variables with joint distribution $p(X)$. Let $G = (V, E)$ be a directed acyclic graph, with $V = [n]$.

DEFINITION I. .1. [FACTORISATION IN G]
We say that $p(X)$ factorizes in G if $p(X) \in \mathcal{L}(G)$.

We prove the following useful and fundamental property of directed graphical models:

PROPOSITION I. .1. [LEAF MARGINALIZATION]
Suppose that $p(X)$ factorizes in G. Then for any leaf[a] ℓ, we have

$$p(x_{V \setminus \{\ell\}}) = \prod_{i \neq \ell} f_i(x_i, x_{\pi_i})$$

Hence $p(X_{V \setminus \{\ell\}})$ factorizes in G' the induced graph on $V \setminus \{\ell\}$.

[a] a *leaf* or *terminal node* of a directed acyclic graph is a node that has no descendant

PROOF Without loss of generality, we can assume that the leaf is indexed by n. Since it is a leaf, we clearly have that $n \notin \pi_i$ for all $i \in [\![1, n-1]\!]$. We have the following computation:

$$\begin{aligned} p(x_1, \dots, x_{n-1}) &= \sum_{x_n} p(x_1, \dots, x_n) = \sum_{x_n} \prod_{i=1}^{n-1} f_i(x_i, x_{\pi_i}) f_n(x_n, x_{\pi_n}) \\ &= \prod_{i=1}^{n-1} f_i(x_i, x_{\pi_i}) \sum_{x_n} f_n(x_n, x_{\pi_n}) = \prod_{i=1}^{n-1} f_i(x_i, x_{\pi_i}) \end{aligned}$$

□

Note that the new graph G' obtained by removing a leaf is still a directed acyclic graph. Indeed, since we only removed edges and nodes, if there was a cycle in G', the same cycle would be present in G, which is not possible since G it is directed acyclic graph.

REMARK I. .1. By induction this result shows that in the definition of factorization we do not need to assume that p is a probability distribution. Indeed, if any function p satisfies the factorisation property then it is a probability distribution, because it is non-negative as a product of non-negative factors and it sums to 1 by using formula proved by induction.

LEMMA I. .2. *Let A, B, C be three sets of nodes such that $C \subset B$ and $A \cap B = \varnothing$. If $p(x_A \mid x_B)$ only depends on (x_A, x_C) then $p(x_A \mid x_B) = p(x_A \mid x_C)$.*

PROOF Let $p(x_A \mid x_B) = f(x_A, x_C)$ for some function f. Then $p(x_A, x_B) = p(x_A \mid x_B)p(x_B) = f(x_A, x_C)p(x_B)$. By summing over $x_{B\setminus C}$, we have:

$$p(x_A, x_C) = \sum_{x_{B\setminus C}} p(x_A, x_B) = f(x_A, x_C) \sum_{x_{B\setminus C}} p(x_B) = f(x_A, x_C)p(x_C)$$

which proves that $p(x_A \mid x_C) = f(x_A, x_C) = p(x_A \mid x_B)$. □

Now we try to characterize the factor functions. The following result will imply that if p factorizes in G, then we have a uniqueness of the factors.

PROPOSITION I. .3. *If $p(X) \in \mathcal{L}(G)$ then $f_i(x_i, x_{\pi_i}) = p(x_i \mid x_{\pi_i})$ for all $i \in [\![1, n]\!]$.*

PROOF Assume, without loss of generality, that the nodes are sorted in a topological order[1]. Consider a node $i \in V$. Since the nodes are in topological order, we can apply the leaf marginalization $n - i$ times to obtain that

$$p(x_1, \dots, x_i) = \prod_{j \leq i} f(x_j, x_{\pi_j})$$

Since we also have $p(x_1, \dots, x_{i-1}) = \prod_{j<i} f(x_j, x_{\pi_j})$, we have taking the ratio:

$$p(x_i \mid x_1, \dots, x_{i-1}) = f(x_i, x_{\pi_i})$$

Since $\pi_i \subset [\![1, i-1]\!]$, this entails by the previous lemma that

$$p(x_i \mid x_1, \dots, x_{i-1}) = p(x_i \mid x_{\pi_i}) = f(x_i, x_{\pi_i})$$

□

Hence we can give an equivalent definition the factorization over a directed acyclic graph:

PROPOSITION I. .4. *$p(X)$ factorizes in G if and only if*

$$\forall x, \quad p(x) = \prod_{i=1}^{n} p(x_i \mid x_{\pi_i})$$

EXAMPLE I. .1.

- **[TRIVIAL GRAPHS]** Assume $E = \varnothing$, i.e. G has no edges. We then have $p(x) = \prod_{i=1}^{n} p(x_i)$, implying the random variables $X_1, \dots, X_n$ are independent. Thus variables are mutually independent if they factorize in the empty graph.
- **[COMPLETE GRAPHS]** Assume now we have a complete graph[2], we have: $p(x) = \prod_{i=1}^{n} p(x_i \mid x_1, \dots, x_{i-1})$, the so-called "chain rule" which is always true. Every probability distribution factorizes in a complete graph. Note that there are $n!$ complete graph possible, and that they are all equivalent.

[1] for any $j \in [\![1, n]\!]$, we have $\pi_j \subset [\![1, j-1]\!]$

[2] thus with $n(n-1)/2$ edges as we need acyclicity for it to be a directed acyclic graph

- **[GRAPHS WITH SEVERAL CONNECTED COMPONENTS]** If G has several connected components $C_1, \dots, C_K$, then one can show that $p \in \mathcal{L}(G)$ implies $p(x) = \prod_{k=1}^{K} p(x_{C_k})$ (exercise). As a consequence, each connected component can be treated separately.

In the rest of the lecture, we will therefore focus on connected graphs.

I. B. Graphs with three nodes

In this subsection we consider all connected graphs with 3 nodes, except for the complete graph, which we have already discussed.

- MARKOV chain: the MARKOV chain on 3 nodes is illustrated on Figure 5.2. For this graph we have
$$p(X, Y, Z) \in \mathcal{L}(G) \quad \Longrightarrow \quad (X \perp\!\!\!\perp Y) \mid Z$$
Indeed we have:
$$p(y \mid z, x) = \frac{p(x, y, z)}{p(x, z)} = \frac{p(x, y, z)}{\sum_{y'} p(y', x, z)} = \frac{p(x)p(z \mid x)p(y \mid z)}{\sum_{y'} p(x)p(z \mid x)p(y' \mid z)} = p(y \mid z)$$
thus
$$p(x, y \mid z) = \frac{p(x, y, z)}{p(z)} = \frac{p(x, y, z)}{p(x, z)} \frac{p(x, z)}{p(z)} = p(y \mid x, z)p(x \mid z) = p(y \mid z)p(x \mid z)$$

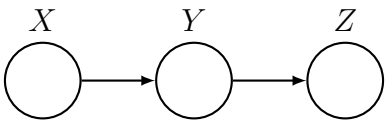


Figure 5.2: Graph of the MARKOV chain on 3 nodes

- Latent cause: it is the type of directed acyclic graph given in Figure 5.3. We show that:
$$p(X, Y, Z) \in \mathcal{L}(G) \quad \Longrightarrow \quad (X \perp\!\!\!\perp Y) \mid Z$$
Indeed:
$$p(x, y \mid z) = \frac{p(x, y, z)}{p(z)} = \frac{p(z)p(y \mid z)p(x \mid z)}{p(z)} = p(x \mid z)p(y \mid z)$$
- Explaining away: represented in Figure 5.4, we can show for this type of graph
$$p(X, Y, Z) \in \mathcal{L}(G) \quad \Longrightarrow \quad X \perp\!\!\!\perp Y$$
It basically stems from:
$$p(x, y) = \sum_z p(x, y, z) = p(x)p(y) \sum_z p(z \mid x, y) = p(x)p(y)$$

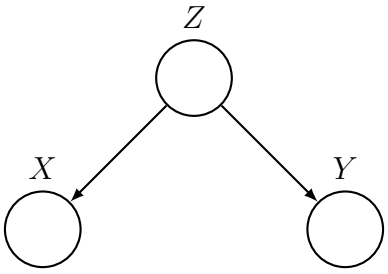


Figure 5.3: Graph of the common latent cause

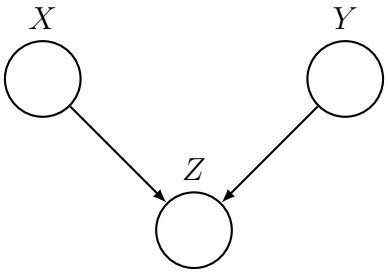


Figure 5.4: Explaining away or v-structure

REMARK I. .2. The word "cause" should here be between quotes and used very carefully, because the same way that correlation is not causation, conditional dependance is not causation either! This is however the historical name for this model. The reason why cause is a bad name, and that *latent factor* might be better, is that the factorisation properties that are encoded by graphical models do not in general correspond to the existence of a causal mechanisms, but only to conditional independence relations.

REMARK I. .3. If p factorizes in the latent cause graph, then $p(x, y, z) = p(z)p(x \mid z)p(y \mid z)$, but using BAYES rule $p(z)p(x \mid z) = p(x)p(z \mid x)$ and so we also have that $p(x, y, z) = p(x)p(z \mid x)p(y \mid z)$ which shows that p is a MARKOV chain, i.e. factorizes in the MARKOV chain graph.

This is an example of *basic edge reversal* that we will discuss in the next section. Note that we proceeded by equivalence, which shows that the MARKOV chain graph, the reversed MARKOV chain graph and the "latent cause" graph are in fact equivalent in the sense that a distribution that factorizes according to one factorizes according to the others. This is what we will call MARKOV equivalence.

REMARK I. .4. In the "explaining away" graph, in general $(X \perp\!\!\!\perp Y) \mid Z$ is not true in the sense that there exist elements in $\mathcal{L}(G)$ such that this statement is violated.

REMARK I. .5. For a fixed graph, $p \in \mathcal{L}(G)$ implies that p satisfies some list of (positive) conditional independence statements (CIS). The fact that $p \in \mathcal{L}(G)$ cannot guarantee that a given CIS does not hold. This should be obvious because the independent distribution belongs to all graphical models and satisfies all CIS ...

It is also important to note that not all lists of CIS correspond to a graph, in the sense that there are lists of CIS for which there exists no graph such that $\mathcal{L}(G)$ is formed exactly of the distributions which satisfy only the conditional independences that are listed or that are consequences of the ones listed. In particular there is no graph G on 3 variables such that $\mathcal{L}(G)$ contains all distributions on (X, Y, Z) that satisfy $X \perp\!\!\!\perp Y$, $Y \perp\!\!\!\perp Z$, $X \perp\!\!\!\perp Z$ and does not contain distributions for which any of these statements is violated[a].

[a] remember that pairwise independence does not imply mutual independence: see Remark I. .3

I. C. Inclusion, reversal and marginalization properties

Inclusion property Here is a quite intuitive proposition about included graphs and their factorization:

PROPOSITION I. .5. *If $G = (V, E)$ and $G' = (V, E')$ then:*

$$E \subset E' \quad \Longrightarrow \quad \mathcal{L}(G) \subset \mathcal{L}(G')$$

PROOF If $p(X) \in \mathcal{L}(G)$, then $p(x) = \prod_{i=1}^{n} p(x_i \mid x_{\pi_i(G)})$. Since $E \subset E'$, it is obvious that $\pi_i(G) \subset \pi_i(G')$, and we can define $f_i(x_i, x_{\pi_i(G')}) = p(x_i \mid x_{\pi_i(G)})$. Then $p(x) = \prod_{i=1}^{n} f_i(x_i, x_{\pi_i(G')})$ and f_i meets the factorization requirements, which proves that $p \in \mathcal{L}(G')$. □

The converse of the previous proposition is not true. In particular, different graphs can define the same set of distributions. We introduce first some new definitions:

DEFINITION I. .2. [MARKOV EQUIVALENCE]
We say that two graphs G and G' are MARKOV equivalent if $\mathcal{L}(G) = \mathcal{L}(G')$.

PROPOSITION I. .6. [BASIC EDGE REVERSAL]
If $G = (V, E)$ is a directed acyclic graph and if for all $(i, j) \in E$, i has no parents and the only parent of j is i, then the graph obtained by reversing the edge (i, j) is MARKOV equivalent to G.

PROOF First note that by reversing such an edge no cycle can be created because the cycle would necessarily contain (j, i) and j has no parent other than i. Using BAYES rule we have

$$p(x_i)p(x_j \mid x_i) = p(x_j)p(x_i \mid x_j)$$

and we convert the factorization w.r.t. G to factorization w.r.t. the graph obtained by edge reversal. □

Informally, the previous result can be reformulated as: an edge reversal that does not remove or creates any v-structure leads to a graph which is MARKOV equivalent.

When applied to the 3-nodes graphs considered earlier, this property proves that the MARKOV chain

and the latent cause graph are equivalent. On the other hand, the fact that the explain away graph has a v-structure is the reason why it is not equivalent to the others.

DEFINITION I. .3. [COVERED EDGE]
An edge (i, j) is said to be covered if $\pi_j = \pi_i \cup \{i\}$.

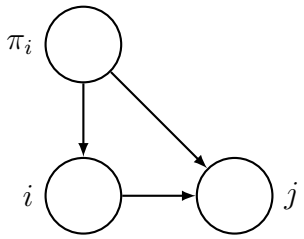


Figure 5.5: Graph where edge (i, j) is covered

PROPOSITION I. .7. [COVERED EDGE REVERSAL]
Let $G = (V, E)$ be a directed acyclic graph and $(i, j) \in E$ a covered edge. Let $G' = (V, E')$ with $E' = (E \setminus \{(i, j)\}) \cup \{(j, i)\}$, then G' is necessarily also a directed acyclic graph and $\mathcal{L}(G) = \mathcal{L}(G')$.

PROOF Exercise. □

Marginalization We have proved in Proposition I. .1 that if $p(x_1, \ldots, x_n)$ factorizes in G, the distribution obtained by marginalizing a leaf i factorizes in the graph G' induced on $V \setminus \{i\}$ by G. A nice property of the obtained graph is that all the conditional independences between variables $X_1, \ldots, X_{n-1}$ that were implied by G are still implied by G': marginalization has lost conditional independences information about X_n but not about the rest of the distribution.

It would be natural to try to generalize this and a legitimate question is: if we marginalise a node i in a distribution of $\mathcal{L}(G)$ is there a simple construction of a graph G' such that the marginalized distribution factorizes in G' and such that all the CIS that hold in G and do not involve X_i are still implied by G'. Unfortunately this is not true. Another less ambitious natural question is then: is there an unique smallest graph G' such that if $p \in \mathcal{L}(G)$ then the distribution obtained by marginalizing i is in $\mathcal{L}(G')$. Unfortunately this is not the case either, as illustrated by the following exemple.

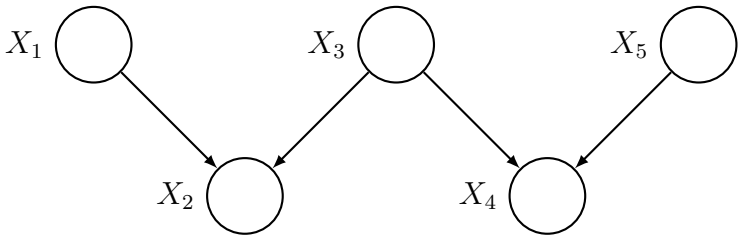


Figure 5.6: Marginalizing X_3 would not result in family of distributions that cannot be exactly represented by a directed graphical model and one can check that there is no unique smallest graph in which the obtained distribution factorizes

Conditional independence with the non-descendents In a MARKOV chain, a well known property is that X_t is independent of the past given X_{t-1}. This result generalizes as follows in a directed graphical model: if $p(X)$ factorizes in G then every single random variable is independent from the set of its non-descendants given its parents.

DEFINITION I. .4. The set of non-descendants of i denoted $\text{nd}(i)$ is the set of nodes that are not descendants of i.

LEMMA I. .8. *For a graph $G = (V, E)$ and a node i, there exists a topological order such that all elements of $\text{nd}(i)$ appear before i.*

PROOF This is easily proved constructively: we construct the topological order in reverse order. At each iteration we remove a node among leaves (of the remaining graph) which we add in the reverse order, and specifically, if some leaves are descendants of i then we remove one of those. If at any iteration there is no leaf that is a descendant of i, it means that all descendants of i have been removed from the graph. Indeed, if there were some descendants of i left in the graph, since all their descendants are descendants of i as well there would exist a leaf node which is a descendant of i. This procedure thus removes all strict descendants of i first, then i and then only all elements of $\text{nd}(i)$. □

With this lemma, we can show our main result:

PROPOSITION I. .9. *If G is a DAG, then:*

$$p(X) \in \mathcal{L}(G) \quad \iff \quad \forall i, \quad (X_i \perp\!\!\!\perp X_{\text{nd}(i)}) \mid X_{\pi_i}$$

PROOF

$\implies$ Based on the previous lemma we can find an order such that $\text{nd}(i) = [\![1, i-1]\!]$. But we have proven in Proposition I. .4 that $p(x_i \mid x_{\pi_i}) = p(x_i \mid x_1, \dots, x_{i-1})$, which given the order chosen is also $p(x_i \mid x_1, \dots, x_{i-1}) = p(x_i \mid x_{\pi_i}, x_{\text{nd}(i)\setminus\pi_i})$, this proves $(X_i \perp\!\!\!\perp X_{\text{nd}(i)\setminus\pi_i}) \mid X_{\pi_i}$, what we wanted to show.

$\impliedby$ Consider a topological order, Then $[\![1, i-1]\!] \subset \text{nd}(i)$: indeed by contradiction, suppose $j \in [\![1, i-1]\!]$ and $j \notin \text{nd}(i)$, then it exists a path from i to j, which contradicts the topological order property as there would be an edge from i to an element of $[\![1, i-1]\!]$.
By the chain rule, we always have $p(x_V) = \prod_{i=1}^{n} p(x_i \mid x_1, \dots x_{i-1})$ but by the conditional independence assumptions $p(x_i \mid x_1, \dots x_{i-1}) = p(x_i \mid x_{\pi_i})$, hence the result by substitution.

□

I. D. d-separation

Given a graph G and A, B, C three subsets of V, it would be useful to be able to answer the question: is $X_A \perp\!\!\!\perp X_B \mid X_C$ true for all $p \in \mathcal{L}(G)$? An answer is provided by the concept of d-separation, or directed separation.

We call a chain a path in the symmetrized graph, i.e. in the graph obtained by ignoring the directionality of the edges.

DEFINITION I. .5. [CHAIN]
Let $a, b \in V$. A chain from a to b is a sequence of nodes, say $(v_1, \dots, v_m)$ such that $v_1 = a$ and $v_m = b$ and for all $i \in [\![0, m-1]\!]$, we have $(v_i, v_{i+1}) \in E$ or $(v_{i+1}, v_i) \in E$.

Assume C is an observed set. We want to define a notion of being "blocked" by this set in order to answer the underlying question above.

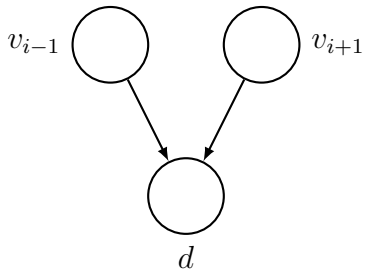


Figure 5.7: d-separation: case $d \in C$ and v-structure

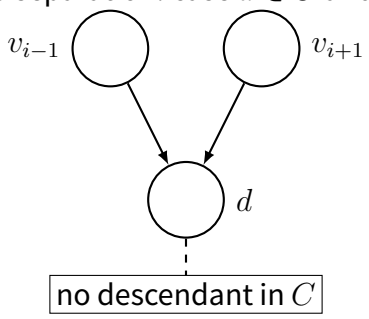


Figure 5.8: d-separation: case $d \notin C$ and v-structure

DEFINITION I. .6. [BLOCKING NODE IN A CHAIN, BLOCKED CHAIN AND d-SEPARATION]

1. A chain from a to b is blocked at $v_i = d$ if:
   - either $d \in C$ and (v_{i-1}, d, v_{i+1}) is not a v-structure,
   - or $d \notin C$ and (v_{i-1}, d, v_{i+1}) is a v-structure and no descendants of d is in C.
2. A chain from a to b is blocked if and only if it is blocked at any node.
3. A and B are said to be d-separated by C if and only if all chains that go from any $a \in A$ to any $b \in B$ are blocked.

EXAMPLE I. .2.

- **[MARKOV CHAIN]** Applying d-separation to the MARKOV chain retrieves the well know results that the future is independent to the past given the present:
- **[HIDDEN MARKOV MODEL]** We can apply it as well to the hidden MARKOV chain graph.

I. E. BAYES ball algorithm

Checking whether 2 nodes are d-separated is not always easy. The BAYES ball algorithm is an intuitive "reachability" algorithm to answer this question. Suppose we want to determine if X is

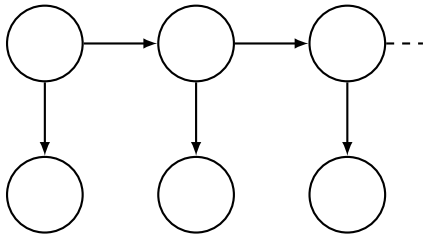


Figure 5.9: Hidden MARKOV Model

conditionally independent from Z, given Y. The principle of the algorithm is to place initially a ball on each of the nodes in X, to then let them bounce around according to some rules described below and to see if any reaches Z. $(X \perp\!\!\!\perp Z) \mid Y$ is true if none reached Z, but not otherwise.

The rules are as follow for the three canonical graph structures. Note that the balls are allowed to travel in either direction along the edges of the graph:

- MARKOV chain: balls pass through when we do not observe Y, but are blocked otherwise.

Figure 5.10: MARKOV chain rule. Left: when Y is observed, balls are blocked. Right: when Y is not observed, balls pass through

- Two children: balls pass through when we do not observe Y, but are blocked otherwise.

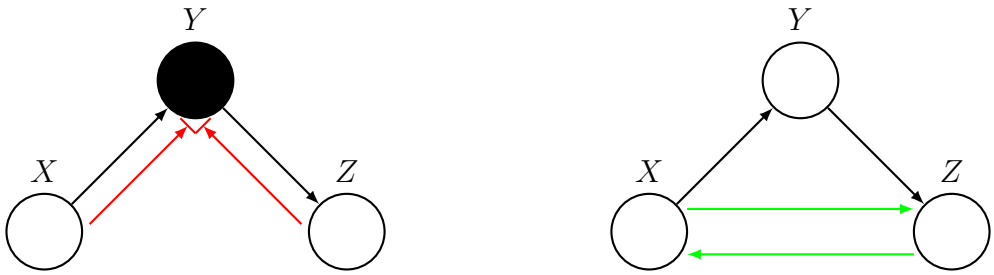


Figure 5.11: Rule when X and Z are Y's children. Left: when Y is observed, balls are blocked. Right: when Y is not observed, balls pass through

- v-structure: balls pass through when we observe Y, but are blocked otherwise.

II. Undirected graphical models

II. A. Definition

Let $G = (V, E)$ be an undirected graph.

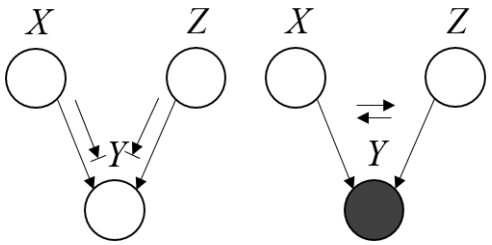


Figure 5.12: v-structure rule. Left: when Y is not observed, balls are blocked. Right: when Y is observed, balls pass through

DEFINITION II. .1. [FACTORIZATION IN AN UNDIRECTED GRAPH]
We denote by $\mathcal{C}$ the set of cliques of G. We say that a probability distribution $p(X)$ factorizes in G and write $p \in \mathcal{L}(G)$ if exists $(\psi_C)_{C\in\mathcal{C}}$ nonnegative functions such that:

$$\forall x, \quad p(x) = \frac{1}{Z}\prod_{C\in\mathcal{C}} \psi_C(x_C) \qquad \text{where } Z = \sum_x \prod_{C\in\mathcal{C}} \psi_C(x_C)$$

The functions $(\psi_C)_{C\in\mathcal{C}}$ are not probability distributions like in the directed graphical models. They are called *potentials*.

REMARK II. .1. With the normalization by Z of this expression, we see that the functions $(\psi_C)_{C\in\mathcal{C}}$ are defined up to a multiplicative constant.

REMARK II. .2. We can restrict $\mathcal{C}$ to $\mathcal{C}_{\max}$ the set of maximal cliques.

REMARK II. .3. This definition can be extended to any function: f is said to factorize in G if and only if $f(x) = \prod_{C\in\mathcal{C}} \psi_C(x_C)$ for all x.

II. B. Trivial graphs

Empty graphs We consider $G = (V, E)$ with $E = \varnothing$. If $p \in \mathcal{L}(G)$, then as $\mathcal{C} = \{(\}\,\{i\} \mid i \in V)$:

$$\forall x, \quad p(x) = \frac{1}{Z}\prod_{i=1}^{n} \psi_i(x_i)$$

thus $X_1, \dots, X_n$ are mutually independent.

Complete graphs We consider $G = (V, E)$ with $E = V \times V$. If $p \in \mathcal{L}(G)$, then as $\mathcal{C} = \{(\}\,\{i\} \mid i \in V)$:

$$\forall x, \quad p(x) = \frac{1}{Z}\psi_V(x_V)$$

This places no constraints on the distribution of $X_1, \dots, X_n$.

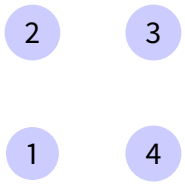


Figure 5.13: The empty graph

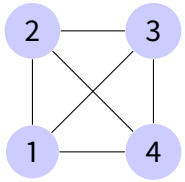


Figure 5.14: The complete graph

II. C. Separation and conditional dependence

PROPOSITION II. .1. *Let $G = (V, E)$ and $G' = (V, E')$ be two undirected graphs. Then:*

$$E \subseteq E' \quad \Longrightarrow \quad \mathcal{L}(G) \subseteq \mathcal{L}(G')$$

PROOF The cliques of G are included in cliques of G'. □

DEFINITION II. .2. [GLOBAL MARKOV PROPERTY]
We say that p satisfies the *global* MARKOV *property* w.r.t. G if and only if for all A, B, S disjoint subsets of V such that A and B are separated by S, then $(X_A \perp\!\!\!\perp X_B) \mid X_S$.

PROPOSITION II. .2. *If $p \in \mathcal{L}(G)$ then p satisfies the global* MARKOV *property w.r.t. G.*

PROOF We suppose without loss of generality that A, B, S are a partition of V, as we could otherwise replace A and B by $A' = A \cup \{a \in V \mid a \text{ and } A \text{ are not separated by } S\}$ and $B' = V \setminus \{S \cup A'\}$, which are also separated by S. Then if we can show that $(X_{A'} \perp\!\!\!\perp X_{B'}) \mid X_S$, then by the decomposition property, we also have that $(X_A \perp\!\!\!\perp X_B) \mid X_S$, giving the required general case.

We consider $C \in \mathcal{C}$. It is not possible to have both $C \cap A \neq \varnothing$ and $C \cap B \neq \varnothing$ as A and B are separated by S and C is a clique. Thus $C \subset A \cup S$ or $C \subset B \cup S$ (or both if $C \subset S$). Let $\mathcal{D}$ be the set of cliques C such that $C \subset A \cup S$ and $\mathcal{D}'$ the set of all other cliques. We have:

$$\forall x, \quad p(x) = \frac{1}{Z} \prod_{C \in \mathcal{D}} \psi_C(x_C) \prod_{C \in \mathcal{D}'} \psi_C(x_C) = f(x_{A \cup S}) g(x_{B \cup S})$$

Thus:

$$p(x_A, x_S) = \frac{1}{Z} f(x_A, x_S) \sum_{x_B} g(x_B, x_S) \quad \Longrightarrow \quad p(x_A \mid x_S) = \frac{f(x_A, x_S)}{\sum_{x'_A} f(x'_A, x_S)}$$

and similarly

$$p(x_B \mid x_S) = \frac{g(x_B, x_S)}{\sum_{x'_B} g(x'_A, x_S)}$$

Hence:

$$p(x_A \,|\, x_S)p(x_B \,|\, x_S) = \frac{\frac{1}{Z}f(x_A, x_S)g(x_B, x_S)}{\frac{1}{Z}\sum_{x'_A} f(x'_A, x_S)\sum_{x'_B} g(x'_A, x_S)} = \frac{p(x_A, x_B, x_S)}{p(x_S)} = p(x_A, x_B \,|\, x_S)$$

i.e. $(X_A \perp\!\!\!\perp X_B) \,|\, X_S$. □

THEOREM II. .3. [HAMMERSLEY-CLIFFORD]
If $p(x) > 0$ for all x, then $p \in \mathcal{L}(G)$ if and only if p satisfies the global MARKOV *property.*

II. D. Marginalization

As for directed graphical models, we also have a marginalization notion in undirected graphs. It is slightly different. If $p(X)$ factorizes in G, then $p(X_1, \dots, X_{n-1})$ factorizes in the graph where the node n is removed and all neighbors are connected:

PROPOSITION II. .4. *Let $G = (V, E)$ be an undirected graph and $G' = (V', E')$ be the graph where n is removed and its neighbors are connected, i.e. . $V' = V \setminus \{n\}$ and E' is obtained from the set E by first connecting together all the neighbours of n and then removing n. If $p(X) \in \mathcal{L}(G)$ then $p(X_1, \dots, X_{n-1}) \in \mathcal{L}(G')$.*

Hence undirected graphical models are closed under marginalization as the construction above is true for any vertex.

We now introduce the notion of MARKOV blanket:

DEFINITION II. .3. [MARKOV BLANKET]
For $i \in V$, the MARKOV *blanket* of G is the smallest set of nodes that makes X_i independent to the rest of the graph.

REMARK II. .4. The MARKOV blanket in an undirected graph for $i \in V$ is the set of its neighbors. For a directed graph, it is the union of all parents, all children and parents of children.

II. E. Relation between directed and undirected graphical models

Since now we have seen that many notions developed for directed graph naturally extended to undirected graphs. The raising question is thus to know whether we can find a theory including both directed and undirected graphs, in particular, is there a way – for instance by symmetrizing the directed graph as we have done repeatedly – to find a general equivalence between those two notions. The answer is no, as we will discuss, though it might work in some special cases described above.

Let G be directed acyclic graph. Can we find G' an undirected graph such that $\mathcal{L}(G) = \mathcal{L}(G')$? $\mathcal{L}(G) \subset \mathcal{L}(G')$?

DEFINITION II. .4. [SYMMETRIZED GRAPH]
The *symmetrized graph* of G is $\tilde{G} = (V, \tilde{E})$, with $\tilde{E} = \{(u, v), (v, u) \,|\, (u, v) \in E\}$, i.e. an edge going the opposite direction is added for every edge in E.

DEFINITION II. .5. [MORALIZED GRAPH]
The *moralized graph* $\overline{G}$ of G is the symmetrized graph where we add edges such that for all $v \in V$, π_v is a clique.

We admit the following proposition:

PROPOSITION II. .5. *Assume G has no v-structure, then $\overline{G} = \tilde{G}$ and $\mathcal{L}(G) = \mathcal{L}(\tilde{G}) = \mathcal{L}(\overline{G})$.*

In case there is a v-structure in the graph, we still have:

PROPOSITION II. .6. *We have $\mathcal{L}(G) \subset \mathcal{L}(\overline{G})$.*

REMARK II. .5. $\overline{G}$ is minimal for the number of edges in the set H of undirected graphs such that $\mathcal{L}(G) \subset \mathcal{L}(H)$

Not all conditional independence structure for random variables can be factorized in a graphical model (directed or undirected).

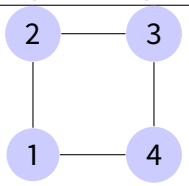
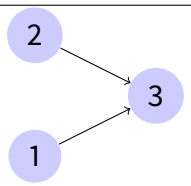
	Directed graphical model	Undirected graphical model
Factorization	$p(x) = \prod_{i=1}^{n} p(x_i \mid x_{\pi_i})$	$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C)$
Set independence	d-separation $(x_i \perp\!\!\!\perp x_{\mathsf{nd}(i)}) \mid x_{\pi_i}$ (and many more)	separation $(X_A \perp\!\!\!\perp X_B) \mid X_S$
Marginalization	not closed in general, only when marginalizing leaf nodes	closed
Difference	grid 	v-structure 

Figure 5.15: Review of the different notions in both the directed and undirected graphical models

CHAPTER 6

Information Theory

DEFINITION . .1. [ENTROPY]
Let X be a random variable taking values in the finite set $\mathcal{X}$, with distribution p. For $x \in \mathcal{X}$, the quantity $I(x) = \log \frac{1}{p(x)}$ is called self-information and the entropy of X is defined as the expectation of $I(X)$:

$$H(X) = \mathbb{E}[I(X)] = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

with convention $0 \log 0 = 0$.

REMARK . .1.

- $I(x)$ can be interpreted as a quantity of information carried by the occurrence of x. $H(X)$ is then the expected amount of information of the random variable X.
- The base of the logarithm is the natural base. We can also use base 2, which can be more consistent with bit coding interpretations of entropy. In this course we will use the natural logarithm, but note that all entropies are proportional.

DEFINITION . .2. [KULLBACK-LEIBLER DIVERGENCE]
Let p, q be two finite distributions on $\mathcal{X}$. The KULLBACK-LEIBLER divergence between p and q is defined by

$$D(p \,||\, q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p\Big[\log \frac{p(X)}{q(X)}\Big] = \mathbb{E}_q\Big[\frac{p(X)}{q(X)} \log \frac{p(X)}{q(X)}\Big]$$

Note that the KULLBACK-LEIBLER divergence is not a distance as it is not symmetric.

PROPOSITION . .1. *We have $D(p \,||\, q) \geq 0$ and equality holds if and only if $p = q$.*

PROOF If there exists $x \in \mathcal{X}$ such that $q(x) = 0$ and $p(x) \neq 0$ then $D(p \,||\, q) = +\infty$. Otherwise, we can without loss of generality assume that $q > 0$ everywhere.

By convexity of the $y \longmapsto y \log y$, we have by JENSEN's inequality:

$$D(p \,||\, q) = \mathbb{E}_q\Big[\frac{p(X)}{q(X)} \log \frac{p(X)}{q(X)}\Big] \geq \mathbb{E}_q\Big[\frac{p(X)}{q(X)}\Big] \log \mathbb{E}_q\Big[\frac{p(X)}{q(X)}\Big] = 0$$

since $\mathbb{E}_q\left[\frac{p(X)}{q(X)}\right] = \sum_{x\in\mathcal{X}} \frac{p(x)}{q(x)} q(x) = \sum_{x\in\mathcal{X}} p(x) = 1$.

Furthermore, by strict convexity $D(p \,||\, q) = 0$ if and only if p/q is constant almost surely. As p and q are two probability distributions, it implies that $p = q$. □

PROPOSITION . .2. *We have the following inequalities:*

(i) $H(X) \geq 0$ *with equality if* X *is constant almost surely,*
(ii) $H(X) \leq \log(\mathrm{card}(\mathcal{X}))$.

PROOF

(i) Since for all $x \in \mathcal{X}$, $p(x) = \mathbb{P}_p(X = x) \leq 1$ then $-p(x)\log p(x) \geq 0$, which implies that $H(X) \geq 0$ with equality if and only if $-p(x)\log p(x) = 0$ for all $x \in \mathcal{X}$, which proves the first point.

(ii) We have for all distribution q

$$0 \leq D(p \,||\, q) = -\Big[\sum_{x\in\mathcal{X}} p(x)\log q(x) - \sum_{x\in\mathcal{X}} p(x)\log p(x)\Big] = -\sum_{x\in\mathcal{X}} p(x)\log q(x) - H(X)$$

Thus $H(X) \leq -\sum_{x\in\mathcal{X}} p(x)\log q(x)$ and taking for q the uniform distribution over $\mathcal{X}$, we obtain

$$H(X) \leq \sum_{x\in\mathcal{X}} p(x)\log(\mathrm{card}(\mathcal{X})) = \log(\mathrm{card}(\mathcal{X}))$$

□

DEFINITION . .3. [MUTUAL INFORMATION]
Let X, Y be two random variables of joint distribution $p_{X,Y}$ and with marginal distributions p_X and p_Y[a]. The mutual information of X and Y is defined by

$$I(X,Y) = D(p_{X,Y} \,||\, p_X p_Y) = \sum_{x,y} p_{X,Y}(x,y)\log\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}$$

[a] we recall that $p_X(x) = \sum_y p_{X,Y}(x,y)$ and $p_Y(y) = \sum_x p_{X,Y}(x,y)$

From Proposition . .1 it directly follows that:

PROPOSITION . .3. $I(X,Y) = 0$ *if and only if* $X \perp\!\!\!\perp Y$.

In general we know that independence implies non correlation but the converse is not true! The first implication comes from the fact that if $X \perp\!\!\!\perp Y$ then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. For the reverse implication, we have the following counter-example: if $\Theta \sim \mathcal{U}([0,1])$ and $X = \sin(2\pi\Theta)$, $Y = \cos(2\pi\Theta)$ then X and Y are not correlated but dependent.

Note that in the case of Gaussian vectors the converse is also satisfied.

Relation between minimum KULLBACK-LEIBLER divergence and maximum likelihood principle
Let $x_1, \ldots, x_n \in \mathcal{X}$ be n i.i.d. observations of X.

DEFINITION . .4. [EMPIRICAL DISTRIBUTION]
The empirical distribution of X derived from the sample $x_1, \dots, x_n$ is the distribution $\hat{p}$ defined by:

$$\forall x \in \mathcal{X}, \quad \hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_{x,x_i}$$

where $\delta_{a,b} = \mathbb{1}_{a=b}$.

Let $\mathcal{P}_\Theta = \{p_\theta \,|\, \theta \in \Theta\}$ be a statistical model.

PROPOSITION . .4. *Maximizing the likelihood $p_\theta(x)$ is equivalent to minimizing the* KULLBACK-LEIBLER *divergence $D(\hat{p} \,||\, p_\theta)$.*

PROOF One has:

$$\begin{aligned} D(\hat{p} \,||\, p_\theta) &= \sum_{x \in \mathcal{X}} \hat{p}(x) \log \frac{\hat{p}(x)}{p_\theta(x)} = -H(\hat{p}) - \sum_{x \in \mathcal{X}} \hat{p}(x) \log p_\theta(x) \\ &= -H(\hat{p}) - \frac{1}{n} \sum_{x \in \mathcal{X}} \sum_{i=1}^{n} \delta_{x,x_i} \log p_\theta(x) = -H(\hat{p}) - \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(x_i) \end{aligned}$$

The second term is equal to the opposite of the $\log$-likelihood. Hence the conclusion. □

REMARK . .2. If $p_\theta(x) = 0$ then $\hat{p}(x) = 0$ but the converse is not true. So we should not try to compute $D(p_\theta \,||\, \hat{p})$ because this would rule out all the values of x that we have not encountered yet (i.e. such that $\hat{p}(x) = 0$).

Maximum entropy principle The maximum entropy principle is a different principle than the maximum likelihood principle and solves a different kind of problem. It assumes that we use the data to specify a constraint on the possible distribution we choose.

The idea is to maximize the entropy $H(p)$ under the constraint that $p \in \mathcal{P}(\mathcal{X})$ a set of possible distribution typically specified from the data.

Let us consider the following examples:

EXAMPLE . .1.

- A study on kangaroos estimated that $p = 3/4$ of the kangaroos are left-handed and $q = 2/3$ drink Foster beer. What is a reasonable estimate of the fraction of kangaroos that are both left-handed and drink Foster beer? The maximum entropy principle can be invoked to choose among all distributions of pairs of binary random variables. In particular, one way to formalize that we want to choose the least specific distribution that satisfies these constraints is to find the distribution with maximal entropy that satisfies the constraints on the marginals.
  If X is the indicator variable of being left-handed and Y the indicator variable of drinking Foster beer, then the problem is formalized as:

$$\begin{array}{ll} \max_{p_{X,Y}} & H(p_{X,Y}) \\ \text{s.t.} & p_{X,Y}(1,0) + p_{X,Y}(1,1) = p \\ & p_{X,Y}(0,1) + p_{X,Y}(1,1) = q \end{array}$$

What is the solution to this problem? (exercise)

- Among all distributions on $[\![1, 10]\!]$ what is the distribution with expected value equal to 2 which has the largest entropy? (exercise)
- It is possible to show that the distribution on $\mathbb{R}$ with fixed mean μ and fixed variance σ^2 that has maximal differential entropy is the Gaussian distribution.
- The principle of maximum entropy is also the principle invoked to construct distribution on angles with fixed mean and variance. It leads to the so-called *wrapped normal distribution*. A related distribution on angle which is also a maximum entropy distribution is the VON MISES distribution.

The maximum entropy principle is used often when working with *contingency tables*.

Entropy and KULLBACK-LEIBLER divergence for continuous random variables Let X be a continuous random variable taking its values in the continuous space $\mathcal{X}$ and let p be its probability density function. We have the following adapted expressions of entropy and KULLBACK-LEIBLER divergence:

DEFINITION . .5. [ENTROPY AND KULLBACK-LEIBLER DIVERGENCE, CONTINUOUS CASE] Let p, q be two probability density functions.

- The differential entropy of p is defined as:

$$H_{\text{diff}}(p) = -\int_{\mathcal{X}} p(x) \log(p(x)) d\mu(x)$$

- The differential KULLBACK-LEIBLER divergence is defined as:

$$D_{\text{diff}}(p \,||\, q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} d\mu(x) = \mathbb{E}_{X \sim p}\Big[\log \frac{p(X)}{q(X)}\Big]$$

In the continuous case, the entropy can be negative!

REMARK . .3. Note that the definition of $H_{\text{diff}}(p)$ depends on the reference measure μ. This means that $H_{\text{diff}}(p)$ does not capture any intrinsic properties of p any more, and loses its "physical interpretation" in terms of quantity of information, at least in an absolute sense. By contrast $D_{\text{diff}}(p \,||\, q)$ does not depend on the choice of the reference measure and has therefore a stronger interpretation.

CHAPTER 7

Exponential families

I. Generalities

Let $x_1, \dots, x_n \in \mathcal{X}$ be n i.i.d. observations of a random variable X.

Definition I. .1. [statistic]
A statistic Φ is just a function of the data $x_1, \dots, x_n \longmapsto \Phi(x_1, ..., x_N)$.

Definition I. .2. [sufficient statistic[a]]
T is a sufficient statistic for a statistical model $\mathcal{P}_\Theta$ if for all $\theta \in \Theta$ there exists some function h and g such that:

$$\forall x, \quad p_\theta(x) = h(x)g(T(x), \theta)$$

[a] statistique exhaustive in french

A sufficient statistic $T(x)$ carries all the information that is relevant to estimate θ from data x using the maximum likelihood principle.

Another way of interpreting what a sufficient statistic is is to take the Bayesian point of view. In Bayesian statistics, the parameter θ is modelled as a random variable and we then have:

$$p(x, \theta) = p(x \mid \theta)p(\theta) = h(x)g(T(x), \theta)p(\theta)$$

which means that $\theta \perp\!\!\!\perp X \mid T(X)$.

Let Θ be an open subset of $\mathbb{R}^d$ and $\mathcal{P}_\Theta$ a family of distributions taking values in a same space $\mathcal{X}$. Let μ be a fixed measure on $\mathcal{X}$.

Definition I. .3. [exponential family]
$\mathcal{P}_\Theta$ is an exponential family if each distribution $p_\theta \in \mathcal{P}_\Theta$ admits a density w.r.t. μ of the form

$$\forall x \in \mathcal{X}, \quad p_\theta(x) = h(x) \exp\Big(b(\theta)^\top \phi(x) - \tilde{A}(\theta)\Big) d\mu(x)$$

where h is the ancillary statistic, $h\mu$ the reference or base measure, ϕ the sufficient statistic, also called feature vector, $\eta = b(\theta)$ the canonical parameter and $\tilde{A}(\theta) = A(\eta)$ the $\log$-partition function.

PROPOSITION I. .1. [EXPRESSION OF THE log-PARTITION FUNCTION]
One has:

$$A(\eta) = \log \int_{\mathcal{X}} h(x) \exp\left(\eta^\top \phi(x)\right) d\mu(x)$$

PROOF It suffices to write that:

$$1 = \int_{\mathcal{X}} p_\theta(x) d\mu(x) = \mathrm{e}^{-A(\eta)} \int_{\mathcal{X}} h(x) \exp\left(\eta^\top \phi(x)\right) d\mu(x)$$

□

DEFINITION I. .4. [CANONICAL EXPONENTIAL FAMILY]
A canonical exponential family is an exponential family which such that $\eta = b(\theta) = \theta$, i.e. :

$$\forall x \in \mathcal{X}, \quad p_\eta(x) = h(x) \exp\left(\eta^\top \phi(x) - A(\eta)\right)$$

DEFINITION I. .5. [DOMAIN]
The domain of an exponential family is defined as $\Omega = \{\eta \in \mathbb{R}^p \,|\, A(\eta) < +\infty\}$.

EXAMPLE I. .1. [MULTINOMIAL MODEL]
Let X be a random variable on $\mathcal{X} = \left\{(0, \dots, 0, 1, 0, \dots, 0) \in \{0,1\}^K\right\}$ following a multinomial distribution of parameter π. Then we have the following density function w.r.t. the counting measure ν:

$$\forall x \in \mathcal{X}, \quad p_\pi(x) = \prod_{k=1}^K \pi_k^{x_k} = \exp\left(\sum_{k=1}^K x_k \log \pi_k\right) = \exp\left(\sum_{k=1}^K x_k \tilde{\eta}_k\right) = \exp(\tilde{\eta}^\top x)$$

where $\tilde{\eta} = (\log \pi_1, \dots, \log \pi_K)^\top$. We are close to identify an exponential family with $h = 1$, $\phi = \mathrm{id}$ and $\eta = \tilde{\eta}$, but we cannot identify $A(\eta)$. Using Proposition I. .1, we have:

$$A(\eta) = \log\left(\sum_{x \in \mathcal{X}} \exp(\eta^\top x)\right) = \log\left(\sum_{k=1}^K \exp(\eta_k)\right)$$

and if the family is an exponential family with $h = 1$ and $\phi = \mathrm{id}$ we can write:

$$p_\pi(x) = \exp\Big(\sum_{k=1}^K x_k \eta_k - A(\eta)\Big) = \exp\Big(\sum_{k=1}^K (\eta_k - A(\eta)) x_k\Big) = \exp\Big(\sum_{k=1}^K \log\Big(\frac{\exp \eta_k}{\sum_{k'=1}^K \exp \eta_{k'}}\Big) x_k\Big)$$

With the above expression we identify η is defined as satisfying $\pi_k = \frac{\exp \eta_k}{\sum_{k'=1}^K \exp \eta_{k'}}$ for all k.

In fact the first expression was showing that we had an exponential family, with $\eta = \tilde{\eta}$ we have $A(\eta) = 0$.

The difference with this new expression is that we now take into account the fact that $\sum_{k=1}^K \pi_k = 1$. This was a hidden constraint on $\tilde{\eta}$. Adding the $A(\eta)$ gives a new expression with no more constraint over the values that η can take.

EXAMPLE I. .2. [GAUSSIAN DISTRIBUTION OVER $\mathbb{R}$]
We have:

$$\forall x \in \mathbb{R}, \quad p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big) = \exp\Big(\frac{-1}{2\sigma^2} x^2 + \frac{\mu}{\sigma^2} x - \Big[\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\Big]\Big)$$

and we recognize an exponential family on the domain: $\{\eta \in \mathbb{R}^2, \eta_2 < 0\}$ where:

$$\phi(x) = (x, x^2)^\top \qquad \eta = \Big(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\Big)^\top \qquad A(\eta) = \frac{1}{2}\log\Big(-\frac{2\pi}{2\eta_2}\Big) - \frac{\eta_1^2}{4\eta_2}$$

EXAMPLE I. .3. Many other common distributions are exponential families. For instance binomial, POISSON ($\mathcal{X} = \mathbb{N}$), DIRICHLET, Gamma and exponential laws.

DEFINITION I. .6. [CURVED EXPONENTIAL FAMILY]
An exponential family is said to be curved if its Jacobian $\Big(\frac{\partial b_j}{\partial \theta_i}(\theta)\Big)_{i,j}$ is not full rank for all $\theta \in \Theta$.

EXAMPLE I. .4. One can check that $\mathcal{P}_\Theta = \{\mathcal{N}(\mu, \mu^2) \,|\, \mu > 0\}$ is a curved exponential family.

II. Link with the graphical models

EXAMPLE II. .1. [ISING MODEL]

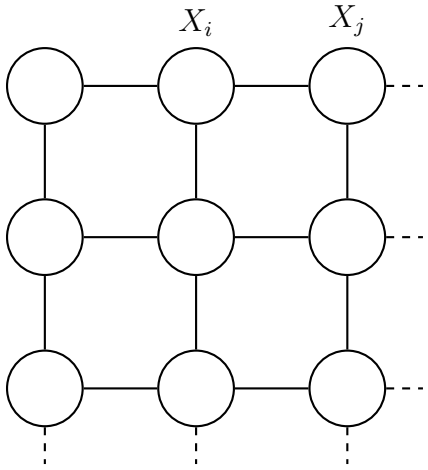


Figure 7.1: ISING model

The ISING model is a model of variables n variables taking values in $\{0, 1\}$ and linked by the graph $G = ([n], E)$ of Figure **??**. A probability distribution of this model is under the following form:

$$\forall x \in \{0,1\}^n, \quad p_\eta(x) = \frac{1}{Z(\eta)} \exp\Big(\sum_{(i,j)\in E} \psi_{ij}^\eta(x_i, x_j)\Big)$$

where $\eta = (V_{ij}^{kk'})_{(i,j)\in E, k,k'\in\{0,1\}}$ and each ψ_{ij} has the following expression[1]:

$$\psi_{ij}(x_i, x_j) = V_{ij}^{11} x_i x_j + V_{ij}^{10} x_i(1-x_j) + V_{ij}^{01}(1-x_i)x_j + V_{ij}^{00}(1-x_i)(1-x_j)$$

[1] we omit the η to avoid heavy notations

It is easy to see that this is an exponential family for which we have

$$\phi(x) = (x_i x_j, x_i(1-x_j), (1-x_i)x_j, (1-x_i)(1-x_j))_{(i,j)\in E}^\top$$

In fact the above expression is overparametrized and we can rewrite the model with just one parameter per node and one per edge under the form:

$$\forall x \in \{0,1\}^n, \quad p_{\tilde\eta}(x) = \frac{1}{Z(\tilde\eta)} \prod_{i\in V} \mathrm{e}^{\tilde\eta_i x_i} \prod_{(i,j)\in E} \mathrm{e}^{\tilde\eta_{ij} x_i x_j}$$

EXAMPLE II. .2. [GENERAL DISCRETE GRAPHICAL MODEL]
In the general case of a discrete graphical model such that $p > 0$ on $\mathcal{X}$, we have:

$$p(x) = \frac{1}{Z}\prod_{c\in\mathcal{C}} \Psi_c(x_c) = \frac{1}{Z}\exp\Big(\sum_{c\in\mathcal{C}} \log \Psi_c(x_c)\Big) = \frac{1}{Z}\exp\Big(\sum_{c\in\mathcal{C}}\sum_{y_c\in\mathcal{X}_c} \delta_{y_c,x_c} \log \Psi_c(y_c)\Big)$$

where $\mathcal{X}_c$ is the set of all possible values of the random variable on the clique c. We identify an exponential family with

$$\phi(x) = (\delta_{x_c,y_c})_{c\in\mathcal{C},y_c\in\mathcal{X}_c} \qquad \text{and} \qquad \eta = (\log \Psi_c(y_c))_{c\in\mathcal{C},y_c\in\mathcal{X}_c}$$

III. Minimal representation

Let $\mathcal{P}_\Theta$ be an exponential family.

REMARK III. .1. The set $\mathcal{N}_\eta = \{x \in \mathcal{X} \mid p_\eta(x) = 0\}$ actually does not depend on η but only on $h(x)$!

In the following we denote by $\mathcal{N} = \{x \mid h(x) = 0\}$ the common set of probability zero.

DEFINITION III. .1. [AFFINELY DEPENDENT STATISTICS]
We denote $\phi(x) = (\phi_1(x), \dots, \phi_K(x))^\top$. The sufficient statistics are said to be affinely dependent if:

$$\exists (c_0, \dots, c_K) \neq 0_{K+1}, \quad \forall x \notin \mathcal{N}, c_0 + \sum_{k=1}^K c_k \phi_k(x) = 0$$

DEFINITION III. .2. [MINIMAL REPRESENTATION OF AN EXPONENTIAL FAMILY]
A vector of sufficient statistics provides a minimal representation of the exponential family these statistics are affinely independent.

THEOREM III. .1. *Every exponential family admits at least one minimal representation (not necessarily unique) of unique minimal dimension K.*

REMARK III. .2. In practice we will quite often use redundant (i.e. not minimal) representations.

IV. Exponential family of an i.i.d. sample

We consider an i.i.d. sample $X_1, \dots, X_n$ of distribution p_η which belongs to an exponential family. Then:

$$p_\eta(x_1, \dots, x_n) = \prod_{i=1}^n p_\eta(x_i) = \prod_{i=1}^n h(x_i) \exp\Big(\eta^\top \phi(x_i) - A(\eta)\Big) = \prod_{i=1}^n h(x_i) \exp\Big(\eta^\top \Big(\sum_{i=1}^n \phi(x_i)\Big) - nA(\eta)\Big)$$

Thus the distribution of our sample belongs to an exponential family with

- sufficient statistics $n\overline{\phi}$ where $\overline{\phi}(x) = \frac{1}{n}\sum_{i=1}^n \phi(x_i)$,
- canonical parameter and domain unchanged,
- log-partition function $nA(\eta)$.

V. Convexity and differentiability in exponential families

We recall the Hölder's inequality:

Lemma V. .1. [Hölder's inequality]
Let $p \in [1, +\infty]$ and μ a measure on $\mathbb{R}^d$. Then for all $f, g : \mathbb{R}^d \longrightarrow \mathbb{R}$, one has:

$$\int |f(x)g(x)|\, d\mu(x) \leq \Big(\int |f(x)|^p\, dx\Big)^{\frac{1}{p}} \Big(\int |g(x)|^q\, dx\Big)^{\frac{1}{q}}$$

where q is such that $\frac{1}{p} + \frac{1}{q} = 1$.

Theorem V. .2. [Convexity properties in an exponential familiy]
In a canonical exponential family, we have the following properties:

(i) The domain Ω is a convex subset of $\mathbb{R}^d$,
(ii) $Z : \eta \longmapsto \int h(x) \exp(\eta^\top \phi(x)) dx$ is a convex function,
(iii) Z is log-convex function, i.e. $A = \log(Z)$ is convex.

Proof

- Let us prove (i) and (ii) together. If $\Omega = \varnothing$, the result is trivial.
  If not, let $\eta = \alpha\eta_1 + (1-\alpha)\eta_2$ where $\eta_1, \eta_2 \in \Omega$ and $\alpha \in [0, 1]$. By convexity we have:
  $$\exp(\eta^\top \phi(x)) \leq \alpha \exp(\eta_1^\top \phi(x)) + (1-\alpha)\exp(\eta_2^\top \phi(x))$$
  Thus:
  $$\int h(x) \exp(\eta^\top \phi(x)) d\mu(x) \leq \alpha \int h(x) \exp(\eta_1^\top \phi(x)) d\mu(x) + (1-\alpha) \int h(x) \exp(\eta_2^\top \phi(x)) d\mu(x)$$
  which is exactly $Z(\eta) \leq \alpha Z(\eta_1) + (1-\alpha) Z(\eta_2)$. Thus Z is convex and as $\eta_1, \eta_2 \in \Omega$, we obtain $Z(\eta) < +\infty$, and thus $\eta \in \Omega$. So Ω is convex.
- To prove (iii), let $\eta = \alpha\eta_1 + (1-\alpha)\eta_2$ where $\eta_1, \eta_2 \in \Omega$ and $\alpha \in [0, 1]$. We can write:
  $$Z(\eta) = \int h(x) \exp(\eta^\top \phi(x)) d\mu(x) = \int \underbrace{[h(x) \exp(\eta_1^\top \phi(x))]^\alpha}_{f(x)} \underbrace{[h(x) \exp(\eta_2^\top \phi(x))]^{1-\alpha}}_{g(x)} d\mu(x)$$

Applying HÖLDER's inequality with $p = 1/\alpha \geq 1$, we obtain:

$$Z(\eta) \leq \Big(\int f(x)^p d\mu(x) \Big)^{\frac{1}{p}} \Big(\int g(x)^q d\mu(x) \Big)^{\frac{1}{q}}$$
$$= \Big(\int h(x) \exp(\eta_1^\top \phi(x))]^\alpha d\mu(x) \Big)^{\alpha} \Big(h(x) \exp(\eta_2^\top \phi(x))]^\alpha d\mu(x) \Big)^{1-\alpha} = Z(\eta_1)^\alpha Z(\eta_2)^{1-\alpha}$$

and Z is $\log$-convex.

□

Thus in a canonical exponential family, the maximum likelihood estimator is the solution of a convex optimization problem! Indeed the $\log$-likelihood is concave:

$$\ell(\eta) = \log p_\eta(x) = \log h(x) + n\eta^\top \overline{\phi}(x) - nA(\eta)$$

The theorem does not hold in any of those two cases: if the family is curve or if ϕ is not fully observed and we consider the marginal likelihood of the observations.

THEOREM V. .3. *If $\eta \in \mathring{\Omega}$, then Z and A are $\mathcal{C}^\infty$ and:*

$$\forall m_1, \dots, m_K \in \mathbb{N}, \quad \frac{\partial^m Z}{\partial \eta_1^{m_1} \dots \partial \eta_K^{m_K}}(\eta) = \mathbb{E}_\eta[\phi_1(x)^{m_1} \dots \phi_K(x)^{m_K}] Z(\eta)$$

PROOF It is a bit technical but standard to show using the dominated convergence theorem that one can exchange differentiation and expectation in the computations of the differentials of Z. One then has for a fixed $k \in [\![1, K]\!]$:

$$\frac{\partial Z}{\partial \eta_k}(\eta) = \int \phi_k(x) h(x) \exp(\eta^\top \phi(x)) d\mu(x)$$
$$= \int \phi_k(x) h(x) \exp(\eta^\top \phi(x) - A(\eta)) d\mu(x) \exp(A(\eta))$$
$$= \mathbb{E}_\eta[\phi_k(x)] Z(\eta)$$

and we obtain the general formula by induction. □

VI. Moment methods

VI. A. Moment vector

DEFINITION VI. .1. [MOMENT VECTOR]
We define the moment vector (or moment parameter) as:

$$\mu(\eta) = \nabla A(\eta) = \mathbb{E}_\eta[\phi(X)]$$

PROOF We have:

$$\nabla A(\eta) = \frac{\int_{\mathcal{X}} h(x)\phi(x) \exp\Big(\eta^\top \phi(x)\Big) d\mu(x)}{\int_{\mathcal{X}} h(x) \exp\Big(\eta^\top \phi(x)\Big) d\mu(x)} = \int_{\mathcal{X}} h(x)\phi(x) \exp\Big(\eta^\top \phi(x) - A(\eta)\Big) d\mu(x) = \mathbb{E}_\eta[\phi(X)]$$

□

EXAMPLE VI. .1.

- For a BERNOULLI distribution, we can write:

$$p(x) = \pi^x(1-\pi)^{1-x} = \exp(x\log\pi - x\log(1-\pi) + \log(1-\pi) = \mathrm{e}^{x\eta - A(\eta)}$$

where $\eta = \log(\frac{\pi}{1-\pi})$ and $A(\eta) = -\log(1-\pi)$.
From this we get that $\pi(1-\pi)\,\mathrm{e}^{\eta}$ and thus $\pi = \sigma(\eta)$ where σ is the sigmoid function[2]. Moreover, we can write $A(\eta) = -\log(1-\pi) = \log(1+\mathrm{e}^{\eta})$ and the moment vector is:

$$\mu(\eta) = \mathbb{E}_\eta[\phi(X)] = \mathbb{E}_\eta[X] = \pi$$

- In the multinomial case we consider $Z \sim \mathcal{M}(1,\pi)$ with $Z \in \{0,1\}^K$. We have $\phi(Z) = Z$ and the moment vector is:

$$\mu(\eta) = \pi$$

- In the Gaussian model we have $\phi(X) = (X, X^2)^\top$ and we obtain:

$$\mu(\eta) = (\mu, \sigma^2 + \mu^2)^\top$$

VI. B. Hessian of A

PROPOSITION VI. .1. *The hessian of A is the covariance matrix of the sufficient statistic:*

$$\nabla^2 A(\eta) = \mathbb{E}[(\phi(X) - \mu(\eta))(\phi(X) - \mu(\eta))^\top] = \mathrm{Cov}(\phi(X))$$

PROOF We can write:

$$\nabla^2 A(\eta) = \nabla\nabla A(\eta) = \nabla\Big(\frac{\nabla Z(\eta)}{Z(\eta)}\Big) = \frac{\nabla^2 Z(\eta)}{Z(\eta)} + \nabla Z(\eta)\Big(-\frac{\nabla Z(\eta)}{Z(\eta)^2}\Big)^\top = \frac{\nabla^2 Z(\eta)}{Z(\eta)} + \frac{\nabla Z(\eta)}{Z(\eta)}\Big(-\frac{\nabla Z(\eta)}{Z(\eta)}\Big)^\top$$

Moreover remark that $\frac{\nabla Z(\eta)}{Z(\eta)} = \mu(\eta)$ and that:

$$(\nabla^2 Z(\eta))_{k,k'} = \mathbb{E}[\phi_k(X)\phi_{k'}(X)]Z(\eta) \qquad \nabla^2 Z(\eta) = \mathbb{E}[\phi(X)\phi(X)^\top]Z(\eta)$$

And consequently:

$$\nabla^2 A(\eta) = \mathbb{E}[\phi(X)\phi(X)^\top] - \mu(\eta)\mu(\eta)^\top = \mathbb{E}[(\phi(X) - \mu(\eta))(\phi(X) - \mu(\eta))^\top] = \mathrm{Cov}(\phi(X))$$

□

COROLLARY VI. .1. *We have the following properties:*

(i) $\nabla^2 A(\eta)$ *is semi-definite positive,*
(ii) A *is convex,*
(iii) A *is strictly convex on* $\mathring{\Omega}$ *if and only if* $\phi(X)$ *is a minimal representation of the exponential family.*

[2] remark that in logistic regression we have $\eta = \mathrm{w}^\top x$

PROOF

(i) One has:

$$\forall c \in \mathbb{R}^K, \quad c^\top \nabla^2 A(\eta) c = \mathbb{E}[c^\top(\phi(X) - \mu(\eta))(\phi(X) - \mu(\eta))^\top c] = \mathbb{E}\Big[\Big\|c^\top(\phi(X)\Big\|^2\Big] \geq 0$$

(ii) It directly follows from (i).

(iii) If A is not strictly convex, then there exists η and $c \in \mathbb{R}^K$ such that $c^\top \nabla^2 A(\eta) c = 0$ therefore, for all $x \in \mathcal{X}$, $\mathrm{Var}(c^\top \phi(x)) = 0$ thus $c^\top \phi(x) = -c_0$ a.s.. We can thus write

$$\forall x \in \mathcal{X}, \quad c_0 + c_1\phi_1(x) + \cdots + c_K\phi_K(x) = 0$$

Since we can go backward, we have the equivalence.

□

VI. C. $\log$-likelihood of an exponential function

With the context and notations of section IV. , we have

$$-\ell(\eta) = -n\eta^\top \overline{\phi} + nA(\eta) \qquad \text{and} \qquad -\nabla\ell(\eta) = -n\overline{\phi} + n\mu(\eta)$$

Consequently we have the following equivalence:

$$\nabla\ell(\eta) = 0 \quad \Longleftrightarrow \quad \mu(\eta) = \overline{\phi}$$

THEOREM VI. .2. [MOMENT MATCHING]
The maximum likelihood estimator η is such that $\phi(x) = \mu(\eta)$.

REMARK VI. .1.

$$\eta \underset{\text{learning}}{\overset{\text{inference}}{\rightleftarrows}} \mu(\eta) = \overline{\phi}$$

VI. D. Link between maximum likelihood and maximum entropy

The maximum entropy principle can be applied: we want to find the distribution p such that $\mathbb{E}_p[\phi(X)] = \overline{\phi}$ and has maximal entropy. We can write this as a convex optimization problem:

$$\begin{array}{ll} \min_p & -H(p) \\ \text{subject to} & \mathbb{E}_p[\phi(X)] = \overline{\phi}, \quad p \geq 0, \quad \sum_x p(x) = 1 \end{array}$$

Let us introduce the Lagrangian of this problem, forgetting in a first time the non-negativity of p:

$$\mathcal{L}(p, \lambda, \nu) = \sum_x p(x)\log p(x) - \lambda\Big(\sum_x p(x)\phi(x) - \overline{\phi}\Big) + \nu\Big(\sum_x p(x) - 1\Big)$$

Since the problem is convex, we have strong duality[3].

Without loss of generality, we can hence assume that $p > 0$ and that the moment condition holds. The gradient of the Lagrangian w.r.t. p is given by:

$$\nabla_p \mathcal{L}(p, \lambda, \nu) = \log p + 1 - \lambda\phi + c$$

and we have:

$$\nabla_p \mathcal{L}(p, \lambda, \nu) = 0 \quad \Longleftrightarrow \quad \forall x, \quad \log p(x) = \lambda\phi(x) - (c+1) \quad \Longleftrightarrow \quad \forall x, \quad p(x) = \mathrm{e}^{-(c+1)}\, \mathrm{e}^{\lambda^\top \phi(x)}$$

We recognize here an exponential family. Reinjecting this value of p and maximizing w.r.t. λ and c, we obtain the maximum likelihood estimator.

We have shown:

THEOREM VI. .3. *If $X_1, \dots, X_n$ is an i.i.d. sample and $\phi(X)$ the sufficient statistic, then the maximum entropy estimator satisfying $\mathbb{E}_p[\phi(X)] = \overline{\phi}$ is the maximum likelihood distribution in the exponential family with sufficient statistic ϕ.*

VI. E. Gaussian graphical models

We consider $X \sim \mathcal{N}(\mu, \Sigma) \in \mathbb{R}^p$.

Canonical parameterization Denoting $\eta = \Sigma^{-1}\mu$ and $\Lambda = \Sigma^{-1}$, we get:

$$\forall x \in \mathbb{R}^p, \quad (x-\mu)^\top \Lambda (x-\mu) = x^\top \Lambda x - 2\eta^\top x + \eta^\top \Sigma \eta$$

from which we deduce:

$$\forall x \in \mathbb{R}^p, \quad p_{\mu,\Lambda}(x) = \exp\left(\eta^\top x - \frac{1}{2} x^\top \Lambda x - A(\eta, \Lambda)\right)$$

where

$$A(\eta, \Lambda) = \frac{1}{2}\eta^\top \Lambda^{-1}\eta + \frac{p}{2}\log(2\pi) - \frac{1}{2}\log(\det(\Lambda))$$

$\theta = (\Lambda, \eta)$ are the canonical parameters. Λ is called the *precision matrix*, and η is the *loading vector*. We have the following sufficient statistic, which is not a minimal representation:

$$\phi(x) = \left(x, -\frac{1}{2} x x^\top\right)^\top$$

The moment of the model is:

$$\nabla_\theta A(\eta, \Lambda) = \mathbb{E}_\theta[\phi(X)] = \left(\mathbb{E}_\theta[X], -\frac{1}{2}\mathbb{E}[XX^\top]\right)$$

[3] SLATER's condition corresponds to the existence of p in the relative interior of the domain of the function that is in $\mathbb{R}_{+*}^{|\mathcal{X}|}$ and such that $\sum_x p(x) = 1$. If we do not find such a p then we can reduce our set taken $\mathcal{X}' = \mathcal{X} \setminus \{x \,|\, p(x) = 0\}$

where:

$$\mathbb{E}_\theta[X] = \nabla_\eta A(\eta, \Lambda) = \Lambda^{-1}\eta = \mu$$

and

$$-\frac{1}{2}\mathbb{E}[XX^\top] = \nabla_\Lambda A(\eta, \Lambda) = -\frac{1}{2}\Lambda^{-1}\eta\eta^\top\Lambda - 1 - \frac{1}{2}\Lambda^{-1} = -\frac{1}{2}[\mu\mu^\top + \Lambda^{-1}]$$

Computing the covariance of X we get:

$$\mathrm{Cov}_\theta(X) = \mathbb{E}_\theta[XX^\top] - \mathbb{E}_\theta[X]\mathbb{E}_\theta[X]^\top = \Lambda^{-1} = \Sigma$$

REMARK VI. .2. We could have also computed the covariance with $\Lambda^{-1} = \nabla^2_\eta A(\eta, \lambda) = \mathrm{Cov}_\theta(XX^\top)$.

Conditioning and marginalization We partition the random variable $X \in \mathbb{R}^p$ into two components $X_1 \in \mathbb{R}^{p_1}$ and $X_2 \in \mathbb{R}^{p_2}$ such that $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ and $p = p_1 + p_2$. We now seek to determine the law of X_1 and $X_2 \mid X_1$

Before doing so, we need to partition the moment parameters μ, Σ and the canonical parameters Λ, η in the same way:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \qquad \eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} \qquad \text{and} \qquad \Lambda = \Sigma^{-1} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}$$

from which we get a partitioned form for the joint distribution:

$$\forall x_1, x_2, \quad p_{\mu,\Sigma}(x_1, x_2) = \frac{1}{\sqrt{(2\pi)^p \det(\Sigma)}} \exp\Big(-\frac{1}{2} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^\top \Lambda \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \Big)$$

In Annex VII. , we introduce a tool to block diagonalize partitioned matrices. This allows us to develop general formulas for marginalization and conditioning in the multivariate Gaussian setting.

Using the WOODBURY-SHERMAN-MORRISON formula, we compute an interesting expression for the quadratic form of the multivariate Gaussian distribution:

$$\begin{aligned}
(x-\mu)^\top\Lambda(x-\mu) &= \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^\top \Sigma^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \\
&= \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^\top \begin{pmatrix} I & -\Sigma_{11}^{-1}\Sigma_{12} \\ 0 & I \end{pmatrix} \begin{pmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & [\Sigma_{/\Sigma_{11}}]^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ -\Sigma_{21}\Sigma_{11}^{-1} & I \end{pmatrix} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \\
&= (x_1 - \mu_1)^\top\Sigma_{11}^{-1}(x_1 - \mu_1) + (x_2 - \mu_2 - b)^\top[\Sigma_{/\Sigma_{11}}]^{-1}(x_2 - \mu_2 - b)
\end{aligned}$$

where we denote $b = \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1)$.

Now recall that $\det(\Sigma) = \det(\Sigma_{11})\det([\Sigma_{/\Sigma_{11}}])$, the joint distribution can be expressed as:

$$p_{\mu,\Sigma}(x_1, x_2) = \underbrace{\frac{1}{\sqrt{(2\pi)^{p_1}\det(\Sigma_{11})}} \mathrm{e}^{-\frac{1}{2}(x_1-\mu_1)^\top\Sigma_{11}^{-1}(x_1-\mu_1)}}_{p(x_1)} \underbrace{\frac{1}{\sqrt{(2\pi)^{p_2}\det([\Sigma_{/\Sigma_{11}}])}} \mathrm{e}^{-\frac{1}{2}(x_2-\mu_2-b)^\top[\Sigma_{/\Sigma_{11}}]^{-1}(x_2-\mu_2-b)}}_{p(x_2 \mid x_1)}$$

from which we deduce that:

$$X_1 \sim \mathcal{N}(\mu_1, \Sigma_{11}) \qquad \text{and} \qquad X_2 \,|\, X_1 \sim \mathcal{N}(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(X_1 - \mu_1), [\Sigma_{/\Sigma_{11}}])$$

Denoting $X_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $X_2 \,|\, X_1 \sim \mathcal{N}(\mu_{2\,|\,1}, \Sigma_{2\,|\,1})$[4], we can then parametrize the canonical parameters:

$$\eta_1 = [\Lambda_{/\Lambda_{22}}]\mu_1 = \eta_2 - \Lambda_{12}\Lambda_{22}^{-1}\eta_2 \qquad \Lambda_1 = \Sigma_{11}^{-1} = [\Lambda_{/\Lambda_{22}}] = \Lambda_{11} - \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}$$
$$\Lambda_{22\,|\,1} = \Lambda_{22} \qquad \eta_{2\,|\,1} = \Lambda_{22\,|\,1}\mu_{2\,|\,1} = \Lambda_{22}\mu_2 - \Lambda_{21}(x_1 - \mu_1) = \eta_2 - \Lambda_{21}x_1$$

We can notice that in the moment parameterization, the marginalization operation is simple and the conditioning is complicated and the opposite holds in the canonical parameterization.

Zeros of the precision matrix and MARKOV properties Let $p(X_1, \ldots, X_p)$ a joint Gaussian distribution. We denote $I = \{i, j\}$ for fixed $i < j$ and we consider $p(x_I \,|\, x_{I^c})$. Using the canonical parameterization:

$$\eta_{I\,|\,I^c} = \begin{pmatrix} \eta_i - \Lambda_{iI^c}x_{I^c} \\ \eta_j - \Lambda_{jI^c}x_{I^c} \end{pmatrix} \qquad \text{and} \qquad \Lambda_{II\,|\,I^c} = \Lambda_{II} = \begin{pmatrix} \Lambda_{ii} & \Lambda_{ij} \\ \Lambda_{ji} & \Lambda_{jj} \end{pmatrix}$$

and we have the following expression for the covariance of $X_I \,|\, X_{I^c}$:

$$\mathrm{Cov}(X_I \,|\, X_{I^c}) = \Sigma_{II\,|\,I^c} = \Lambda_{II\,|\,I^c}^{-1} = \frac{1}{\det(\Lambda_{II})}\begin{pmatrix} \Lambda_{jj} & -\Lambda_{ji} \\ -\Lambda_{ij} & \Lambda_{ii} \end{pmatrix}$$

Hence $\mathrm{Cov}(X_i, X_j \,|\, X_{I^c}) \overset{[?]}{=} -\frac{\Lambda_{ij}}{\sqrt{\Lambda_{ii}\Lambda_{jj}}}$ and $\Lambda_{ij} = 0$ implies that $(X_i \perp\!\!\!\perp X_j) \,|\, X_{I^c}$.

PROPOSITION VI. .4. *The non zero coefficients in Λ correspond to edges in the underlying graphical model.*

Indeed the distribution is proportional to $\exp(\eta^\top x - \frac{1}{2}x^\top \Lambda x) = \prod_{1\le i\le p}\exp(\eta_i x_i)\prod_{1\le i,j\le p}\exp(-\frac{1}{2}x_i\Lambda_{ij}x_j)$.

[4] nota that $\Sigma_{2\,|\,1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$

CHAPTER 8

Sum-product algorithm

I. Motivations

Inference, along with *estimation* and *decoding*, are the three key operations one must be able to perform efficiently in graphical models.

Given a discrete GIBBS model of the form:

$$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

where $\mathcal{C}$ is the set of cliques of the graph, inference enables computation of:

- the marginal $p(x_i)$ for a fixed i or more generally $p(x_C)$,
- the partition function Z,
- the conditional marginal $p(x_i \mid x_j, x_k)$.

and as a consequence computation of:

- the gradient in a exponential family,
- the expected value of the $\log$-likelihood of an exponential family at step E of the EM algorithm (for example for a hidden MARKOV model).

A first example: the ISING model We consider the ISING model and denote by $G = (V, E)$ the associated graph, with $|V| = n$. Let $X = (X_i)_{i \in V}$ be a vector of random variables, taking value in $\{0, 1\}^n$, of which the exponential form of the distribution is:

$$\forall x \in \{0,1\}^n, \quad p(x) = \mathrm{e}^{-A(\eta)} \prod_{i \in V} \mathrm{e}^{\eta_i x_i} \prod_{(i,j) \in E} \mathrm{e}^{\eta_{ij} x_i x_j}$$

The $\log$-likelihood is then:

$$\ell(\eta) = \sum_{i \in V} \eta_i x_i + \sum_{(i,j) \in E} \eta_{ij} x_i x_j - A(\eta)$$

We can therefore take as sufficient statistic:

$$\phi(x) = \left((x_i)_{i \in V}, (x_i x_j)_{(i,j) \in E}\right)^\top$$

But we have seen that for exponential families:

$$\ell(\eta) = \phi(x)^\top \eta - A(\eta) \qquad \text{and} \qquad \nabla_\eta \ell(\eta) = \phi(x) - \mathbb{E}_\eta[\phi(X)]$$

We therefore need to compute $\mathbb{E}_\eta[\phi(X)]$. In our case, we get:

$$\forall i \in V, \quad \mathbb{E}_\eta[X_i] = p_\eta(X_i = 1) \qquad \text{and} \qquad \forall (i,j) \in E, \quad \mathbb{E}_\eta[X_i X_j] = p_\eta(X_i = 1, X_j = 1)$$

This is one of the motivations for solving the problem of inference: in order to be able to compute the gradient of the $\log$-likelihood, we need to know the marginal laws.

Another example: the POTTS model Let $(X_i)_{i \in V}$ be random variables such that X_i takes values in $[\![1, K_i]\!]$. Denoting $\Delta_{ik} = \mathbb{1}_{X_i = k}$ and $\delta_{ik} = \mathbb{1}_{x_i = k}$, the POTTS model is such that

$$p_\eta(\delta) = \exp\Big(\sum_{i \in V} \sum_{k=1}^{K_i} \eta_{ik} \delta_{ik} + \sum_{(i,j) \in E} \sum_{k=1}^{K_i} \sum_{k'=1}^{K_j} \eta_{ijkk'} \delta_{ik} \delta_{jk'} - A(\eta) \Big)$$

This is an exponential family with sufficient statistic $\phi(\delta) = ((\delta_{ik})_{i,k}, (\delta_{ik}\delta_{jk'})_{i,j,k,k'})^\top$. Then we have:

$$\mathbb{E}_\eta[\Delta_{ik}] = p_\eta[X_i = k] \qquad \text{and} \qquad \mathbb{E}_\eta[\Delta_{ik}\Delta_{jk'}] = p_\eta[X_i = k, X_j = k']$$

Those two examples illustrate the need to perform inference.

But a main problem is that in general the inference problem is NP-hard!

How to deal with this problem depends on the kind of graphs:

- for trees, the inference problem is efficient as it is linear in n,
- for "tree-like" graphs, we use the *Junction Tree Algorithm* which enables us to bring the situation back to that of a tree,
- for the general case, we are forced to carry out approximative inference.

II. Inference on a chain

We consider the following graphical model where $(X_i)_{i \in V}$ are random variables with $V = [n]$ taking values in $[\![1, K]\!]$, with joint distribution defined as:

$$p(x) = \frac{1}{Z} \prod_{i=1}^{n} \psi_i(x_i) \prod_{i=2}^{n} \psi_{i-1,i}(x_{i-1}, x_i)$$

We wish to compute $p(x_j)$ for a certain $j \in [\![1, n]\!]$. The naive solution would be to compute the marginal with the greedy formula

$$p(x_j) = \sum_{x_{V \setminus \{j\}}} p(x_1, \dots, x_n)$$

Unfortunately, this type of calculation is of complexity $\mathcal{O}(K^n)$.

We therefore develop the expression[1]:

$$\begin{aligned}
p(x_j) &= \frac{1}{Z} \sum_{x_{V\setminus\{j\}}} \prod_{i=1}^{n} \psi_i(x_i) \prod_{i=2}^{n} \psi_{i-1,i}(x_{i-1}, x_i) \\
&= \frac{1}{Z} \sum_{x_{V\setminus\{j\}}} \prod_{i=1}^{n-1} \psi_i(x_i) \prod_{i=2}^{n-1} \psi_{i-1,i}(x_{i-1}, x_i) \psi_n(x_n) \psi_{n-1,n}(x_{n-1}, x_n) \\
&= \frac{1}{Z} \sum_{x_{V\setminus\{j,n\}}} \sum_{x_n} \prod_{i=1}^{n-1} \psi_i(x_i) \prod_{i=2}^{n-1} \psi_{i-1,i}(x_{i-1}, x_i) \psi_n(x_n) \psi_{n-1,n}(x_{n-1}, x_n) \\
&= \frac{1}{Z} \sum_{x_{V\setminus\{j,n\}}} \prod_{i=1}^{n-1} \psi_i(x_i) \prod_{i=2}^{n-1} \psi_{i-1,i}(x_{i-1}, x_i) \sum_{x_n} \psi_n(x_n) \psi_{n-1,n}(x_{n-1}, x_n)
\end{aligned}$$

which allows us to bring out the messaged passed by node n to node $n-1$:

$$\mu_{n\to n-1}(x_{n-1}) = \sum_{x_n} \psi_n(x_n) \psi_{n-1,n}(x_{n-1}, x_n)$$

When continuing, we obtain by induction:

$$p(x_j) = \frac{1}{Z} \sum_{x_1,\dots,x_{j-1}} \prod_{i=1}^{j} \psi_i(x_i) \prod_{i=2}^{j} \psi_{i-1,i}(x_{i-1}, x_i) \mu_{j+1\to j}(x_j)$$

where we have the following definitions of the descending messages, for $i \in [\![1, n-1]\!]$

$$\mu_{i+1\to i}(x_i) = \sum_{x_{i+1}} \psi_{i+1}(x_{i+1}) \psi_{i,i+1}(x_i, x_{i+1}) \mu_{i+2\to i+1}(x_{i+1})$$

with convention $\mu_{n+1\to n} = 1$.

We can do the same method in an ascending way, by defining the messages for $i \in [\![1, n-1]\!]$

$$\mu_{i\to i+1}(x_{i+1}) = \sum_{x_i} \psi_i(x_i) \psi_{i,i+1}(x_i, x_{i+1}) \mu_{i-1\to i}(x_i)$$

with convention $\mu_{0\to 1} = 1$. Finally we obtain:

$$p(x_j) = \frac{1}{Z} \mu_{j-1\to j}(x_j) \psi_j(x_j) \mu_{j+1\to j}(x_j)$$

Each of the messages is computed with complexity $\mathcal{O}((n-1)K^2)$, and with those $2(n-1)$ messages calculated, one can easily compute $p(x_j)$ for all j and x_j. We also obtain Z by summing:

$$Z = \sum_{x_j} p(x_j) = \sum_{x_j} \mu_{j-1\to j}(x_j) \psi_j(x_j) \mu_{j+1\to j}(x_j)$$

[1] if $j < n$

III. Inference in undirected trees

We consider the following general joint probability:

$$p(x) = \frac{1}{Z} \prod_{i \in V} \psi_i(x_i) \prod_{(i,j) \in E} \psi_{i,j}(x_i, x_j)$$

We note i the vertice for which we want to compute the marginal law $p(x_i)$. We set i to be the root of our tree. For all $j \in V$, we note $\mathcal{C}_j$ and $\mathcal{D}_j$ respectively the set of children and the set of descendants of j.

For a tree with at least two vertices, we define by recurence if $j \in \mathcal{C}_i$:

$$F(x_i, x_j, x_{\mathcal{D}_j}) = \psi_{i,j}(x_i, x_j)\psi_j(x_j) \prod_{k \in \mathcal{C}_j} F(x_j, x_k, x_{\mathcal{D}_k})$$

where $F(x_j, x_k, x_{\mathcal{D}_k}) = 1$ if $\mathcal{D}_k = \varnothing$.

Then by reformulating the marginal:

$$\begin{aligned}
p(x_i) &= \frac{1}{Z} \sum_{x_{V \setminus \{i\}}} \psi_i(x_i) \prod_{j \in \mathcal{C}_i} F(x_i, x_j, x_{\mathcal{D}_j}) \\
&= \frac{1}{Z} \psi_i(x_i) \prod_{j \in \mathcal{C}_i} \sum_{x_j, x_{\mathcal{D}_j}} F(x_i, x_j, x_{\mathcal{D}_j}) \\
&= \frac{1}{Z} \psi_i(x_i) \prod_{j \in \mathcal{C}_i} \sum_{x_j, x_{\mathcal{D}_j}} \psi_{i,j}(x_i, x_j)\psi_j(x_j) \prod_{k \in \mathcal{C}_j} F(x_j, x_k, x_{\mathcal{D}_k}) \\
&= \frac{1}{Z} \psi_i(x_i) \prod_{j \in \mathcal{C}_i} \sum_{x_j} \psi_{i,j}(x_i, x_j)\psi_j(x_j) \prod_{k \in \mathcal{C}_j} \underbrace{\sum_{x_k, x_{\mathcal{D}_k}} F(x_j, x_k, x_{\mathcal{D}_k})}_{\mu_{k \to j}(x_j)} \\
&= \frac{1}{Z} \psi_i(x_i) \prod_{j \in \mathcal{C}_i} \underbrace{\sum_{x_j} \psi_{i,j}(x_i, x_j)\psi_j(x_j) \prod_{k \in \mathcal{C}_j} \mu_{k \to j}(x_j)}_{\mu_{j \to i}(x_i)}
\end{aligned}$$

which leads us to the recurrence relation for the Sum Product Algorithm (SPA): if $j \in \mathcal{C}_i$

$$\mu_{j \to i}(x_i) = \sum_{x_j} \psi_{i,j}(x_i, x_j)\psi_j(x_j) \prod_{k \in \mathcal{C}_j} \mu_{k \to j}(x_j)$$

IV. Sum Product Algorithm (SPA)

IV. A. Sequential SPA for a rooted tree

We deduce the following algorithm:

Algorithm 4: Sum Product Algorithm for a rooted tree

Input : $G, (\psi_i)_i, (\psi_{i,j})_{i,j}$, root i, x_i
Output: $p(x_i)$

for *all leaf* ℓ **do**
  Send message $\mu_{\ell \to \pi_\ell}(x_{\pi_\ell}) = \sum_{x_\ell} \psi_\ell(x_\ell)\psi_{\ell,\pi_\ell}(x_\ell, x_{\pi_\ell})$ for all x_{π_ℓ}
end
while i *did not receive all messages from its children* **do**
  for *all node* $k \in [n]$ *such that* k *has received all messages from its children* **do**
    Send message $\mu_{k \to \pi_k}(x_{\pi_k})$ for all x_{π_k}
  end
end
Compute $p(x_i) = \frac{1}{Z}\psi_i(x_i) \prod_{j \in \mathcal{C}_i} \mu_{j \to i}(x_i)$

This algorithm only enables us to compute $p(x_i)$ at the root. To be able to compute all the marginals (as well as the conditional marginals), one must not only *collect* all the messages from the leafs to the root, but then also *distribute* them back to the leafs. In fact, the algorithm can then be written independently from the choice of a root.

IV. B. SPA for an undirected tree

The case of undirected trees is slightly different:

Algorithm 5: Sum Product Algorithm for an undirected tree

Input : $G, (\psi_i)_i, (\psi_{i,j})_{i,j}$, node i_0, x_{i_0}
Output: $p(x_{i_0})$

for *all leaf* ℓ **do**
  Send message $\mu_{\ell \to \pi_\ell}(x_{\pi_\ell})$ for all x_{π_ℓ}
end
while *at least one edge has not been used to transmit a message* **do**
  for *all node* $j \in [n]$ *such that* j *has not send a message to one of its neighbors, say* i*, and has received messages from all its other neighbors* **do**
    Send $\mu_{j \to i}(x_i) = \sum_{x_j} \psi_{i,j}(x_i, x_j)\psi_j(x_j) \prod_{k \in \mathcal{N}(j) \setminus \{i\}} \mu_{k \to j}(x_j)$ for all x_i
  end
end
Compute $p(x_{i_0}) =$ [???]

IV. C. Parallel SPA (flooding) [todo]

1. Initialise the messages randomly
2. At each step, each node sends a new message to each of its neighbours, using the messages received at the previous step.

IV. D. Marginal laws

Once all messages have been passed, we can easily calculate all the marginal laws

$$\forall\, i \in V,\; p(x_i) = \frac{1}{Z}\, \psi_i(x_i) \prod_{k \in \mathcal{N}(i)} \mu_{k \to i}(x_i) \tag{8.1}$$

$$\forall\, (i,j) \in E,\; p(x_i, x_j) = \frac{1}{Z}\, \psi_i(x_i)\, \psi_j(x_j)\, \psi_{j,i}(x_i, x_i) \prod_{k \in \mathcal{N}(i) \setminus j} \mu_{k \to i}(x_i) \prod_{k \in \mathcal{N}(j) \setminus i} \mu_{k \to j}(x_j) \tag{8.2}$$

IV. E. Conditional probabilities

We can use a clever notation to calculate the conditional probabilities. Suppose that we want to compute

$$p(x_i \,|\, x_5 = 3, x_{10} = 2) \propto p(x_i, x_5 = 3, x_{10} = 2)$$

We can set

$$\tilde{\psi}_5(x_5) = \psi_5(x_5)\, \delta(x_5, 3)$$

Generally speaking, if we observe $X_j = x_{j0}$ for $j \in J_{\text{obs}}$, we can define the modified potentials:

$$\tilde{\psi}_j(x_j) = \psi_j(x_j)\, \delta(x_j, x_{j0})$$

such that

$$p(x \,|\, X_{\text{Jobs}} = x_{\text{Jobs0}}) = \frac{1}{\tilde{Z}} \prod_{i \in V} \tilde{\psi}_i(x_i) \prod_{(i,j) \in E} \psi_{i,j}(x_i, x_j) \tag{8.3}$$

Indeed we have

$$p(x \,|\, X_{\text{Jobs}} = x_{\text{Jobs0}})\, p(X_{\text{Jobs}} = x_{\text{Jobs0}}) = p(x) \prod_{j \in J_{\text{obs}}} \delta(x_j, x_{j0}) \tag{8.4}$$

so that by dividing the equality by $p(X_{\text{Jobs}} = x_{\text{Jobs0}})$ we obtain the previous equation with $\tilde{Z} = Zp(X_{\text{Jobs}} = x_{\text{Jobs0}})$.

We then simply apply the SPA to these new potentials to compute the marginal laws $p(x_i \,|\, X_{\text{Jobs}} = x_{\text{Jobs0}})$

V. Remarks

- The SPA is also called *belief propagation* or *message passing*. On trees, it is an exact inference algorithm.
- If G is not a tree, the algorithm doesn't converge in general to the right marginal laws, but sometimes gives reasonable approximations. We then refer to "Loopy belief propagation", which is still often used in real life.
- The only property that we have used to construct the algorithm is the fact that $(\mathbb{R}, +, \times)$ is a semi-ring. It is interesting to notice that the same can therefore also be done with $(\mathbb{R}_+, \max, \times)$ and $(\mathbb{R}, \max, +)$.

  Example For $(\mathbb{R}_+, \max, \times)$ we define the Max-Product algorithm, also called "Viterbi algorithm" which enables us to solve the *decoding* problem, namely to compute the most probable configuartion of the variables, given fixed parameters, thanks to the messages

$$\mu_{j\to i}(x_i) = \max_{x_j}\left[\psi_{i,j}(x_i,x_j)\psi_j(x_j)\prod_{x_k}\mu_{k\to j}(x_j)\right] \tag{8.5}$$

  If we run the Max-Product algorithm with respect to a chosen root, the *collection* phase of the messages to the root enables us to compute the maximal probability over all configurations, and if at each calculation of a message we have also kept the argmax, we can perform a *distribution* phase, which instead of propagating the messages, will consist of recursively calculating one of the configurations which will reach the maximum.
- In practice, we may be working on such small values that the computer will return errors. For instance, for k binary variables, the joint law $p(x_1, x_2 ... x_n) = \dfrac{1}{2^n}$ can take infinitesimal values for a large k. The solution is to work with logarithms: if $p = \sum_i p_i$, by setting $a_i = \log(p_i)$ we have:

$$\begin{aligned}\log(p) &= \log\left[\sum_i e^{a_i}\right] \\ \log(p) &= a_i^* + \log\left[\sum_i e^{(a_i - a_i^*)}\right]\end{aligned} \tag{8.6}$$

  With $a_i^* = \max_i a_i$. Using logarithms ensures a numerical stability.

VI. Proof of the algorithm

We are going to prove that the SPA is correct by recurrence. In the case of two nodes, we have:

$$p(x_1, x_2) = \frac{1}{Z}\psi_1(x_i)\psi_2(x_2)\psi_{1,2}(x_1, x_2)$$

We marginalize, and we obtain

$$p(x_1) = \frac{1}{Z}\psi_1(x_1)\underbrace{\sum_{x_2}\psi_{1,2}(x_1,x_2)\psi_2(x_2)}_{\mu_{2\to 1}(x_1)}$$

We can hence deduct

$$p(x_1) = \frac{1}{Z}\psi_1(x_1)\mu_{2\to 1}(x_1)$$

And

$$p(x_2) = \frac{1}{Z}\psi_2(x_2)\mu_{1\to 2}(x_2)$$

We assume that the result is true for trees of size $n-1$, and we consider a tree of size n. Without loss of generality, we can assume that the nodes are numbered, so that the $n-$th be a leaf, and we will call π_n its parent (which is unique, the graph being a tree). The first message to be passed is:

$$\mu_{n\to\pi_n}(x_{\pi_n}) = \sum_{x_n} \psi_n(x_n)\psi_{n,\pi_n}(x_n, x_{\pi_n}) \tag{8.7}$$

And the last message to be passed is:

$$\mu_{\pi_n\to n}(x_n) = \sum_{x_{\pi_n}} \psi_{\pi_n}(x_{\pi_n})\psi_{n,\pi_n}(x_n, x_{\pi_n}) \prod_{k\in\mathcal{N}(\pi_n)\setminus\{n\}} \mu_{k\to\pi_n}(x_{\pi_n}) \tag{8.8}$$

We are going to construct a tree $\tilde{T}$ of size $n-1$, as well as a family of potentials, such that the $2(n-2)$ messages passed in T (i.e. all the messages except for the first and the last) be equal to the $2(n-2)$ messages passed in $\tilde{T}$. We define the tree and the potentials as follows:

- $\tilde{T} = (\tilde{V}, \tilde{E})$ with $\tilde{V} = \{1, \dots, n-1\}$ and $\tilde{E} = E \setminus \{n, \pi_n\}$ (i.e., it is the subtree corresponding to the $n-1$ first vertices).
- The potentials are all the same as those of T, except for the potential

$$\tilde{\psi}_{\pi_n}(x_{\pi_n}) = \psi_{\pi_n}(x_{\pi_n})\mu_{n\to\pi_n}(x_{\pi_n}) \tag{8.9}$$

- The root is unchanged, and the topological order is also kept.

We then obtain two important properties:

1) The product of the potentials of the tree of size $n-1$ is equal to:

$$\begin{aligned}
\tilde{p}(x_1, \dots, x_{n-1}) &= \frac{1}{Z} \prod_{i\neq n,\pi_n} \psi_i(x_i) \prod_{(i,j)\in E\setminus\{n,\pi_n\}} \psi_{i,j}(x_i, x_j)\tilde{\psi}_{\pi_n}(x_{\pi_n}) \\
&= \frac{1}{Z} \prod_{i\neq n,\pi_n} \psi_i(x_i) \prod_{(i,j)\in E\setminus\{n,\pi_n\}} \psi_{i,j}(x_i, x_j) \sum_{x_n} \psi_n(x_n)\psi_{\pi_n}(x_{\pi_n})\psi_{n,\pi_n}(x_n, x_{\pi_n}) \\
&= \sum_{x_n} \frac{1}{Z} \prod_{i=1}^{n} \psi_i(x_i) \prod_{(i,j)\in E} \psi_{i,j}(x_i, x_j) \\
&= \sum_{x_n} p(x_1, \dots, x_{n-1}, x_n)
\end{aligned}$$

which shows that these new potentials define on $(X_1, \dots, X_{n-1})$ exactly the distribution induced by p when marginalizing X_n.

2) All of the messages passed in $\tilde{T}$ correspond to the messages passed in T (except for the first and the last).

Now, with the recurrence hypothesis that the SPA is true for trees of size $n-1$, we are going to show that it is true for trees of size n. For nodes $i \neq n, \pi_n$, the result is obvious, as all messages passed are the same:

$$\forall\, i \in V \setminus \{n, \pi_n\},\ p(x_i) = \frac{1}{Z}\,\psi_i(x_i) \prod_{k \in \mathcal{N}(i)} \mu_{k \to i}(x_i) \tag{8.10}$$

For the case $i = \pi_n$, we deduct:

$$\begin{aligned}
p(x_{\pi_n}) &= \frac{1}{Z}\tilde{\psi}_{\pi_n}(x_{\pi_n}) \prod_{k \in \tilde{\mathcal{N}}(\pi_n)} \mu_{k \to \pi_n}(x_{\pi_n}) \quad \text{(product over the neighbours of } \pi_n \text{ in } \tilde{T}) \\
&= \frac{1}{Z}\tilde{\psi}_{\pi_n}(x_{\pi_n}) \prod_{k \in \mathcal{N}(\pi_n) \setminus \{n\}} \mu_{k \to \pi_n}(x_{\pi_n}) \\
&= \frac{1}{Z}\psi_{\pi_n}(x_{\pi_n})\mu_{n \to \pi_n}(x_{\pi_n}) \prod_{k \in \mathcal{N}(\pi_n) \setminus \{n\}} \mu_{k \to \pi_n}(x_{\pi_n}) \\
&= \frac{1}{Z}\psi_{\pi_n}(x_{\pi_n}) \prod_{k \in \mathcal{N}(\pi_n)} \mu_{k \to \pi_n}(x_{\pi_n})
\end{aligned}$$

For the case $i = n$, we have:

$$p(x_n, x_{\pi_n}) = \sum_{x_{V \setminus \{n, \pi_n\}}} p(x) = \psi_n(x_n)\psi_{\pi_n}(x_{\pi_n})\psi_{n,\pi_n}(x_n, x_{\pi_n}) \underbrace{\sum_{x_{V \setminus \{n,\pi_n\}}} \frac{p(x)}{\psi_n(x_n)\psi_{\pi_n}(x_{\pi_n})\psi_{n,\pi_n}(x_n, x_{\pi_n})}}_{\alpha(x_{\pi_n})}$$

Therefore:

$$p(x_n, x_{\pi_n}) = \psi_{\pi_n}(x_{\pi_n})\alpha(x_{\pi_n})\psi_n(x_n)\psi_{n,\pi_n}(x_n, x_{\pi_n}) \tag{8.11}$$

Consequently:

$$p(x_{\pi_n}) = \psi_{\pi_n}(x_{\pi_n})\alpha(x_{\pi_n}) \underbrace{\sum_{x_n} \psi_n(x_n)\psi_{n,\pi_n}(x_n, x_{\pi_n})}_{\mu_{n \to \pi_n}(x_{\pi_n})}$$

Hence:

$$\alpha(x_{\pi_n}) = \frac{p(x_{\pi_n})}{\psi_{\pi_n}(x_{\pi_n})\mu_{n\to\pi_n}(x_{\pi_n})} \tag{8.12}$$

By using (7.31), (7.32) and the previous result, we deduct that:

$$\begin{aligned} p(x_n, x_{\pi_n}) &= \psi_{\pi_n}(x_{\pi_n})\psi_n(x_n)\psi_{n,\pi_n}(x_n, x_{\pi_n})\frac{p(x_{\pi_n})}{\psi_{\pi_n}(x_{\pi_n})\mu_{n\to\pi_n}(x_{\pi_n})} \\ &= \psi_{\pi_n}(x_{\pi_n})\psi_n(x_n)\psi_{n,\pi_n}(x_n, x_{\pi_n})\frac{\frac{1}{Z}\psi_{\pi_n}(x_{\pi_n})\prod_{k\in\mathcal{N}(\pi_n)}\mu_{k\to\pi_n}(x_{\pi_n})}{\psi_{\pi_n}(x_{\pi_n})\mu_{n\to\pi_n}(x_{\pi_n})} \\ &= \frac{1}{Z}\psi_{\pi_n}(x_{\pi_n})\psi_n(x_n)\psi_{n,\pi_n}(x_n, x_{\pi_n})\prod_{k\in\mathcal{N}(\pi_n)\setminus\{n\}}\mu_{k\to\pi_n}(x_{\pi_n}) \end{aligned}$$

By summing with respect to x_{π_n}, we get the result for $p(x_n)$:

$$p(x_n) = \sum_{x_{\pi_n}} p(x_n, x_{\pi_n}) = \frac{1}{Z}\psi_n(x_n)\mu_{\pi_n\to n}(x_n)$$

VI. A. Proposition:

Let $p \in \mathcal{L}(G)$, for $G = (V, E)$ a tree, then we have:

$$p(x_1, \dots, x_n) = \frac{1}{Z}\prod_{i\in V}\psi(x_i)\prod_{(i,j)\in E}\frac{p(x_i, x_j)}{p(x_i)p(x_j)} \tag{8.13}$$

Proof: we prove it by reccurence. The case $n = 1$ is trivial. Then, assuming that n is a leaf, and we can write $p(x_1, \dots, x_n) = p(x_1, \dots, x_{n-1})p(x_n \,|\, x_{\pi_n})$. But multiplying by $p(x_n \,|\, x_{\pi_n}) = \frac{p(x_n, x_{\pi_n})}{p(x_n)p(x_{\pi_n})}p(x_n)$ boils down to adding the edge potential for (n, π_n) and the node potential for the leaf n. The formula is hence verified by reccurence.

VI. B. Junction tree

Junction tree is an algorithm designed to tackle the problem of *inference on general graphs*. The idea is to look at a general graph from far away, where it can be seen as a tree. By merging nodes, one will hopefully be able to build a tree. When this is not the case, one can also think of adding some edges to the graph (i.e., cast the present distribution into a larger set) to be able to build such a graph.

The trap is that if one collapses too many nodes, the number of possibles values will explode, and as such the complexity of the whole algorithm. The *tree width* is the smallest possible clique size. For instance, for a 2D regular grid with n points, the tree width is equal to $\sqrt{n}$.

CHAPTER 9

Hidden MARKOV Model

The Hidden MARKOV Model[1] is one of the most used graphical models.

We consider $z_0, z_1, \dots, z_T$ states corresponding to latent variables and $y_0, y_1, \dots, y_T$ states corresponding to observed variables. The model assume that:

- $(z_t)_{0 \leq t \leq T}$ is a MARKOV chain (hence the name of the model),
- each z_t takes K possible values, denoted by $[\![1, K]\!]$,
- z_0 follows a multinomial distribution $\mathcal{M}(1, \pi_0)$,
- the transition probabilities are homogeneous: $p(z_t = k \mid z_{t-1} = k')$ does not depend on t. We denote by A the transition matrix,
- the emission probabilities $p(y_t \mid z_t)$ are homogeneous, i.e. $p(y_t \mid z_t) = f(y_t, z_t)$,
- The joint probability distribution function can be written as:

$$p(z_0, \dots, z_T, y_0, \dots, y_T) = p(z_0) \prod_{t=0}^{T-1} p(z_{t+1} \mid z_t) \prod_{t=0}^{T} p(y_t \mid z_t)$$

We want to perform different tasks that on this model:

- filtering: compute $p(z_t \mid y_1, \dots, y_{t-1})$,
- smoothing: compute $p(z_t \mid y_1, \dots, y_T)$,
- decoding: find $\max_{z_0, \dots, z_T} p(z_0, \dots, z_T \mid y_0, \dots, y_T)$.

All these tasks can be performed with a sum-product or max-product algorithm.

I. Sum-product

From now on, we note the observations $\overline{y} = \{\overline{y}_1, \dots, \overline{y}_T\}$. The distribution on y_t simply becomes the delta function $\delta_{y_t, \overline{y}_t}$.

To use the sum-product algorithm, we define z_T as the root, i.e. we send all forward messages to z_T and go back afterwards.

[1] modèle de MARKOV caché in french

Forward We compute the following messages:

$$\forall t \in [\![0, T]\!], \quad \mu_{y_t \to z_t}(z_t) = \sum_{y_t} \delta_{y_t = \overline{y}_t} p(y_t \mid z_t) = p(\overline{y}_t \mid z_t)$$

and recursively:

$$\mu_{z_{-1} \to z_0}(z_0) = p(z_0) \qquad \text{and} \qquad \forall t \in [\![0, T-1]\!], \quad \mu_{z_t \to z_{t+1}}(z_{t+1}) = \sum_{z_t} p(z_{t+1} \mid z_t) \mu_{z_{t-1} \to z_t}(z_t) \mu_{y_t \to z_t}(z_t)$$

With those messages we will be able to compute some conditional probabilities. Indeed let us introduce the "α-message":

$$\forall t \in [\![0, T]\!], \quad \alpha_t(z_t) = \mu_{y_t \to z_t}(z_t) \mu_{z_{t-1} \to z_t}(z_t)$$

We have the following property, due to the definition of the messages: $\alpha_t(z_t)$ represents a marginal of the distribution corresponding to the sub-HMM $\{z_0, \dots, z_t\}$:

PROPOSITION I. .1. *We have:*

$$\forall t \in [\![0, T]\!], \quad \alpha_t(z_t) = p(z_t, \overline{y}_0, \dots, \overline{y}_t)$$

PROOF It is easy to obtain the following recursion formule for the α-messages:

$$\forall t \in [\![0, T-1]\!], \quad \alpha_{t+1}(z_{t+1}) = p(\overline{y}_{t+1} \mid z_{t+1}) \sum_{z_t} p(z_{t+1} \mid z_t) \alpha_t(z_t)$$

Note that the result is true for $t = 0$. Thus by induction, we have if $\alpha_t(z_t) = p(z_t, \overline{y}_0, \dots, \overline{y}_t)$ for some $t \in [\![0, T-1]\!]$:

$$\begin{aligned}
\alpha_{t+1}(z_{t+1}) &= p(\overline{y}_{t+1} \mid z_{t+1}) \sum_{z_t} p(z_{t+1} \mid z_t) \alpha_t(z_t) && \text{by } \alpha\text{-recursion} \\
&= \sum_{z_t} p(\overline{y}_{t+1} \mid z_{t+1}, z_t, \overline{y}_0, \dots, \overline{y}_t) p(z_{t+1} \mid z_t, \overline{y}_0, \dots, \overline{y}_t) p(z_t, \overline{y}_0, \dots, \overline{y}_t) && \text{by independences} \\
&= \sum_{z_t} p(z_{t+1}, z_t, \overline{y}_0, \dots, \overline{y}_{t+1}) && \text{by chain rule} \\
&= p(z_{t+1}, \overline{y}_0, \dots, \overline{y}_{t+1})
\end{aligned}$$

□

Backward We compute recursively the following messages:

$$\mu_{z_{T+1} \to z_T}(z_T) = 1 \qquad \text{and} \qquad \forall t \in [\![1, T]\!], \quad \mu_{z_t \to z_{t-1}}(z_{t-1}) = \sum_{z_t} p(z_t \mid z_{t-1}) \mu_{z_{t+1} \to z_t}(z_t) \mu_{\overline{y}_t \to z_t}(z_t)$$

Defining the "β-message":

$$\forall t \in [\![0, T]\!], \quad \beta_t(z_t) = \mu_{z_{t+1} \to z_t}(z_t)$$

we have:

PROPOSITION I. .2.

$$\forall t \in [\![0, T]\!], \quad \beta_t(z_t) = p(\overline{y}_{t+1}, \dots, \overline{y}_T \,|\, z_t)$$

PROOF Use the recursion formula:

$$\forall t \in [\![0, T-1]\!], \quad \beta_t(z_t) = \sum_{z_{t+1}} p(z_{t+1} \,|\, z_t) p(\overline{y}_{t+1} \,|\, z_{t+1}) \beta_{t+1}(z_{t+1})$$

□

Using both α and β messages allow to compute several quantities:

PROPOSITION I. .3. *For all $t \in [\![0, T]\!]$, we have:*

$$p(z_t, \overline{y}_0, \dots, \overline{y}_T) = \alpha_t(z_t)\beta_t(z_t)$$

from which we can easily deduce:

$$p(\overline{y}_0, \dots, \overline{y}_T) = \sum_{z_t} \alpha_t(z_t)\beta_t(z_t) \qquad \textit{and} \qquad p(z_t \,|\, \overline{y}_0, \dots, \overline{y}_T) = \frac{\alpha_t(z_t)\beta_t(z_t)}{\sum_{z_t} \alpha_t(z_t)\beta_t(z_t)}$$

We also have for $t \in [\![0, T-1]\!]$:

$$p(z_t, z_{t+1} \,|\, \overline{y}_0, \dots, \overline{y}_T) = \frac{1}{p(\overline{y}_0, \dots, \overline{y}_T)} \alpha_t(z_t)\beta_{t+1}(z_{t+1}) p(z_{t+1} \,|\, z_t) p(\overline{y}_{t+1} \,|\, z_{t+1})$$

Implementation is not difficult, but requires to avoid errors in the indices! Also, in order to prevent numerical errors, it is better to code them using $\log$-probabilities.

II. EM algorithm

With the previous notations and assumptions, we write the complete $\log$-likelihood $\ell_c(\theta)$ where θ is the vector of parameters of the model[2]:

$$\begin{aligned}
\ell_c(\theta) &= \log\left(p(z_0) \prod_{t=0}^{T-1} p(z_{t+1} \,|\, z_t) \prod_{t=0}^{T} p(\overline{y}_t \,|\, z_t)\right) \\
&= \log p(z_0) + \sum_{t=0}^{T-1} \log p(z_{t+1} \,|\, z_t) + \sum_{t=0}^{T} \log p(\overline{y}_t \,|\, z_t) \\
&= \sum_{k=1}^{K} \delta_{z_0,k} \log(\pi_0)_k + \sum_{t=0}^{T-1} \sum_{k,k'=1}^{K} \delta_{z_{t+1},k}\delta_{z_t,k'} \log A_{k,k'} + \sum_{t=0}^{T} \sum_{k=1}^{K} \delta_{z_t,k} \log f(\overline{y}_t, z_t)
\end{aligned}$$

When applying EM algorithm to estimate the parameters of this HMM, we use JENSEN's inequality to obtain a lower bound on the $\log$-likelihood:

$$\log p(\overline{y}_0, \dots, \overline{y}_T) \geq \mathbb{E}_q[\log p(z_0, \dots, z_T, \overline{y}_0, \dots, \overline{y}_T)] = \mathbb{E}_q[\ell_c(\theta)]$$

[2] containing π_0, A but also parameters for f

At step i, we use q defined by $q(z_0, \dots, z_T) = p_{\theta^{(i)}}(z_0, \dots, z_T \,|\, \overline{y}_0, \dots, \overline{y}_T)$. Thus the E-step consists in replacing the δ values in the $\log$-likelihood expression by their expectation. For instance $\delta_{z_0,k}$ is replaced by $p_{\theta^{(i)}}(z_0 = k \,|\, \overline{y})$.

For the M-step, we maximize the obtained expression w.r.t. θ in the usual manner to obtain a new estimator θ^{i+1}. The key is that everything will decouple, thus maximizing is simple and can be done in closed form.

Addressing practical implementation issues

Since α_t and β_t are respectively joint probabilities of $t+1$ and $T-t$ variables they tend to become exponentially small respectively for t large and t small. A naive implementation of the forward-backward algorithm therefore typically leads to rounding errors. It is therefore necessary to work on a logarithmic scale.

So when considering operations on quantities say $a_1, \dots, a_n$ whose logarithms are $\ell_i = \log(a_i)$, the $\log$ of the product is easily computed as $\ell_\Pi = \log \prod_i a_i = \sum_i \ell_i$ and the $\log$ of the sum can be computed with the smallest amount of numerical errors by factoring the largest element. Precisely if $i_* = \operatorname{argmax}_i a_i$ and $\ell_* = \log a_{i_*}$ then:

$$\ell_\Sigma = \log \sum_i a_i = \log \sum_i \exp(\ell_i) = \log \left(\exp(\ell_*) \sum_i \exp(\ell_i - \ell_*) \right) = \ell_* + \log \left(1 + \sum_{i \neq i_*} \exp(\ell_i - \ell_*) \right)$$

which provides a stable way of computing the logarithm of the sum.

For hidden MARKOV models, remember that the max-product algorithm[3] allows to compute the most probable sequence for hidden states.

[3] a.k.a. VITERBI algorithm

CHAPTER 10

Back to classification

[todo]

I. Principal Component Analysis (PCA)

Framework: $x_1, \dots, x_N \in \mathbb{R}^d$
Goal: put points on a closest affine subspace

a. Analysis view

$$\text{Find } w \in \mathbb{R}^d \text{ such that } \mathrm{Var}(x^T w) \text{ is maximal, with } ||w|| = 1$$

With centered data, i.e. $\frac{1}{N}\sum_{n=1}^N x_n = 0$, the empirical variance is:

$$\hat{\mathrm{Var}}(x^T w) = \frac{1}{N}\sum_{n=1}^N (x_n^T w)^2 = \frac{1}{N} w^T (X^T X) w$$

where $X \in \mathbb{R}^{N \times d}$ is the design matrix. In this case: w is the eigenvector of $X^T X$ with largest eigenvalue. It is not obvious *a priori* that this is the direction we care about.
If more than one direction is required, one can use *deflation*:

1. Find w
2. Project x_n onto the orthogonal of $\mathrm{Vect}(w)$
3. Start again

b. Synthesis view

$$\min_w \sum_{n=1}^N d(x_n, \{w^T x = 0\})^2 \text{ with } w \in \mathbb{R}^D, \ ||w|| = 1.$$

Advantage: if one wants more than 1 dimension, replace $\{w^T x = 0\}$ by any subspace.

c. Probabilistic approach: Factor Analysis

Model:

- $\Lambda = (\lambda_1, \dots, \lambda_K) \in \mathbb{R}^{d \times k}$
- $X \in \mathbb{R}^k \sim \mathcal{N}(0, I)$
- $\epsilon \sim \mathcal{N}(0, \Psi)$, $\epsilon \in \mathbb{R}^d$ independent from X with Ψ diagonal.
- $Y \in \mathbb{R}^d$: $Y = \Lambda X + \mu + \epsilon$

We have $Y \mid X \sim \mathcal{N}(\Lambda X + \mu, \Psi)$.
Problem: get $X \mid Y$.

(X, Y) is a Gaussian vector on $\mathbb{R}^{d+k}$ which satisfies:

- $\mathbb{E}[X] = 0 = \mu_X$
- $\mathbb{E}[Y] = \mathbb{E}[\Lambda X + \mu + \epsilon]\mu = \mu_Y$
- $\Sigma_{XX} = I$
- $\Sigma_{XY} = \mathrm{Cov}(X, \Lambda X + \mu + \epsilon) = \mathrm{Cov}(X, \Lambda X) = \Lambda^T$
- $\Sigma_{YY} = \mathrm{Var}(\Lambda X + \epsilon, \Lambda X + \epsilon) = \mathrm{Var}(\Lambda X, \Lambda X) + \mathrm{Var}(\epsilon, \epsilon) = \Lambda\Lambda^T + \Psi$

Thanks to the results we know on exponential families, we know how to compute $X \mid Y$:

$$\begin{aligned} \mathbb{E}[X \mid Y = y] &= \mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}(y - \mu_Y) \\ \mathrm{Cov}[X \mid Y = y] &= \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} \end{aligned}$$

In our case, we therefore have:

$$\begin{aligned} \mathbb{E}[X \mid Y = y] &= \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}(y - \mu) \\ \mathrm{Cov}[X \mid Y = y] &= I - \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}\Lambda^T \end{aligned}$$

To apply EM, one needs to write down the complete log-likelihood.

$$\log p(X, Y)\alpha - \frac{1}{2}X^T X - \frac{1}{2}(Y - \Lambda X - \mu)^T\Psi^{-1}(Y - \Lambda X - \mu) - \frac{1}{2}\log\det\Psi$$

Trap: $\mathbb{E}[XX^T \mid Y] \neq \mathrm{Cov}(X \mid Y)$
Rather, $\mathbb{E}[XX^T \mid Y] = \mathrm{Cov}(X \mid Y) + \mathbb{E}[X \mid Y]\mathbb{E}[X \mid Y]^T$

> REMARK I. .1.
>
> - $\mathrm{Cov}(X) = \Lambda\Lambda^T + \Psi$: our parameters are not identifiable, $\Lambda \leftarrow \Lambda R$ with R a rotation gives the same results (in other words, a subspace has different orthonormal bases).
> - Why do we care ?
>
> 1. A probabilistic interpretation allows to model in a finer way the problem.
> 2. It is very flexible and therefore allows to combine multiple models.

II. Multiclass classification

We return briefly to classification to mention two simple yet classical and useful models for multi-class classification: the naive Bayes model and the multiclass logistic regression. We consider clas-

sification problems where the input data is in $\mathcal{X} = \mathbb{R}^p$ and the output variable is a binary indicator in $\mathcal{Y} = \{y \in \{0,1\}^K \mid y_1 + \ldots + y_K = 1\}$.

II. A. Naive Bayes classifier

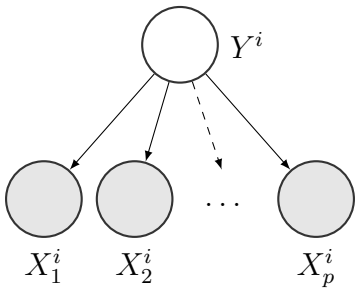
The naive Bayes classifier is relevant when modeling the joint distribution of $p(x \mid y)$ is too complicated. We will present it the special case where the input data is a vector of binary random variable. $X^i : \Omega \mapsto \{0,1\}^p$

A practical example of classification problem in this setting is the problem of classification of documents based on a bag of word representation. In the bag-of-word approach, a document is represented as a long binary vector which indicates for each word of a reference dictionary whether that word is present in the document considered or not. So the document i would be represented by a vector $x^i \in \{0,1\}^p$, with $x^i_j = 1$ iff word j of the dictionary is present in the ith document.

As we saw in the second lecture, it is possible to approach the problem using directly a *conditional model* of $p(y \mid x)$ or using a *generative model* of the joint distribution modeling separately $p(y)$ and $p(x \mid y)$ and computing $p(y \mid x)$ using Bayes rule. The naive Bayes model is an instance of a generative model. By contrast the multi class logistic regression of the following section is an example of a conditional model.

Y^i is naturally modeled as a multinomial distribution with $p(y^i) = \prod_{k=1}^K \pi_k^{y^i_k}$. However $p(x^i \mid y^i) = p(x^i_1, \ldots, x^i_p \mid y^i)$ has a priori $2^p - 1$ parameters. The key assumption made in the naive Bayes model is that $X^i_1, \ldots, X^i_p$ are all independent conditionally on Y^i. This assumption is not realistic and simplistic, hence the term "naive". This assumption is clearly not satisfied in practice for documents where one would expect that there would be correlations between words that are not just explained by a document category. The corresponding modeling strategy is nonetheless working well in practice.

These conditional independence assumptions correspond to the following graphical model:



The distribution of Y^i is a multinomial distribution which we parameterize with $(\pi_1, ..., \pi_K)$, and we write $\mu_{jk} = P(X^{(i)}_j = 1 \mid Y^{(i)}_k = 1)$ We then have

$$p(X^i = x^i, Y^i = y^i) = p(x_i, y_i) = p(x^i \mid y^i)p(y^i) = \prod_{j=1}^{p} p(x^i_j \mid y^i)p(y^i)$$

which leads to

$$p(x^i, y^i) = \Big[\prod_{j=1}^{p} \prod_{k=1}^{K} \mu_{jk}{}^{x_j^i y_k^i} (1 - \mu_{jk})^{(1-x_j^i) y_k^i} \Big] \prod_{k=1}^{K} \pi_k^{y_k^i}$$

and

$$\log p(x^i, y^i) = \sum_{k=1}^{K} \Big(\sum_{j=1}^{p} \Big(x_j^i y_k^i \log \mu_{jk} + (1 - x_j^i) y_k^i \log(1 - \mu_{jk}) \Big) + y_k^i \log(\pi_k) \Big)$$

We can then use Bayes' rule (hence the "Bayes" in "Naive Bayes"), which leads to

$$\log p(y^i \mid x^i) = \eta(x^i)^\top y^i - A(\eta(x^i))$$

with $\eta(x) = (\eta_1(x), \dots, \eta_K(x)) \in \mathbb{R}^K$ and

$$\eta_k(x) = w_k^\top x + b_k, \quad w_k \in \mathbb{R}^p, \quad [w_k]_j = \log \frac{\mu_{jk}}{1 - \mu_{jk}}, \quad b_k = \log \pi_k.$$

Note that, in spite of the name the naive Bayes classifier is not a Bayesian approach to classification.

a. Multiclass logistic regression

In the light of the course on exponential families, the logistic regression model can be seen as resulting from a linear parameterization as a function of x of the natural parameter $\eta(x)$ of the Bernoulli distribution corresponding to the conditional distribution of Y given $X = x$. Indeed for binary classification, we have that $Y \mid X = x \sim \mathsf{Ber}(\mu(x))$ and in the logistic regression model we set $\mu(x) = \exp(\eta(x) - A(\eta(x))) = (1 + \exp(-\eta(x))^{-1}$ and $\eta(x) = w^\top x + b$.

It is then natural to consider the generalization to a multiclass classification setting. In that case, $Y \mid X = x$ is multinomial distribution with natural parameters $(\eta_1(x), \dots, \eta_K(x))$. To again parameterize them linearly as a function of x, we need to introduce parameters $w_k \in \mathbb{R}^p$ and $b_k \in \mathbb{R}$, for all $1 \le k \le K$ and set $\eta_k(x) = w_k^\top x + b_k$. We then have

$$\mathsf{p.p.}(Y_k = 1 \mid X = x) = \exp(\eta_k(x) - A(\eta(x))) = \frac{e^{\eta_k(x)}}{\sum_{k'=1}^{K} e^{\eta_{k'}(x)}} = \frac{e^{w_k^\top x + b_k}}{\sum_{k'=1}^{K} e^{w_{k'}^\top x + b_{k'}}},$$

and thus

$$\log \ \mathsf{p.p.}(Y_k = y \mid X = x) = \sum_{k=1}^{K} y_k (w_k^\top x + b_k) - \log \Big[\sum_{k'=1}^{K} e^{w_{k'}^\top x + b_{k'}} \Big].$$

Like for binary logistic regression, the maximum likelihood principle can used to learn $(w_k, b_k)_{1 \le k \le K}$ using numerical optimization methods such as the IRLS algorithm.

Note that the form of the parameterization obtained is the same as for the Naive Bayes model; however, the Naive Bayes model is learnt as a generative model, while the logistic regression is learnt as conditional model.

We have not talked about the multi class generalization of Fisher's linear discriminant. It exists as well as the multi class counterpart of the model seen for binary regression. It relies like in the binary case on the assumption that $p(x \mid y)$ is Gaussian. This is good exercise to derive it.

III. Learning on graphical models

III. A. ML principle for general Graphical Models

Directed graphical model

Proposition : Let G be a directed graph with p nodes. Assume that $(X^1, ...X^n)$ are i.i.d., with p features : i.e $\forall i \in \{1, .., n\}, X_i \in \mathbb{R}^p$, and that are fully observed, i.e., there is no latent or hidden variable among them. Then the ML principle decouples in p optimisation problems.

Proof : Let us assume we have a decoupled model $\mathcal{P}_\Theta$, i.e. :

$$\mathcal{P}_\Theta := \left\{ p_\theta(x) = \prod_j p(x_j \mid x_{\pi_j}, \theta_j) \mid \theta = (\theta_1, ..., \theta_p) \in \Theta = \Theta_1 \times ... \times \Theta_p \right\}$$

$$L(\theta) = \prod_{i=1}^{n} p(x^i \mid \theta) = \prod_{i=1}^{n} \prod_{j=1}^{p} p(x_j^i \mid x_{\pi_j}^i, \theta_j)$$

$$\ell(\theta) = \sum_{j=1}^{p} \sum_{i=1}^{n} \log p(x_j^i \mid x_{\pi j}^i, \theta_j).$$

Then the ML principle reduces to solving p optimization problems of the form

$$\max_{\theta_j} \ell_j(\theta_j) \quad \text{s.t} \quad \theta_j \in \Theta_j, \qquad \text{with} \qquad \ell_j(\theta_j) := \sum_{i=1}^{n} \log p(x_j^i \mid x_{\pi_j}^i, \theta_j).$$

Undirected graphical model

→ The ML problem is convex with respect to canonical parameters if: the data is fully observed (no latent or hidden variable), and the parameters are decoupled.
→ In general, if the data is not fully observed, the EM scheme or similar scheme is used.
If the parameters are coupled, the problem remains convex in some cases (e.g linear coupling), but not in general.
→ If the model is a tree, one can reformulate the model as a directed tree to get back to the directed case.
→ In general, to compute the gradient of the log partition function and thus to compute the gradient of the log-likelihood, it is necessary to perform *probabilistic inference* on the model (i.e. to compute $\nabla A(\theta) = \mu(\theta) = \mathbb{E}_\theta[\phi(X)]$). If the model is a tree, this can be done with the sum-product algorithm and if the model is a close to a tree, the junction tree theory can be leveraged to perform *probabilistic inference*; however in general *probabilistic inference* is NP-hard and so one needs to use approximate probabilistic inference techniques.

CHAPTER 11

Approximate inference with MONTE-CARLO methods

I. Sampling methods

We often need to compute the expectation of a function f under a distribution p that cannot be computed. This corresponds to compute $\mu = \mathbb{E}[f(X)]$ where X is a random variable following distribution p.

EXAMPLE I. .1. For $X = (X_1, \dots, X_n)$ the vector of variables corresponding to a graphical model, we can consider $f : X \longmapsto \delta_{X_I = x_I}$ for a fixed subset I of $[\![1, n]\!]$, then:

$$\mathbb{E}[f(X)] = p(X_I = x_I)$$

If we know how to sample from p, we can use the following method:

Algorithm 6: MONTE-CARLO estimation

Input : p, n
Output: $\hat{\mu}$

1 Draw $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} p$
2 $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$

This method relies on the two following results:

PROPOSITION I. .1. [LAW OF LARGE NUMBERS (LLN)]
If X is an integrable random variable ($\mathbb{E}[X]$ is defined and finite), and $(X_i)_{i \in \mathbb{N}^*} \overset{\text{i.i.d.}}{\sim} X$, then:

$$\frac{1}{n} \sum_{i=1}^{n} X_i \underset{n \to +\infty}{\longrightarrow} \mathbb{E}[X] \quad \text{a.s.}$$

THEOREM I. .2. [CENTRAL LIMIT THEOREM (CLT)]
If X is a random variable such that $\mathrm{Var}(X) = \sigma^2 < +\infty$, and $(X_i)_{i \in \mathbb{N}^*} \overset{\text{i.i.d.}}{\sim} X$, then:

$$\sqrt{n}\Big(\frac{1}{n} \sum_{i=1}^{n} X_i - \mathbb{E}[X]\Big) \underset{n \to +\infty}{\xrightarrow{\mathcal{D}}} \mathcal{N}(0, \sigma^2)$$

from which we deduce that $\hat{\mu} \underset{n\to+\infty}{\longrightarrow} \mu$ a.s. and $\mathbb{E}[|\hat{\mu} - \mu|^2] = \sigma^2/n$.

The hard question is: how to sample from a specific distribution?

I. A. Inverse transform sampling

Assume we can draw a uniform $U \sim \mathcal{U}([0,1])$ distribution[1].

We can easily draw a $\mathcal{B}(p)$ distribution by taking $X = \mathbb{1}_{U \leq p}$.

DEFINITION I. .1. [INVERSE TRANSFORM]
For a distribution p with cumulative distribution function F, we define

$$F^{-1} : u \longmapsto \inf\{x \in \mathbb{R} \mid F(x) \geq u\}$$

PROPOSITION I. .3. *If $U \sim \mathcal{U}([0,1])$ then $X = F^{-1}(U) \sim p$.*

PROOF If F is inversible we have:

$$p(X \leq x) = p(F^{-1}(U) \leq x) = p(U \leq F(y)) = F(y)$$

Otherwise we admit the result. □

EXAMPLE I. .2. For an exponential distribution[2] $p : x \longmapsto \lambda \mathrm{e}^{-\lambda x} \mathbb{1}_{\mathbb{R}_+}(x)$, we have

$$F^{-1} = -\frac{\log}{\lambda} \qquad \text{and} \qquad X = -\frac{1}{\lambda}\ln(U) \sim p$$

I. B. Ancestral sampling

Consider the problem of sampling from a directed graphical model whose distribution takes the form

$$p(x_1, \dots, x_n) = \prod_{i=1}^{n} p(x_i \mid x_{\pi_i})$$

We assume, without loss of generality, that the variables are indexed in a topological order.

Consider the following algorithm:

PROPOSITION I. .4. *The random variable $(x_1, \dots, x_n)$ returned by the ancestral sampling algorithm follows exactly the joint distribution p.*

PROOF We prove the result by induction. It is clearly obvious for a graph with a single node.

For two nodes corresponding the pair of variable (X_1, X_2), then either X_1 and X_2 are independent and we are back to the single node case. Or $\pi_2 = \{1\}$ and then x_1 is drawn from p_{X_1} and, given

[1] use `rand` in Python
[2] one of the rare cases admitting an explicit inverse CDF

Algorithm 7: Ancestral sampling

Input : n, $p(X_i \mid X_{\pi_i})$ for all $i \in [\![1, n]\!]$
Output: $(x_1, \dots, x_n)$

for $i = 1$ *to* n **do**
  Draw x_i from $p(X_i \mid X_{\pi_i} = x_{\pi_i})$
end

the value x_1 obtained, x_2 is drawn from the conditional distribution $p(X_2 \mid X_1 = x_1)$, then the pair (x_1, x_2) follows the joint distribution p_{X_1,X_2}.

Now, assuming the result is true for $n-1$ nodes we prove that it is also true for n nodes. First note that after sampling $x_1, \dots, x_{n-1}$, we know that $(x_1, \dots, x_{n-1})$ follows the distribution given by $\prod_{i=1}^{n-1} p(x_i \mid x_{\pi_i})$ which is exactly the marginal distribution of $(X_1, \dots, X_{n-1})$. But then x_n is drawn according to the distribution $p(X_n \mid X_{\pi_n} = z_{\pi_n})$ which by the MARKOV property is equal to $p(X_n \mid (X_1, \dots, X_{n-1}) = (z_1, \dots, z_{n-1}))$. Applying the two nodes case to $\tilde{X}_2 = X_n$ and $\tilde{X}_1 = (X_1, \dots, X_{n-1})$, we obtain that $(x_1, \dots, x_{n-1})$ is indeed drawn from the joint distribution $(X_1, \dots, X_n)$. By induction the result is proven. □

I. C. Rejection sampling

Assume that $p(X)$ the distribution of X admits a density w.r.t. some measure μ[3], known up to a normalizing constant, i.e. we know $\tilde{p}$ such that $p = \frac{\tilde{p}}{Z}$.

Assume that we can construct and compute a probability distribution q such that $\tilde{p} \leq kq$ and assume we can sample from q. We define the rejection sampling algorithm as :

Algorithm 8: Rejection Sampling

Input : $n, \tilde{p}, q, k$
Output: x

$accept = 0$
while $accept = 0$ **do**
  Draw x from q
  Draw $accept$ from $\mathcal{B}(\frac{\tilde{p}(x)}{kq(x)})$
end

PROPOSITION I. .5. *The returned sample x returned by the rejection sampling follow distribution p.*

PROOF We write the proof for the case of a discrete random variable. We have:

$$p(X = x, X \text{ is accepted}) = p(X \text{ is accepted} \mid X = x)p(X = x) = \frac{\tilde{p}(x)}{kq(x)}q(x) = \frac{\tilde{p}(x)}{k}$$

[3] typically the LEBESGUE measure for a continuous random variable and the counting measure for a discrete variable

and so

$$p(X \text{ is accepted}) = \sum_x \frac{\tilde{p}(x)}{k} = \frac{Z_p}{k}$$

which gives:

$$p(X = x \mid X \text{ is accepted}) = \frac{\tilde{p}(x)}{k}\frac{k}{Z_p} = p(x)$$

To write the general version of this proof formally for any random variable admiting a density w.r.t. μ, we would need to define A to be the BERNOULLI random variable such that $A = \mathbb{1}_{X \text{ is accepted}}$ and to consider $p_{X,A}$ the joint density of (X, A) w.r.t. the product measure $\mu \times \nu$, where ν is the counting measure on $\{0, 1\}$. The proof is then the same as above. □

REMARK I. .1. In practice, finding q and k such that acceptance has a reasonably large probability is hard, because it requires to find a fairly tight bound on $p(x)$ over the entire space.

I. D. Importance sampling

Assume $X \sim p$ and $Y \sim q$. We aim at computing the expectation of a function $f(X)$. One has:

$$\mathbb{E}[f(X)] = \int f(x)p(x)dx = \int \frac{f(x)p(x)}{q(x)}q(x)dx = \mathbb{E}\Big[f(Y)\frac{p(Y)}{q(Y)}\Big] = \mathbb{E}[g(Y)]$$

where $g = fp/q$. Thus is we can sample from q, we can approximate $\mathbb{E}[f(X)]$ by a MONTE-CARLO estimation:

$$\mathbb{E}[f(X)] \simeq \hat{\mu} = \frac{1}{n}\sum_{i=1}^n g(Y_i) = \frac{1}{n}\sum_{i=1}^n f(Y_i)\frac{p(Y_i)}{q(Y_i)}$$

where $(Y_i)_{1 \leq i \| eqn} \overset{\text{i.i.d.}}{\sim} q$. The weights $(w(Y_i) = \frac{p(Y_j)}{q(Y_j)})_{1 \leq i \leq n}$ are called *importance weights*. We have:

$$\mathbb{E}[\hat{\mu}] = \frac{1}{n}\sum_{i=1}^n \int f(x)\frac{p(x)}{q(x)}q(x)dx = \int f(x)p(x)dx = \mu$$
$$\mathrm{Var}(\hat{\mu}) = \frac{1}{n}\mathrm{Var}\left(\frac{f(Y)p(Y)}{q(Y)}\right)$$

LEMMA I. .6. *Assume $|f| \leq M$ a.s., then:*

$$\mathrm{Var}(\hat{\mu}) \leq \frac{M^2}{n}\mathbb{E}\Big[\frac{p(X)}{q(X)}\Big]$$

PROOF It simply comes from $\mathrm{Var}(Z) \leq \mathbb{E}[Z^2]$. □

REMARK I. .2. We have:

$$\mathbb{E}\Big[\frac{p(X)}{q(X)}\Big] = \int \frac{p(x)^2}{q(x)}dx = \int \frac{(p(x)-q(x))^2}{q(x)}dx + \int 2p(x) - q(x)dx = \underbrace{\int \frac{(p(x)-q(x))^2}{q(x)}dx}_{\chi^2 \text{ divergence between } p \text{ and } q} + 1$$

Hence, importance sampling will give good results if q has mass where p has. Indeed, if for some y, $q(y) \ll p(y)$, our estimator may have a very large variance.

Extension of importance sampling Assume we only know p and q up to a constant : $p = \frac{\tilde{p}}{Z_p}$ and $q = \frac{\tilde{q}}{Z_q}$, with $\tilde{p}$ and $\tilde{q}$ known. Then:

$$\mathbb{E}\Big[f(Y)\frac{\tilde{p}(Y)}{\tilde{q}(Y)}\Big] = \frac{Z_p}{Z_q}\mathbb{E}\Big[f(Y)\frac{p(Y)}{q(Y)}\Big] = \frac{Z_p}{Z_q}\mu$$

and the LLN gives

$$\hat{\tilde{\mu}} = \frac{1}{n}\sum_{i=1}^{n} f(Y_i)\frac{\tilde{p}(Y_i)}{\tilde{q}(Y_i)} \underset{n\to+\infty}{\longrightarrow} \frac{Z_p}{Z_q}\mu \quad \text{a.s.}$$

Taking $f = 1$, we get

$$\hat{Z}_{p/q} = \frac{1}{n}\sum_{i=1}^{n} \frac{p(Y_i)}{q(Y_i)} \underset{n\to+\infty}{\longrightarrow} \frac{Z_p}{Z_q} \quad \text{a.s.}$$

Thus we obtain:

$$\hat{\mu} = \frac{1}{\hat{Z}_{p/q}}\hat{\tilde{\mu}} \underset{n\to+\infty}{\longrightarrow} \mu \quad \text{a.s.}$$

REMARK I. .3. Even if $Z_p = Z_q = 1$, in practice renormalizing by $\hat{Z}_{p/q}$ often improves the estimation.

II. MARKOV chain MONTE-CARLO

Unfortunately, the previous techniques are often insufficient, especially for complex multivariate distributions, so that it is not possible to draw exactly from the distribution of interest or to obtain a reasonably good estimates based on importance sampling. The idea of MCMC is that in many cases, even though it is not possible to sample directly from a distribution of interest, it is possible to construct a MARKOV chain $(X_t)_{t\geq 0}$ whose distribution $q_t = p(X_t)$ converges to a target distribution $p(Y)$.

The idea is then that if T_0 is sufficiently large, we can consider that for all $t > T_0$, X_t follows approximately the distribution p and so:

$$\frac{1}{T-T_0}\sum_{t=T_0+1}^{T} f(X_t) \simeq \frac{1}{T-T_0}\sum_{t=T_0+1}^{T} f(Y_t) \underset{T\to+\infty}{\longrightarrow} \mathbb{E}[f(Y)]$$

where $(Y_t)_{t\geq 0} \overset{\text{i.i.d.}}{\sim} Y$.

Note that there is a double approximation: one due to the the approximation $q_t \approx p$ for t sufficiently large and the second due to the use of the law of large numbers. Note also that the draws of $(X_t)_{t\geq 0}$ are not independent (but this is not necessary here to have a law of large numbers).

The period before T_0 is often called the burn in period. The most classical procedure to obtain such a MARKOV chain in the context of graphical models is called GIBBS sampling. We will see it in more details later.

> NOTATION II. .1. In this whole section we write X_t instead of $X^{(t)}$ which would match better with other sections. Indeed, here X_t should be thought typically as the whole vector of variables corresponding to a graphical model $X_t = (X_{t,j})_{1\leq j\leq n}$. We write t as an index just to simplify notations.

In the following we assume that we work with random variables taking values in a set $\mathcal{X}$ with $|\mathcal{X}| = K < +\infty$. However K is typically very large since it corresponds to all the configurations that the set of variables of a graphical model can take.

One can find a review on MARKOV chains in Annex VI. .

METROPOLIS-HASTINGS algorithm We consider a proposal distribution $Z \mid X$ that we can sample, denoted by R, and we define an acceptance probability $\alpha(x, z)$ of accepting $Z = z$ when $X = x$.

Algorithm 9: METROPOLIS-HASTINGS

Input : q, R, T, α
Output: $(x_t)_{t\in[\![0,T]\!]}$

Draw x_0 from q
for $t \in [\![1, T]\!]$ **do**
  Draw z_t from $p(Z \mid X_{t-1} = x_{t-1}) = R(x_{t-1}, .)$
  Set $x_t = z_t$ with probability $\alpha(x_{t-1}, z_t)$, otherwise set $x_t = x_{t-1}$
end

> **PROPOSITION II. .1.** *Assume $\mathcal{X}$ is finite and R is the transition matrix of an irreducible* **[aperiodic?]** MARKOV *chain such that $R(x, z) > 0 \Longrightarrow R(z, x) > 0$ for any $x, z \in \mathcal{X}$ and $p(x) > 0$ for any $x \in \mathcal{X}$.*
>
> *Then it exists a choice of α such that the* METROPOLIS-HASTINGS *algorithm defines a* MARKOV *chain that converges to p.*

PROOF We define S the transition matrix of $(X_t)_{t\geq 0}$. We have:

$$\forall z \neq x, S(x, z) = \begin{cases} R(x, z)\alpha(x, z) & \text{if } x \neq z \\ R(x, x) + \sum_{z'\neq x} R(x, z')(1 - \alpha(x, z')) & \text{otherwise} \end{cases}$$

We want to choose S such that we have *detailed balance*[4]: we just need to have it for every $x \neq z$

[4] i.e. S is reversible

as it is automatic for (x, x). Then

$$p(x)S(x,z) = p(z)S(z,x) \iff p(x)R(x,z)\alpha(x,z) = p(z)R(z,x)\alpha(z,x) \iff \frac{\alpha(x,z)}{\alpha(z,x)} = \frac{p(z)R(z,x}{p(x)R(x,z}$$

Defining α as :

$$\forall x, z, \quad \alpha(x,z) = \min\left(1, \frac{p(z)R(z,x)}{p(x)R(x,z)}\right)$$

then α takes values in $[0,1]$ and the last equation of the above equation is satisfied for all $x \neq z$. Thus p is a reversible distribution of the chain, so the chain converges to p as we can show □

III. Approximate inference with MCMC

III. A. GIBBS sampling

Let us consider an undirected graph and its associated distribution p from which we want to sample (in order to do inference for example). We assume that:

- It is difficult to sample directly from p.
- It is easy to sample from[5] $\mathbb{P}_p(X_i \mid X_{-i})$.

The idea consists in using the MARKOV property so that:

$$\mathbb{P}_p(X_i \mid X_{-i}) = \mathbb{P}_p(X_i \mid X_{N_i})$$

where N_i is the MARKOV blanket of node i. Based on this, GIBBS sampling is a process that converges in distribution to p.

The most classical version of the GIBBS sampling algorithm is the *cyclic scan* GIBBS *sampling*:

Algorithm 10: cyclic scan GIBBS sampling

Input : $T, (\mathbb{P}_p(X_i \mid X_{-i}))_{1 \leq i \leq n}$
Output: $x^{(T)}$

Initialize $x^{(0)}$ and $t = 0$
while $t < T$ **do**
  for $i \in [\![1, n]\!]$ **do**
    $t = t + 1$
    Draw $x_i^{(t)}$ from $\mathbb{P}_p(X_i \mid X_{-i} = x_{-i}^{(t-1)})$
    Set $x_j^{(t)} = x_j^{(t-1)}$ for $j \neq i$
  end
end

Another version of the algorithm called *random scan* GIBBS *sampling* consists in picking the index i at random at each step t:

[5] recall that $X_{-i} = X_{[n]\setminus\{i\}}$

Algorithm 11: Random scan GIBBS sampling

Input : $T, (\mathbb{P}_p(X_i \mid X_{-i}))_{1 \leq i \leq n}$
Output: $x^{(T)}$

Initialize $x^{(0)}$
for $t \in [\![1, T]\!]$ **do**
  Draw i from $\mathcal{U}([\![1, n]\!])$
  Draw $x_i^{(t)}$ from $\mathbb{P}_p(X_i \mid X_{-i} = x_{-i}^{t-1})$
  Set $x_j^{(t)} = x_j^{(t-1)}$ for $j \neq i$
end

III. B. Application to the ISING model

Let us now consider the ISING model on a graph $G = (V, E)$ with $V = [n]$. X is a random variable which takes values in $\{0, 1\}^n$ with a probability distribution that depends on some parameter η:

$$\forall x \in \{0,1\}^n, \quad p_\eta(x) = \exp\Big(\sum_{i \in V} \eta_i x_i + \sum_{(i,j) \in E} \eta_{ij} x_i x_j - A(\eta)\Big)$$

To apply the Gibbs sampling algorithm, we need to compute $\mathbb{P}(X_i \mid X_{-i})$.

We have

$$p(x) = p(x_i, x_{-i}) = \frac{1}{Z(\eta)} \exp\left(\eta_i x_i + \sum_{j \in N_i} \eta_{ij} x_i x_j + \sum_{j \neq i} \eta_j x_j + \sum_{(j,j') \in E \mid j,j' \neq i} \eta_{jj'} x_j x_{j'}\right)$$

and thus

$$p(x_{-i}) = \frac{1}{Z(\eta)} \sum_{z \in \{0,1\}} \exp\left(\eta_i z + \sum_{j \in N_i} \eta_{ij} z x_j + \sum_{j \neq i} \eta_j x_j + \sum_{(j,j') \in E \mid j,j' \neq i} \eta_{jj'} x_j x_{j'}\right)$$

Taking the ratio of the two previous quantities, the two last terms of the exponentials cancel out and we get

$$\mathbb{P}(x_i \mid x_{-i}) = \frac{\exp\left(\eta_i x_i + \sum_{j \in N_i} \eta_{ij} x_i x_j\right)}{1 + \exp\left(\eta_i + \sum_{j \in N_i} \eta_{ij} x_j\right)}$$

In particular:

$$\mathbb{P}(X_i = 1 \mid x_{-i}) = \frac{1}{1 + \exp\Big(-\big(\eta_i + \sum_{j \in N_i} \eta_{ij} x_j\big)\Big)} = \sigma\Big(\eta_i + \sum_{j \in N_i} \eta_{ij} x_j\Big)$$

where σ is the logistic function $\sigma : z \longmapsto (1 + \mathrm{e}^{-z})^{-1}$.

Without surprise, the conditional distribution $\mathbb{P}(X_i = x_i \mid X_{-i} = x_{-i})$ only depends on the variables that are neighbors of i in the graph and that form its MARKOV blanket, since we must have

$$\mathbb{P}(X_i = x_i \mid X_{-i} = x_{-i}) = \mathbb{P}(X_i = x_i \mid X_{N_i} = x_{N_i})$$

Since the conditional distribution of X_i given all other variables is a BERNOULLI distribution, it is easy to sample it using a uniform random variable.

PROPOSITION III. .1. *Random scan* GIBBS *sampling satisfies detailed balance for* π *the* GIBBS *distribution of interest (i.e. the distribution of the graphical model).*

PROOF Let us consider one step of the random scan GIBBS sampling algorithm starting from π the distribution of the graphical model. The idea is to prove the reversibility.

We first prove the result for an index i fixed, that is we prove that the transition $q_{i,\text{GIBBS}}(x^{(t+1)} \mid x^{(t)})$ that only resamples the ith coordinate of $x^{(t)}$ is reversible for π.

We write $p_\pi(x_i \mid x_{-i})$ the conditional distribution $p_\pi(x_i \mid x_{-i}) = \frac{\pi(x_i, x_{-i})}{\pi(x_i)}$ of the GIBBS distribution π. We have:

$$\begin{aligned}
\pi(x^{(t)}) q_{i,\text{GIBBS}}(x^{(t+1)} \mid x^{(t)}) &= \pi(x^{(t)}) \delta_{x_{-i}^{(t+1)}, x_{-i}^{(t)}} p_\pi(x_i^{(t+1)} \mid x_{-i}^{(t)}) \\
&= \pi(x_{-i}^{(t)}) p_\pi(x_i^{(t)} \mid x_{-i}^{(t)}) \delta_{x_{-i}^{(t+1)}, x_{-i}^{(t)}} p_\pi(x_i^{(t+1)} \mid x_{-i}^{(t)}) \\
&= \pi(x_{-i}^{(t+1)}) p_\pi(x_i^{(t)} \mid x_{-i}^{(t+1)}) \delta_{x_{-i}^{(t)}, x_{-i}^{(t+1)}} p_\pi(x_i^{(t+1)} \mid x_{-i}^{(t+1)}) \\
&= \pi(x^{(t+1)}) q_{i,\text{GIBBS}}(x^{(t)} \mid x^{(t+1)})
\end{aligned}$$

and detailed balance for $q_{i,\text{GIBBS}}$ is valid for any i. In the random scan case, the index i being chosen at random uniformly with probability $\frac{1}{n}$, the GIBBS transition is in fact:

$$\frac{1}{d} \sum_{i=1}^{d} q_{i,\text{GIBBS}}$$

The result is then obtained by taking the average over i in the previous derivation. Thus π is a stationary distribution of the random scan GIBBS transition. □

PROPOSITION III. .2. *If the* GIBBS *distribution* p *satisfies* $p(x) > 0$ *for all* $x \in \mathcal{X}$*, the* MARKOV *chain defined by the* GIBBS *sampling algorithm (cyclic or random) converges in distribution to* π*.*

EXERCICE III. .1. Extend GIBBS method to POTTS model.

EXERCICE III. .2. Prove that the GIBBS transition is a special case of METROPOLIS-HASTINGS proposal that is always accepted.

CHAPTER 12

Variational inference

I. Overview

The goal is to do approximate inference without using sampling. Indeed, algorithms such as METROPOLIS-HASTINGS or GIBBS sampling can be very slow to converge, besides in practice it is very difficult to find a good stopping criterion. People working on MCMC methods try to find clever tricks to speed up the process, hence the motivation for variational methods.

Let us consider a distribution on $\mathcal{X}$ finite (but usually very large) and Q an exponential family with $q_\eta(x) = \exp(\eta^\top \phi(x) - A(\eta))$. Let us assume that the distribution of interest p, that is for example the distribution of our graphical model that we are working with, is in Q. The goal is to compute $\mathbb{E}_p[\phi(x)]$.

Computing this expectation corresponds to probabilistic inference in general. For example, for the POTTS model, we have

$$\phi(x) = ((x_{ik})_{i\in V, 1\leq k\leq K}, (x_{ik}x_{jk'})_{(i,j)\in E, 1\leq k,k'\leq K})^\top$$

We recall that $p = \operatorname{argmin}_q D(q \,\|\, p)$ where:

$$D(q \,\|\, p) = \sum_{x\in\mathcal{X}} q(x) \log \frac{q(x)}{p(x)} = \mathbb{E}_q[-\log p(X)] - H(q)$$

Since $p \in Q$, it is associated with a parameter η thus:

$$\mathbb{E}_q[-\log p(X)] = \mathbb{E}_q[-\eta^\top \phi(X) + A(\eta)] = -\eta^\top \mathbb{E}_q[\phi(X)] + A(\eta) = -\eta^\top \mu(q) + A(\eta)$$

where $\mu(q)$ is the moment parameter. Thus we have:

$$-D(p \,\|\, q) = \eta^\top \mu(q) + H(q) - A(\eta)$$

which is a nonpositive guantity thus, $A(\eta) \geq \eta^\top \mu(q) + H(q)$ for all q. Maximizing with respect to q in the exponential family leads to:

$$A(\eta) = \max_{q\in Q} \eta^\top \mu(q) + H(q)$$

and the unique value of q that attains the maximum is p.

REMARK I. .1. It is possible here to get rid of q and express things only in terms of the moment. It is indeed a way to parameterize the distribution q : for a realizable μ in the exponential family there is a single associated distribution q_μ. The maximization problem becomes:

$$\max_{\mu \in \mathcal{M}} \eta^\top \mu + \tilde{H}(\mu)$$

where $\tilde{H}(\mu) = H(q_\mu)$ and where $\mathcal{M}$ is called the marginal polytope and is the set of all possible moments[a]. The maximum is only attained for $\mu^* = \mu(p) = \mathbb{E}_p[\phi(X)]$, which is exactly the expectation that needs to be computed.

It turns out that it is possible to show that $\tilde{H}$ is always a concave function, so that the optimization problem above is a convex optimization problem.

It is interesting to note that we have thus turned the probabilistic inference problem, which, a priori, required to compute expectations, that is integrals, into an optimization problem, which is furthermore convex. Unfortunately this convex optimization problem is NP-hard to solve in general because it solves the NP-hard probabilistic inference problem, and it is not possible to escape the fact that the latter is NP-hard. This optimization problem is thus in general intractable and this is because of two reasons:

- for a general graph the marginal polytope $\mathcal{M}$ has number of faces which is exponential in the tree width of the graph.
- $\tilde{H}$ can be extremely complicated to write explicitly.

[a] We have seen in the course on exponential families that the distribution of maximum entropy q under the moment constraint $\mathbb{E}_q[\phi(X)] = \mu$ is also, when it exists, the distribution of maximum likelihood in the exponential family associated with the sufficient statistic ϕ. This essentially – but not exactly – shows that for any moment μ there exists a member q of the exponential family such that $\mu = \mu(q)$. In fact, to be rigorous one has to be careful about what happens at points of the boundary of the set $\mathcal{M}$: the above statement is correct for μ in the interior of $\mathcal{M}$. The points on the boundary of $\mathcal{M}$ are only corresponding to limits of distributions of the exponential family that can be degenerate, like the BERNOULLIdistribution with probability 1 (or 0) for example in the BERNOULLIfamily case, which are themselves not in the family

II. Mean field

In order to approximate the optimization problem it is possible either to change the set of distribution Q, the moments $\mathcal{M}$ or to change the definition of the entropy $\tilde{H}$. The mean field technique consists in choosing q in a set that makes all variables independent.

For a graphical model on variables $x_1, \dots x_n$, let us consider:

$$Q_{\text{indep}} = \left\{ q \,|\, \forall x, q(x) = \prod_{i=1}^{n} q_i(x_i) \right\}$$

the collection of distributions that make the variables $X_1, \dots, X_n$ independent.

We consider the optimization problem:

$$\max_{q \in Q_{\text{indep}}} \eta^\top \mu(q) + H(q)$$

Note that in general $p \notin Q_{\text{indep}}$ so that the solution cannot be exactly $\mu(p)$.

In order to write this optimization problem for the POTTS model, we need to write explicitly $\eta^\top \mu(q)$ and $H(q)$.

Moments in the mean field formulation Let $q \in Q_{\text{indep}}$. We have:

$$\eta^\top \mu(q) = \eta^\top \mathbb{E}_q[\phi(X)] = \sum_{i \in V, 1 \leq k \leq K} \eta_{ik} \mathbb{E}_q[X_{ik}] + \sum_{(i,j) \in E, 1 \leq k,k' \leq K} \eta_{ijkk'} \mathbb{E}_q[X_{ik} X_{jk'}]$$

By independance of the variables

$$\mathbb{E}_q[X_{ik}] = \mathbb{E}_{q_i}[X_{ik}] = \mu_{ik}(q) \qquad \text{and} \qquad \mathbb{E}_q[X_{ik} X_{jk'}] = \mathbb{E}_{q_i}[X_{ik}] \mathbb{E}_{q_j}[X_{j\ell}] = \mu_{ik} \mu_{jk'}$$

Note that if we had not constrained q to make these variables independent, we would in general have a moment here of the form $\mathbb{E}_q[X_{ik} X_{jk'}] = \mu_{ijkk'}(q)$. This is the main place where the mean field approximation departs from the exact variational formulation.

Entropy $H(q)$ in the mean field formulation By independence of the variables one has $H(q) = \sum_{i=1}^n H(q_i)$. Recall that q_i is the distribution on a single node, and that X_i is a multinomial random variable, one has:

$$H(q_i) = -\sum_{k=1}^K q_i(X_{ik} = 1) \log q_i(X_{ik} = 1) = -\sum_{k=1}^K \mu_{ik} \log \mu_{ik}$$

Mean field formulation for the POTTS model In the end, putting everything together the optimization problem can be written as

$$\begin{array}{ll} \max_\mu & \sum_{i \in V, 1 \leq k \leq K} \eta_{ik} \mu_{ik} + \sum_{(i,j) \in E, 1 \leq k,k' \leq K} \eta_{ijkk'} \mu_{ik} \mu_{jk'} - \sum_{i \in V, 1 \leq k \leq K} \mu_{ik} \log \mu_{ik} \\ \text{subject to} & \mu \geq 0, \quad \forall i \in V, \sum_{k=1}^K \mu_{ik} = 1 \end{array}$$

The problem is simple to express, however we cannot longer expect that it will solve our original problem, because by restricting to the set Q_{indep}, we have restrained the forms that the moment parameters $\mu_{ijkk'} = \mathbb{E}[X_{ik} X_{jk'}]$ can take. In particular since $p \notin Q_{\text{indep}}$ in general, the optimal solution of the mean field formulation does not retrieve the correct moment parameter $\mu(p)$. The approximation will be reasonable if $\mu(p)$ is not too far from the sets of moments that are achievable by moments of distributions in Q_{indep}, since the moments of p are approximated by the moments of the closest independent distribution. Note however that the mean field approximation is much more subtle than ignoring the binary potentials in the model, which would be a too naive way of finding an "approximation" with an independent distribution.

One difficulty though is that the objective function is no longer concave, because of the products $\mu_{ik} \mu_{jk'}$ which arise because of the independence assumption from the mean field approximation. Coordinate descent on each of the μ_i (not the μ_{ik}) is an algorithm of choice to solve this kind of problem. To present the algorithm we consider the case of the ISING model, which is a special case of the POTTS model with 2 states for each variable.

Mean field formulation for the Ising model When working with the Ising model is simple to reduce the number of variables by using the fact that $\mu_{i2} = 1 - \mu_{i1}$, we therefore write μ_i for μ_{i1} and the mean field optimization problem becomes:

$$\begin{array}{ll} [???] \max_\mu & \sum_i \eta_i \mu_i + \sum_{i,j} \eta_{ij} \mu_i \mu_j - \sum_i \Big(\mu_i \log \mu_i + (1-\mu_i) \log(1-\mu_i) \Big) \\ \text{subject to} & \mu \in [0,1]^n \end{array}$$

The stationary points for each coordinate correspond to the zeros of the partial derivatives. As

$$\frac{\partial f}{\partial \mu_i} = \eta_i + \sum_{j \in N_i} \eta_{ij} \mu_j - \log \frac{\mu_i}{1-\mu_i}$$

we obtain

$$\frac{\partial f}{\partial \mu_i} = 0 \quad \iff \quad \log \mu_i / (1-\mu_i) = \eta_i + \sum_{j \in N_i} \eta_{ij} \mu_j \quad \iff \quad \mu_i^* = \sigma\Big(\eta_i + \sum_{j \in N_i} \eta_{ij} \mu_j \Big)$$

where σ is the logistic function.

Note that in Gibbs sampling $x_i^{(t+1)} = 1$ with probability $\sigma(\eta_i + \sum_{j \in N_i} \eta_{ij} x_j)$. This is called mean field because the sampling is replaced by an approximation where it is assumed that the sample value is equal to its expectation, which for the physicists correspond to the mean field in the ferromagnetic Ising model.

Finally, let us insist that the mean field formulation is only one of the formulations for variational inference, there are several other ones, among which structured mean field, expectation propagation, loopy belief propagation (which can be reinterpreted as a solving variational formulation as well), tree-reweighted variational inference, …

CHAPTER 13

Model Selection

[todo]

I. Model Selection

I. A. Introduction

Let's consider two models M_1, M_2 such that $M_1 \subset M_2$ and $\Theta_1 \subset \Theta_2$. We define for $i \in [\![1, 2]\!]$:

$$\hat{\theta}_{M_i} = \text{argmax}_{\theta \in \Theta_i} \log p_\theta(x_{1:n})$$

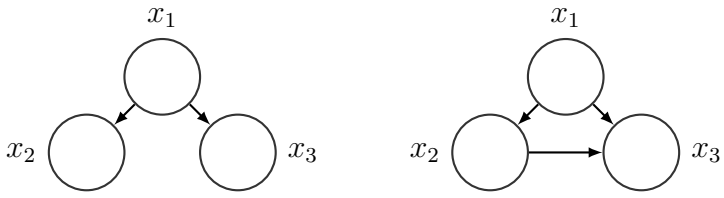


Figure 13.1: Example of Model Section for $n = 2$ (M_1 on the l.h.s and M_2 on the r.h.s)

We want to select the best model. For this, we need to define some kind of model score. We can't use the maximum likelihood as a score since we have by definition:

$$\log p_{\hat{\theta}_{M_2}} \geq \log p_{\hat{\theta}_{M_1}}$$

We are interested in the capacity of the generalisation of the model: we'd like to avoid over-fitting. Commonly, one way of dealing with that task is to select the size of the model by cross-validation. Here, we'll not develop it furthermore.

In this part we present the BAYES factors, which give us the main BAYES principal for selecting models. Also we will show the link with the penalised version BIC (Bayesian Information Criterion) which is used by the frequentists so as to "correct" the maximum likelihood and which has good proprieties. The issue with the selection model task is the issue with the selection of the variables which are an active topic of research. Note that there are others ways of penalising the maximum likelihood and of selecting models.

If p_0 is the distribution of the real data, we wish to choose between different models $(M_i)_{i \in I}$ by maximising $\mathbb{E}_{p_0}[\log p_{M_i}(X^* \mid D)]$ where X^* is a new test sample distributed as p_0 (in fact, it is still the maximum likelihood principle but we take the expectation on new data).

In the Bayesian framework, we can compute the marginal probability of data for a given model

$$\int p(x_{1:n} \mid \theta) p(\theta \mid M_i) d\theta = p(D \mid M_i)$$

and, by applying the BAYES rule, compute the a posteriori probability of the model:

$$p(M_i \mid D) = \frac{p(D \mid M_i) p(M_i)}{p(D)}$$

I. B. BAYES factor

Let's introduce the BAYES factors, which enable us to compare two models:

$$\frac{p(M_i \mid D)}{p(M_j \mid D)} = \frac{p(D \mid M_i) p(M_j)}{p(D \mid M_j) p(M_j)}$$

The marginal probability of data $p(D \mid M) = p(x_{1:n} \mid M)$ can decompose itself in a sequential way by using:

$$p(x_n \mid x_{1:n-1}, M) = \int p(x_n \mid \theta) p(\theta \mid x_{1:n-1}, M) d\theta$$

Indeed, we get:

$$p(D \mid M) = p(x_n \mid x_{1:n-1}, M) p(x_{n-1} \mid x_{1:n-2}, M) \dots p(x_1 \mid M)$$

such as

$$\frac{1}{n} \log p(D \mid M) = \frac{1}{n} \sum_{i=1}^{n} \log p(x_i \mid x_{1:i-1}, M) \simeq \mathbb{E}_{p_0}[\log p_M(X \mid D)]$$

I. C. Bayesian Information Criterion

The Bayesian score is approximated by the BIC:

$$\log p(D \mid M) = \log p_{\hat{\theta}_{\text{MV}}}(D) - \frac{K}{2} \log(n) + O(1)$$

where $p_{\hat{\theta}_{\text{MV}}}(D)$ is the data's distribution when the parameter is the maximum likelihood estimator $\hat{\theta}_{\text{MV}}$, K is the number of parameters of the model and n the number of observations.

In the following section, we outline the proof of this result in the case of an exponential family given by $p(x \mid \theta) = \exp(\langle \theta, \phi(X) \rangle - A(\theta))$.

I. D. LAPLACE's method

$$\begin{aligned} p\left(D \mid M\right) &= \int \prod_{i=1}^{n} p\left(x_i \mid \theta\right) p\left(\theta\right) d\theta \\ &= \int \exp\left(\left\langle \theta, n\overline{\phi} \right\rangle - nA\left(\theta\right)\right) p\left(\theta\right) d\theta \end{aligned}$$

$$\begin{aligned} \langle \theta, n\overline{\phi} \rangle - nA(\theta) = \langle \widehat{\theta}, n\overline{\phi} \rangle - nA(\widehat{\theta}) + \langle \theta - \widehat{\theta}, n\overline{\phi} \rangle \\ - n(\theta - \widehat{\theta})^T \nabla_\theta A(\widehat{\theta}) - \frac{1}{2}(\theta - \widehat{\theta})^T n \nabla_\theta^2 A(\widehat{\theta})(\theta - \widehat{\theta}) \\ + \mathrm{R}_n \end{aligned}$$

where R_n is a negligible rest.

But the maximum likelihood is the dual of the maximum entropy: $\max H(p_\theta)$ such that $\mu(\theta) = \overline{\phi}$.

$$\mu(\widehat{\theta}) = \overline{\phi}$$

$$p(D \mid M) \simeq \exp(\langle \widehat{\theta}, n\overline{\phi} \rangle - nA(\widehat{\theta})) \times \int \exp\left(-\frac{1}{2}(\theta - \widehat{\theta})^T n\widehat{\Sigma}(\theta - \widehat{\theta})\right) p(\theta) d\theta$$

However:

1. the information of fisher is equal to $\widehat{\Sigma}^{-1}$
2. $\int \exp\left(-\frac{1}{2}\left(\theta - \widehat{\theta}\right)^T n\widehat{\Sigma}\left(\theta - \widehat{\theta}\right)\right) p\left(\theta\right) d\theta \simeq c\sqrt{\left(2\pi\right)^k \left|\frac{\widehat{\Sigma}^{-1}}{n}\right|}$

Thus:

$$\begin{aligned} \log p\left(D \mid M\right) &= \log p_{\widehat{\theta}}\left(X\right) + \frac{1}{2}\log\left(\left(2\pi\right)^k \left|\frac{\widehat{\Sigma}^{-1}}{n}\right|\right) \\ &= \log p_{\widehat{\theta}}\left(X\right) + \frac{k}{2}\log\left(2\pi\right) + \frac{1}{2}\log\left(\left(\frac{1}{n}\right)^k \left|\widehat{\Sigma}^{-1}\right|\right) \\ &= \log p_{\widehat{\theta}}\left(X\right) + \frac{k}{2}\log\left(2\pi\right) - \frac{k}{2}\log\left(n\right) + \frac{1}{2}\log\left(\left|\widehat{\Sigma}^{-1}\right|\right) \end{aligned}$$

The main reason why presenting the BIC is that a theorem prove the consistency of the BIC. In other words, when the number of observations is sufficient, thanks to this criterion we choose with a probability that converges to 0, a model that satisfies:

$$M_k \in \mathrm{Argmax}_M \mathbb{E}_{p_0}\left[\log\left(p_{\widehat{\theta}_{MV}}\left(X; M\right)\right)\right]$$

To bring a quick clarification about the notations used in this part (model selection), please read below. The notation is a bit confusing (it was used for example in Bishop's book, but is a bit sloppy).

From the Bayesian perspective, we could treat the model choice as a random variable M. In the M_1 vs. M_2 vs. M_3 example, there are only 3 models, and thus M is a discrete variable with 3 possible values ($M = M_1$, $M = M_2$ or $M = M_3$).

Therefore, when we were writing quantities like the Bayes factor $p(M_1 \mid D)/p(M_2 \mid D)$, It really meant $p(M = M_1 \mid D)/p(M = M_2 \mid D)$. It did not mean that M_1 and M_2 were two different random variables which can take complicated values (someone asked what space M_1 was in and it seemed very complicated – what is meant is just that M is an index in possible (few) models).

D was the data random variable as usual. The mixing of random variables (here M) vs. their possible values ($M = 1, 2$ etc) in the same notation (like $p(M_1 \mid D)$) is usual but confusing; better to use the explicit $p(M = M_1 \mid D)$ notation to distinguish a value vs. a generic random variable. . . .

However, in general, M could be as complicated as we want. For example, it could be a vector of hyper-parameters for the prior distributions. Or it could also have binary component indicating the absence or presence of an edge in graphical model, etc. It does not have to just be an index. It could even be a continuous objects !

It is also fine to have infinite dimensional objects[1]. For example, consider the latent variable model: x is observed, θ and α are latent variables; and M decides the prior over α. I.e. suppose $p(x \mid \theta, \alpha, M) = Multi(\theta, 1)$, $p(\theta \mid \alpha, M) = Dir(\theta \mid \alpha)$, and $p(\alpha \mid M) = M(\alpha)$ i.e. M ranges over possible distributions over the positive vector α. M here is quite a complicated object, but this is fine. . .

II. Example of model

II. A. Bernoulli variable

Let's consider random variables $X_i \in \{0, 1\}$. We'll assume that the X_i are i.i.d. conditionally to θ. Then they follow a Bernoulli law:

$$p(x \mid \theta) = \theta^x (1 - \theta)^{1-x}$$

II. B. Priors

Let's introduce the *distribution* Beta whose density on $[0, 1]$ is

$$p(\theta; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}$$

Where $B(\alpha, \beta)$ is a short-name of the Beta *function*:

$$\forall \alpha > 0, \forall \beta > 0, B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1} (1 - \theta)^{\beta-1} \, d\theta$$

And the Gamma function:

$$\Gamma(x) = \int_0^{+\infty} t^{x-1} \exp(-t) \, dt$$

[1] This would be in the "non-parametric setting" – non-parametric = infinite dimensional.

We can show that $B(\alpha, \beta)$ is symmetric and satisfies:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

We choose as the prior distribution on θ the Beta distribution:

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$p(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

II. C. A posteriori

$$p(\theta \mid x) = \frac{p(x, \theta)}{p(x)} \propto p(x, \theta)$$

But:

$$p(x, \theta) = \theta^{x}(1-\theta)^{1-x}\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

Hence:

$$p(\theta \mid x) \propto \frac{\theta^{x+\alpha-1}(1-\theta)^{1-x+\beta-1}}{B(\alpha, \beta)}$$

$$p(\theta \mid x) = \frac{\theta^{x+\alpha-1}(1-\theta)^{1-x+\beta-1}}{B(x+\alpha, 1-x+\beta)}$$

Thus, if instead of considering a unique variable , we observe an i.i.d. sample of data, the joint distribution can be written as:

$$\theta^{\alpha-1}(1-\theta)^{\beta-1}\prod_{i=1}^{n}\theta^{x_i}(1-\theta)^{1-x_i}.$$

Let's introduce:

$$k = \sum_{i=1}^{n} x_i$$

Then we get:

$$p(\theta \mid x_1, x_2, \ldots, x_n) = \frac{\theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}}{B(k+\alpha, n-k+\beta)}$$

III. Special case of the Beta distribution

We remind that:

$$\theta \sim Beta\left(\alpha, \beta\right)$$

For $\alpha = \beta = 1$, we get a uniform prior.

For $\alpha = \beta > 1$, we get a bell curve.

For $\alpha = \beta < 1$, we get a U curve.

$\mathbb{E}\left[\theta\right] = \frac{\alpha}{\alpha+\beta}$

$\mathbb{V}\left[\theta\right] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{\alpha}{(\alpha+\beta)} \times \frac{\beta}{(\alpha+\beta)} \times \frac{1}{(\alpha+\beta+1)}$

For $\alpha > 1$ and $\beta > 1$, we get the mode: $\frac{\alpha-1}{\alpha+\beta-2}$.

In the case, let's write D for the data:

$$\theta_{post} = \mathbb{E}\left[\theta \mid D\right] = \frac{\alpha + k}{\alpha + \beta + n} = \frac{\alpha}{(\alpha + \beta)} \times \frac{(\alpha + \beta)}{(\alpha + \beta + n)} + \frac{n}{(\alpha + \beta + n)} \times \frac{k}{n}$$

We can see that the a posteriori expectation of the parameter is a convex combination of the maximum likelihood estimator and the prior expectation. It converges asymptotically to the maximum likelihood estimator .

If we use a uniform prior distribution, $\mathbb{E}\left[\theta \mid D\right] = \frac{k+1}{n+2}$. Laplace proposed to correct the frequentist estimator, it seemed odd to him that he was not defined in the absence of data. He proposed to add two virtual observation (0 and 1) such that in the absence of data the estimator equals $\frac{1}{2}$. This correction is known as *Laplace's correction*.

The variance of the a posteriori distribution decrease in $\frac{1}{n}$.

$$\mathbb{V}\left[\theta \mid D\right] = \theta_M \left(1 - \theta_M\right) \frac{1}{(\alpha + \beta + n)}$$

We have chosen a sharper distribution around θ_M, in the same way than in a frequentist approach, the confidence intervals narrow around the estimator when the number of observations increase.

III. A. Playful propriety

$$p\left(x_1, x_2, \ldots, x_n\right) = \frac{B\left(k + \alpha, n - k + \beta\right)}{B\left(\alpha, \beta\right)} = \frac{\Gamma\left(\alpha + k\right)\Gamma\left(\beta + n - k\right)\Gamma\left(\alpha + \beta\right)}{\Gamma\left(\alpha + \beta + n\right)\Gamma\left(\alpha\right)\Gamma\left(\beta\right)} \tag{13.1}$$

Let's use this well-known property of the Gamma function:

$$\Gamma\left(n + 1\right) = n!$$

$$\text{and} \qquad \forall x > -1, \Gamma(x+1) = x\Gamma(x)$$

such that

$$\Gamma(\alpha + k) = (\alpha + k - 1)(\alpha + k - 2)\dots\alpha\Gamma(\alpha)$$

let's write $\alpha^{[k]} = \alpha(\alpha+1)\dots(\alpha+k-1)$ and simplify the expression 13.1:

$$p(x_1, x_2, \dots, x_n) = \frac{\alpha^{[k]}\beta^{[n-k]}}{(\alpha+\beta)^{[n]}}$$

We shall note the analogy with the Polya urn model: let us consider $(\alpha+\beta)$ balls of colour: α are black, β are white. When drawing a first black ball, the probability of the event is:

$$\mathbb{P}(X_1 = 1) = \frac{\alpha}{\alpha+\beta}$$

After the drawing, we put back the ball in the urn and we add a ball of the same colour. Let's imagine that we draw again a black ball then the probability of this event is:

$$\mathbb{P}(X_1 = 1, X_2 = 1) = \mathbb{P}(X_1 = 1)\,\mathbb{P}(X_2 = 1 \mid X_1 = 1) = \frac{\alpha}{\alpha+\beta} \times \frac{\alpha+1}{\alpha+\beta+1}$$

However:

$$\mathbb{P}(X_1 = 1, X_2 = 0) = \frac{\alpha}{\alpha+\beta} \times \frac{\beta}{\alpha+\beta+1}$$

In more general case , we show by recurrence that the marginal probability of obtaining some sequence of colours by drawing from a Polya urn is exactly the marginal probability of obtaining the same result from the marginal model, obtained by integrating on a priori $theta$. First, this show that drawings from a Polya urn are exchangeable; Secondly, the mechanism of this type of urn, and its exchangeability, we'll be useful for the Gibbs sampling and for the same type of Bayesian models.

III. B. Conjugate priors

Let $\mathbb{F}$ be a set. We assume that $p(x \mid \theta)$ known, we deduce from that: $p(\theta) \in \mathbb{F}$ such that $p(\theta \mid x) \in \mathbb{F}$. We say that $p(\theta)$ is conjugated to the model $p(x \mid \theta)$.

a. Exponential model

Let's consider:

$$p(x \mid \theta) = \exp(\langle\theta, \phi(x)\rangle - A(\theta))$$
$$p(\theta) = \exp(\langle\alpha, \theta\rangle - \tau A(\theta) - B(\alpha, \tau))$$

For $p(x \mid \theta)$, θ is the canonical parameter. For $p(\theta)$, α is the canonical parameter and θ is the sufficient statistic. Let us note that B do not stand for the Beta distribution.

$$p(\theta \mid x) \propto p(x \mid \theta) p(\theta) \propto \exp\left(\langle \theta, \phi(x) \rangle - A(\theta) + \langle \alpha, \theta \rangle - \tau A(\theta) - B(\alpha, \tau)\right)$$

Let us define:

$$\overline{\phi} = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i)$$

Then:

$$p(\theta \mid x_i) \propto \exp\left(\langle \theta, \alpha + \phi(x_i) \rangle - (\tau + 1) A(\theta) - B(\alpha + \phi(x_i), \tau + 1)\right)$$

$$p(\theta \mid x_1, x_2, \ldots, x_n) \propto \exp\left(\left\langle \theta, \alpha + n\overline{\phi} \right\rangle - (\tau + n) A(\theta) - B\left(\alpha + n\overline{\phi}, \tau + n\right)\right)$$

$$p(x_1, x_2, \ldots, x_n) \propto \exp\left(B(\alpha, \tau) - B\left(\alpha + n\overline{\phi}, \tau + n\right)\right)$$

Since the family is an exponential one,

$$\nu_{post} = \mathbb{E}[\theta \mid D] = \nabla_\alpha B\left(\alpha + n\overline{\phi}, \tau + n\right)$$

θ_{MAP} results from:

$$\nabla_\theta p(\theta \mid x_1, x_2, \ldots, x_n) = 0$$
$$\alpha + n\overline{\phi} = (\tau + n) \nabla_\theta A(\theta) = (\tau + n) \mu(\theta)$$

Thus we get $\mu_{MAP} = \mu(\theta)$ in the previous equation. Consequently:

$$\mu_{MAP} = \frac{\alpha + n\overline{\phi}}{\tau + n} = \frac{\alpha}{\tau} \times \frac{\tau}{\tau + n} + \frac{n}{\tau + n} \overline{\phi}$$

b. Univariate Gaussian

i. With and a priori on μ but not on σ^2

$$p\left(x \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2}\right)$$

$$p\left(\mu \mid \mu_0, \tau^2\right) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2} \frac{(\mu - \mu_0)^2}{\tau^2}\right)$$

Thus:

$$\begin{aligned} p\left(D \mid \mu, \sigma^2\right) &= p\left(x_1, x_2, \ldots, x_n \mid \mu, \sigma^2\right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{\left(x_i-\mu\right)^2}{\sigma^2}\right) \end{aligned}$$

$$\begin{aligned} p\left(\mu \mid D\right) &= p\left(\mu \mid x_1, x_2, \ldots, x_n\right) \\ &= \exp\left(-\frac{1}{2}\left(\frac{\left(\mu-\mu_0\right)^2}{\tau^2} + \sum_{i=1}^{n}\frac{\left(x_i-\mu\right)^2}{\sigma^2}\right)\right) \\ &= \exp\left(-\frac{1}{2}\left(\frac{\mu^2-2\mu\mu_0+\mu_0^2}{\tau^2} + \sum_{i=1}^{n}\frac{\mu^2-2\mu x_i+x_i^2}{\sigma^2}\right)\right) \\ &= \exp\left(-\frac{1}{2}\left(\mu^2\Lambda - 2\mu\eta + \left(\frac{\mu_0^2}{\tau^2} + \sum_{i=1}^{n}\frac{x_i^2}{\sigma^2}\right)\right)\right) \end{aligned}$$

Where:

$$\begin{aligned} \Lambda &= \frac{1}{\tau^2} + \frac{n}{\sigma^2} \\ \eta &= \frac{\mu_0}{\tau^2} + \frac{n\overline{x}}{\sigma^2} \\ \overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \end{aligned}$$

Thus:

$$\begin{aligned} \mu_{post} &= \mathbb{E}\left[\mu \mid D\right] \\ &= \frac{\eta}{\Lambda} \\ &= \frac{\frac{\mu_0}{\tau^2} + \frac{n\overline{x}}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}} \\ &= \frac{\sigma^2\mu_0 + n\tau^2\overline{x}}{\sigma^2 + n\tau^2} \\ &= \frac{\sigma^2}{\sigma^2+n\tau^2}\mu_0 + \frac{n\tau^2}{\sigma^2+n\tau^2}\overline{x} \end{aligned}$$

And:

$$\begin{aligned} \widehat{\Sigma}_{post}^2 &= \mathbb{V}\left[\mu \mid D\right] \\ &= \frac{1}{\Lambda} \\ &= \frac{\sigma^2\tau^2}{\sigma^2+n\tau^2} \end{aligned}$$

Indeed, the variance decreases in $\frac{1}{n}$.

ii. With an a priori on σ^2 but not onμ We get $p(\sigma^2)$ as an Inverse Gamma form.

iii. With an a priori on μ and σ^2 Gaussian a priori on x and μ, Inverse Gamma a priori on σ^2. Please refer to the chapter 9 of the course handout (Jordan's polycopié).

IV. A posteriori Maximum (MAP)

$$\begin{aligned}\theta_{MAP} &= \arg\max_{\theta} p(\theta \mid x_1, x_2, \ldots, x_n) \\ &= \arg\max_{\theta} p(x_1, x_2, \ldots, x_n \mid \theta)\, p(\theta)\end{aligned}$$

Because, with the Bayes rule:

$$p(\theta \mid x_1, x_2, \ldots, x_n) = \frac{p(x_1, x_2, \ldots, x_n \mid \theta)\, p(\theta)}{p(x)}$$

The a posteriori maximum is not really Bayesian, it's rather a slight modification brought to the frequentist estimator.

IV. A. Predictive probability

In the Bayesian paradigm, the probability of a future observation x^* will be estimated by the *Predictive probability*:

$$\begin{aligned}p(x^* \mid D) &= p(x^* \mid x_1, x_2, \ldots, x_n) \\ &= \int p(x^* \mid \theta)\, p(\theta \mid x_1, x_2, \ldots, x_n)\, d\theta\end{aligned}$$

$$\begin{aligned}p(\theta \mid x_1, x_2, \ldots, x_n) &\propto p(x_n \mid \theta)\, p(x_1 \mid \theta)\, p(x_2 \mid \theta) \ldots p(x_{n-1} \mid \theta)\, p(\theta) \\ &\propto p(x_n \mid \theta)\, p(\theta \mid x_1, x_2, \ldots, x_{n-1})\, p(x_1, x_2, \ldots, x_{n-1}) \\ &\propto p(x_n \mid \theta)\, p(\theta \mid x_1, x_2, \ldots, x_{n-1})\, \frac{p(x_1, x_2, \ldots, x_{n-1})}{p(x_1, x_2, \ldots, x_n)}\end{aligned}$$

A sequential calculus is possible since:

$$p(\theta \mid x_1, x_2, \ldots, x_n) = \frac{p(x_n \mid \theta)\, p(\theta \mid x_1, x_2, \ldots, x_{n-1})}{p(x_n \mid x_1, x_2, \ldots, x_{n-1})}$$

Vocabulary:

- a priori information: $p(\theta \mid x_1, x_2, \ldots, x_{n-1})$
- likelihood: $p(x_n \mid \theta)$
- a posteriori information: $p(\theta \mid x_1, x_2, \ldots, x_n)$

$$p(x_1, x_2, \ldots, x_n) = \int \prod_{i=1}^{n} p(x_i \mid \theta)\, p(\theta)\, d\theta$$

V. Naive Bayes

V. A. Introduction

Remarque: Contrary to its name, "Naive Bayes" is *not* a Bayesian method.

Let's Consider the following problem of classification $x \in \mathbb{X}^p \longmapsto y \in \{1, 2, \ldots, M\}$.

Here, $x = (x_1, x_2, \ldots, x_p)$ is a vector of descriptors (or features): $\forall i \in \{1, 2, \ldots, p\}, x_i \in \mathbb{X}$, with $\mathbb{X} = \{1, 2, \ldots, K\}$ (or $\mathbb{X} = \mathbb{R}$).

Goal: Learn $p(y \mid x)$.

A very naive method will trigger off a combinatorial explosion: $\theta \in \mathbb{R}^{Kp}$.

Bayes formula gets us:

$$p(y \mid x) = \frac{p(x \mid y)\, p(y)}{p(x)}$$

The Naive Bayes method consists in assuming that the features x_i are all conditionally independent from the class, hence:

$$p(x \mid y) = \prod_{i=1}^{p} p(x_i \mid y)$$

Then, the Bayes formula gives us:

$$p(y \mid x) = \frac{p(y) \prod_{i=1}^{p} p(x_i \mid y)}{p(x)} = \frac{p(y) \prod_{i=1}^{p} p(x_i \mid y)}{\sum_{y'} p(y') \prod_{i=1}^{p} p(x_i \mid y')}$$

We consider the case where the features take discrete values. Consequently the new graphical model contains only discrete random variables. Then, we can write a discrete model as an exponential family. Indeed we can write:

$$\log p(x_i = k \mid y = k') = \delta(x_i = k, y = k')\, \theta_{ikk'}$$

and

$$\log p\left(y = k'\right) = \delta\left(y = k'\right)\theta_{k'}$$

We can see that the dummy functions $\delta(x_i = k, y = k')$ and $\delta(y = k')$ are the *sufficient statistics* of the joint distribution model for y and the variables x_i, where $\theta_{ikk'}$ and $\theta_{k'}$ are *canonical parameters*. Thus , we can write:

$$\log p(y, x_1, \ldots, x_p) = \sum_{i,k,k'} \delta(x_i = k, y = k')\theta_{ikk'} + \sum_{k'} \delta(y = k')\theta_{k'} - A((\theta_{ikk'})_{i,k,k'}, (\theta_{k'})_{k'})$$

Where $A((\theta_{ikk'})_{i,k,k'}, (\theta_{k'})_{k'})$ is the log-partition function.

We have rewritten the joint distribution model of $(y, x_1, \ldots, x_p)$ as an exponential family. Given that the maximum of likelihood estimator of an exponential family, where the canonical parameters are not combined, is also the maximum entropy estimator; as seen in a previous course and provided that the statistical moments of the sufficient statistics equal their empirical moments.

Thus, if we introduce

$$N_{ikk'} = \#\left\{(x_i, y) = (k, k')\right\}$$
$$N = \sum_{i,k,k'} N_{ikk'},$$

The maximum likelihood estimator must satisfy the moment constraints

$$\widehat{p}\left(y = k'\right) = \frac{\sum\limits_{i,k} N_{ikk'}}{N} \qquad \text{et} \qquad \widehat{p}\left(x_i = k \,|\, y = k'\right) = \frac{N_{ikk'}}{\sum\limits_{k''} N_{ik''k'}},$$

which define them completely.

Then, we can write the estimators of the canonical parameters as:

$$\widehat{\theta}_{ikk'} = \log \widehat{p}\left(x_i = k \,|\, y = k'\right) \qquad \text{et} \qquad \widehat{\theta}_{k'} = \log \widehat{p}\left(y = k'\right).$$

However, our goal is to obtain a classification model, that is to say, a model of only the conditional probability law. From the approximated generative model and applying the Bayes rule we can get:

$$\log \widehat{p}\left(y = k' \,|\, x\right) = \sum_{i=1}^{p} \log \widehat{p}\left(x_i \,|\, y = k'\right) + \log \widehat{p}\left(y = k'\right) - \log \sum_{k'} \left(\widehat{p}\left(y = k'\right) \prod_{i=1}^{p} \widehat{p}\left(x_i \,|\, y = k'\right) \right)$$

We can re write the conditional model as an exponential family

$$\log p\left(y \,|\, x\right) = \sum_{i,k,k'} \delta(x_i = k, y = k')\theta_{ikk'} + \sum_{k'} \delta(y = k')\theta_{k'} - \log p(x)$$

Its sufficient statistics and canonical parameters are equal to those of the generative model, but seen as functions of the random variable y, given that x is fixed (we could write $\phi_{x,i,k,k'}(y) = \delta(x_i = k, y = k')$). As for the log-partition function, it is now equal to $\log p(x)$.

Warning: $\widehat{\theta}_{ikk'}$ is the maximum likelihood estimator in the generative model which, usually, is not equal to the maximum likelihood estimator in the conditional model.

V. B. Advantages and Drawbacks

Advantages:

- Doable in line.
- Computationally tractable solution.

Drawbacks:

- Generative: generative models produce good estimator whenever the model is "true", or in statistical words *well specified*, which means that the process that generate the real data induce a distribution equal to the one of the generative model. When the model is not *well specified* (which is the most common case) we'd better use a discriminative method.

V. C. Discriminative method

The problem that we have considered in the previous section is the generative model for classification in K classes. How to learn, in a discriminatory way , a classifier in K classes? Is it possible to use an exponential family?

We have already seen the logistic regression for 2 classes classification:

$$p\left(y = 1 \mid x\right) = \frac{\exp\left(\omega^T x\right)}{1 + \exp\left(\omega^T x\right)}$$

Let's study the K-multiclass logistic regression:

$$\begin{aligned}
p\left(y = k' \mid x\right) &= \frac{\exp\left(\sum_{i=1}^{p}\sum_{k=1}^{K}\delta\left(x_i = k\right)\theta_{ikk'}\right)}{\sum_{k''=1}^{M}\exp\left(\sum_{i=1}^{p}\sum_{k=1}^{K}\delta\left(x_i = k\right)\theta_{ikk''}\right)} \\
&= \exp\left(\sum_{i=1}^{p}\sum_{k=1}^{K}\delta\left(x_i = k\right)\theta_{ikk'} - \log\left(\sum_{k''=1}^{M}\exp\left(\sum_{i=1}^{p}\sum_{k=1}^{K}\delta\left(x_i = k\right)\theta_{ikk''}\right)\right)\right) \\
&= \exp\left(\theta_{k'}^T\phi\left(x\right) - \log\left(\sum_{k''=1}^{M}\exp\left(\theta_{k''}^T\phi\left(x\right)\right)\right)\right) \\
&= \frac{\exp\left(\theta_{k'}^T\phi\left(x\right)\right)}{\sum_{k''=1}^{M}\exp\left(\theta_{k''}^T\phi\left(x\right)\right)}
\end{aligned}$$

Although we have built the model from different staring consideration, the resulting modelling (that is the set of possible distribution) is of the same exponential family than the Naive Bayes model.

Nonetheless, the fitted model in a discriminatory approach will be different from the one fitted in a generative approach: the fitting of the K-multiclass logistic regression results from the maximisation of the likelihood of the classes $y^{(j)}$ of a set of learning, given that $x^{(j)}$ are fixed. In other words, the fitting is obtained by computing the maximum likelihood estimator in the conditional model. Unlike what happens in the generative model, the estimator can't be obtained in a analytical form and the learning requires solving a numerical optimisation problem.

Annex

I. Review on probabilities

In this section we recall some basic notations and properties of random variables.

> NOTATION I. .1. The probability that a random variable X takes the value x is denoted $p(X = x)$. In this document, we simply write $p(X)$ to denote a distribution over the random variable X, or $p(x)$ to denote the distribution evaluated for the particular value x. It is similar for more variables.

Fundamental rules For two random variables X, Y we have

- Sum rule:
$$p(X) = \sum_Y p(X, Y)$$
- Product rule:
$$p(X, Y) = p(Y \mid X)p(X)$$
- BAYES formula[2]:
$$p(X \mid Y) = \frac{p(Y \mid X)p(X)}{p(Y)}$$

I. A. Joint distributions

Let $X_1, X_2, \ldots, X_n$ be random variables with joint distribution $\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = p_X(x_1, \ldots, x_n) = p(x)$ where x stands for $(x_1, \ldots, x_n)$.

Given $A \subset [\![1, n]\!]$, we denote the marginal distribution of x_A by:
$$p(x_A) = \sum_{x_{A^c}} p(x_A, x_{A^c})$$
With this notation, we can write the conditional distribution as:
$$p(x_A \mid x_{A^c}) = \frac{p(x_A, x_{A^c})}{p(x_{A^c})}$$
We also recall the so-called "chain rule" stating:
$$p(x) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2) \ldots p(x_n \mid x_1, \ldots, x_{n-1})$$

[2] note that BAYES formula is not a Bayesian formula in the sense of Bayesian statistics

I. B. Independence and conditional independence

Let A, B, and C be disjoint subsets of $[\![1, n]\!]$. We will say that X_A is (marginally) independent of X_B and write $X_A \perp\!\!\!\perp X_B$ if

$$\forall (x_A, x_B), \quad p(x_A, x_B) = p(x_A)p(x_B) \qquad \text{or equivalently} \qquad p(x_B) > 0 \implies p(x_A \mid x_B) = p(x_A)$$

Similarly we will say that X_A is independent from X_B conditionally on X_C (or given X_C) and we will write $X_A \perp\!\!\!\perp X_B \mid X_C$ if

$$\forall x_A, x_B, x_C, \quad p(x_C) > 0 \implies p(x_A, x_B \mid x_C) = p(x_A \mid x_C)p(x_B \mid x_C)$$

or equivalently if

$$\forall x_A, x_B, x_C, \quad p(x_B, x_C) > 0 \implies p(x_A \mid x_B, x_C) = p(x_A \mid x_C)$$

More generally we will say that the $(X_{A_i})_{1 \leq i \leq k}$ are *mutually independent* if

$$\forall x_{A_1}, \ldots, x_{A_k}, \quad p(x_{A_1}, \ldots, x_{A_k}) = \prod_{i=1}^{k} p(x_{A_i})$$

and that they are *mutually independent conditionally on* X_C (or given X_C) if

$$\forall x_{A_1}, \ldots, x_{A_k}, x_C, \quad p(x_C) > 0 \implies p(x_{A_1}, \ldots, x_{A_k} \mid x_C) = \prod_{i=1}^{k} p(x_{A_i} \mid x_C)$$

REMARK I. .2. Note that the conditional probability $p(x_A, x_B \mid x_C)$ is the probability distribution over (X_A, X_B) if X_C is known to be equal to x_C. In practice, it means that if the value of X_C is *observed* (e.g. via a measurement) then the distribution over (X_A, X_B) is $p(x_A, x_B \mid x_C)$. The conditional independence statement $X_A \perp\!\!\!\perp X_B \mid X_C$ should therefore be interpreted as "when the value of X_C is observed (or given), X_A and X_B are independent".

REMARK I. .3. **[PAIRWISE INDEPENDENCE VS MUTUAL INDEPENDENCE]**
Consider a collection of random variables $(X_1, \ldots, X_n)$. We say that these variables are pairwise independent if $X_i \perp\!\!\!\perp X_j$ for all $1 \leq i < j \leq n$. Note that this is different than assuming that $X_1, \ldots, X_n$ are mutually (or jointly or globally) independent. A standard counter-example is as follows: given two variables X, Y that are independent coin flips define Z via the XOR function $\oplus$ with $Z = X \oplus Y$. Then, the three random variables X, Y, Z are *pairwise independent*, but not *mutually independent* (exercise). The notations presented for *pairwise independence* could be generalized to collections of variables that are *mutually independent*.

Three facts about conditional independence

- It is possible to repeat the conditional variable: $X \perp\!\!\!\perp (Y, Z) \mid Z, W$ is the same as $(X, Z) \perp\!\!\!\perp Y \mid Z, W$. The repetition is redundant but may be convenient notation.
- We have decomposition: if $X \perp\!\!\!\perp (Y, Z) \mid W$ then $X \perp\!\!\!\perp Y \mid W$ and $X \perp\!\!\!\perp Z \mid W$.
- The chain rule applies to conditional distributions:

$$p(x, y \mid z) = p(x \mid y, z)p(y \mid z)$$

Independent and identically distributed A set of random variables is independent and identically distributed (i.i.d.) if each random variable has the same probability distribution as the others and all are mutually independent.

II. Review on LAGRANGE duality

Lagrangian Consider the following convex optimization problem:

$$\begin{array}{ll} \min_{x\in\mathcal{X}} & f(x) \\ \text{subject to} & Ax = b \end{array}$$

where f is a convex function, $\mathcal{X} \subset \mathbb{R}^p$ is a convex set included in the domain[3] of f, $A \in \mathcal{M}_{n\,p}$ and $b \in \mathbb{R}^n$.

The *Lagrangian* associated with this optimization problem is defined as

$$\begin{array}{rccc} \mathcal{L}: & \mathcal{X} \times \mathbb{R}^n & \longrightarrow & \mathbb{R} \\ & x, \lambda & \longmapsto & f(x) + \lambda^\top (Ax - b) \end{array}$$

The vector λ is called the LAGRANGE multiplier vector.

Lagrange dual function The LAGRANGE dual function is defined as $g : \lambda \in \mathbb{R}^n \longmapsto \min_{x\,\in\mathcal{X}} \mathcal{L}(x, \lambda)$. The problem of maximizing g is known as the LAGRANGE dual problem.

max-min inequality For any $f : \mathcal{W} \times \mathcal{Z} \subset \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}$, we have

$$\begin{array}{rcl} \forall w \in \mathcal{W}, f(w,z) \leq \max\limits_{z\in\mathcal{Z}} f(w,z) & \Longrightarrow & \min\limits_{w\in\mathcal{W}} f(w,z) \leq \min\limits_{w\in\mathcal{W}} \max\limits_{z\in\mathcal{Z}} f(w,z) \\ & \Longrightarrow & \max\limits_{z\in\mathcal{Z}} \min\limits_{w\in\mathcal{W}} f(w,z) \leq \min\limits_{w\in\mathcal{W}} \max\limits_{z\in\mathcal{Z}} f(w,z) \end{array}$$

Duality It is easy to show that $\max_\lambda \mathcal{L}(x,\lambda) = \begin{cases} f(x) & \text{if } Ax = b \\ +\infty & \text{otherwise} \end{cases}$ which gives us $\min_{x\in\mathcal{X}} f(x) = \min_{x\in\mathcal{X}} \max_\lambda \mathcal{L}(x,\lambda)$. For the above equations we have:

$$\max_\lambda g(\lambda) = \max_\lambda \min_{x\in\mathcal{X}} \mathcal{L}(x,\lambda) \leq \min_{x\in\mathcal{X}} \max_\lambda \mathcal{L}(x,\lambda) = \min_{x\in\mathcal{X}} f(x)$$

This inequality says that the optimal value d^* of the LAGRANGE dual problem always lower-bounds the optimal value p^* of the original problem. This property is called the *weak duality*. If the equality $d^* = p^*$ holds, then we say that the *strong duality* holds. Strong duality means that the order of the minimization over $x \in \mathcal{X}$ and the maximization over λ can be switched without affecting the result.

[3] the domain of a function is the set on which the function is well-defined and finite

SLATER's constraint qualification lemma If there exists an x in the relative interior of $\mathcal{X} \cap \{Ax = b\}$ then strong duality holds.

Note that all the above notions and results are stated for the fixed problem introduced above. For a more general problem and more details about LAGRANGE duality, please refer to [**?**] (chapter 5).

III. Review on differentials

III. A. Generalities

Differentiable function A function f is differentiable at $x \in \mathbb{R}^d$ if there exists a linear form df_x such that:

$$\forall h \in \mathbb{R}^d, \quad f(x+h) = f(x) + df_x(h) + o(\|h\|)$$

Since $\mathbb{R}^d$ is a HILBERT space, we know in that case that there exists $g \in \mathbb{R}^d$ such that $df_x(h) = \langle g \mid h \rangle$. We call g the gradient of f and denote it by $\nabla f(x)$.

EXAMPLE III. .1.

- If $f : x \longmapsto a^\top x + b$ then we have $f(x+h) = f(x) + a^\top h$ and thus f is differentiable and $\nabla f(x) = a$.
- If $f : x \longmapsto \frac{1}{2} x^\top A x$ then we have:

  $$f(x+h) - f(x) = \frac{1}{2}(x+h)^\top A(x+h) - \frac{1}{2} x^\top A x = \frac{1}{2}(x^\top A h + h^\top A x) + o(\|h\|)$$

  The gradient is then $\nabla f(x) = \frac{1}{2}(Ax + A^\top x)$.

Composition of differentials If f and g are differentiable respectively at $g(x)$ and x, then $f \circ g$ is differentiable at x and:

$$d(f \circ g)_x(h) = df_{g(x)}(dg_x(h)) = df_{g(x)} \circ dg_x(h)$$

III. B. Some practical differentials

- Let $f : \mathcal{S}_d^{++}(\mathbb{R}) \longrightarrow \mathbb{R}, \Lambda \longmapsto \log(\det \Lambda)$. We have for $H \in \mathcal{S}_d^{++}(\mathbb{R})$:

  $\log(\det(\Lambda+H)) = \log(\det(\Lambda^{\frac{1}{2}}(I_d + \Lambda^{-\frac{1}{2}} H \Lambda^{-\frac{1}{2}})\Lambda^{-\frac{1}{2}})) = \log(\det \Lambda) + \log(\det(I_d + \Lambda^{-\frac{1}{2}} H \Lambda^{-\frac{1}{2}}))$

  As $\tilde{H} = \Lambda^{-\frac{1}{2}} H \Lambda^{-\frac{1}{2}}$ is symmetric, it it diagonalizable and if $(\lambda_i)_{1 \le i \le d}$ are its eigenvalues:

  $$\log(\det(I_d + \tilde{H})) = \sum_{i=1}^d \log(1+\lambda_j) = \sum_{i=1}^d \lambda_j + o\Big(\big\|\tilde{H}\big\|\Big) = \mathrm{Tr}(\tilde{H}) + o(\|H\|)$$

  Thus f is differentiable at Λ and:

  $$df_\Lambda(H) = \mathrm{Tr}(\tilde{H}) = \mathrm{Tr}(H\Lambda^{-1}) \qquad \nabla f(\Lambda) = \Lambda^{-1}$$

- Let $f : \Lambda \longmapsto \mathrm{Tr}(\Lambda A)$ where A is a fixed symmetric matrix. We have:

  $$f(\Lambda + H) - f(\Lambda) = \mathrm{Tr}(HA)$$

  Thus f is differentiable is $\nabla f(\Lambda) = A$.

IV. Optimization methods

IV. A. First-order methods

Let $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ be the convex $\mathcal{C}^1$ function that we want to minimize. A *descent direction* at point x is a vector d such that $\langle d \mid \nabla f(x) \rangle < 0$. The minimization of f can be done by applying a *descent algorithm*, which iteratively takes a step in a descent direction, leading to an iterative scheme of the form

$$x^{t+1} = x^t + \varepsilon^t d^t$$

where ε^t is the *stepsize*. The direction d^t is often chosen as the opposite of the gradient of f at point x^t:

$$d^t = -\nabla f(x^t)$$

There are several choices for ε^t:

- a constant step: $\varepsilon^t = \varepsilon$. But the scheme does not necessarily converge,
- a decreasing step size: $\varepsilon^t \propto \frac{1}{k}$ with $\sum_k \varepsilon^t = +\infty$ and $\sum_k (\varepsilon^t)^2 < +\infty$. In that case the scheme is guaranteed to converge.
- one can determine ε^t by doing a *line search* which tries to find $\min_{\varepsilon > 0} f(x^t + \varepsilon d^t)$:
  - either exactly but this is costly and rather useless in many situations,
  - or approximately (ARMIJO line search). This is a very effective method.

IV. B. Second-order methods

Assume now that f is a $\mathcal{C}^2$ function. We write the second-order TAYLOR-expansion of f at a point x^t:

$$f(x) = \underbrace{f(x^t) + (x - x^t)^\top \nabla f(x^t) + \frac{1}{2}(x - x^t)^\top Hf(x^t)(x - x^t)}_{g_t(x)} + o\Big(\left\| x - x^t \right\|^2 \Big)$$

We know that a local optimum x^* is reached when

$$\nabla f(x^*) = 0 \qquad \text{and} \qquad H(f(x^*)) \succ 0$$

In order to solve such a problem, we are going to use the NEWTON's method. If f is a convex function, then $\nabla g_t(x) = \nabla f(x^t) + Hf(x^t)(x - x^t)$ and we only need to find x^* so that $\nabla g_t(x) = 0$, i.e. we set $x^{t+1} = x^\top - (Hf(x^t))^{-1} \nabla f(x^t)$. If the Hessian matrix is not invertible, we can regularize the problem and minimize $g_t(x) + \lambda \|x - x^\top\|^2$ instead.

In general the previous update, called the *pure* NEWTON *step* does not lead to a convergent algorithm even if the function is convex!

In general it is necessary to use the so-called *damped* NEWTON *method*, to obtain a convergent algorithm which consists in doing the following iterations:

$$x^{t+1} = x^t - \varepsilon^t (Hf(x^t))^{-1} \nabla f(x^t)$$

where ε^t is set with the ARMIJO line search.

This method may be computationally costly in high dimension because of the inverse of the hessian matrix that needs to be computed at each iteration. For some functions, however, the pure NEWTON's method does converge. This is the case for logistic regression.

In the context of non-convex optimization, the situation is more complicated because the Hessian can have negative eigenvalues. In that case, so-called trust region methods are typically used.

V. Review on graphs

DEFINITION V. .1. [GRAPH]
A *graph* is a pair $G = (V, E)$ comprising a set V of vertices or nodes together with a set $E \subset V \times V$ of edges or arcs, which are 2-element subsets of V.

REMARK V. .1. In this course we only consider graphs without self-loop.

V. A. Undirected graphs

DEFINITION V. .2. [UNDIRECTED GRAPH]
$G = (V, E)$ is an *undirected graph* if for all $u \neq v \in V \times V$ we have:

$$(u, v) \in E \iff (v, u) \in E$$

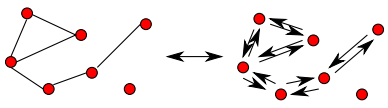


Figure 2: Two different ways to represent an undirected graph

DEFINITION V. .3. [NEIGHBOUR]
We define $\mathcal{N}(u)$, the set of the *neighbours* of u, as

$$\mathcal{N}(u) = \{v \in V \mid (v, u) \in E\}$$

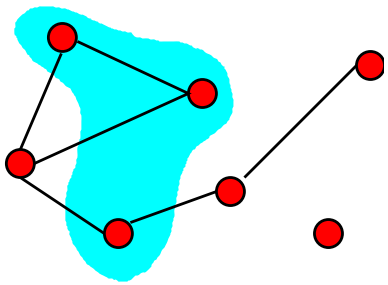


Figure 3: A vertex and its neighbours

DEFINITION V. .4. [CLIQUE]
A totally connected subset of vertices or a singleton is called a *clique*.

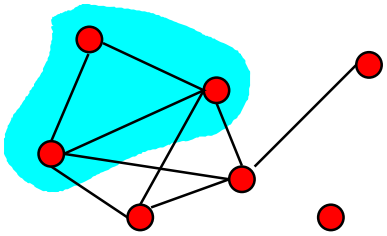


Figure 4: A clique

DEFINITION V. .5. [MAXIMAL CLIQUE]
A *maximal clique* C is a clique which is maximal for the inclusion order, i.e. C is a clique and for all $v \notin C, C \subset \{v\}$ is not a clique.

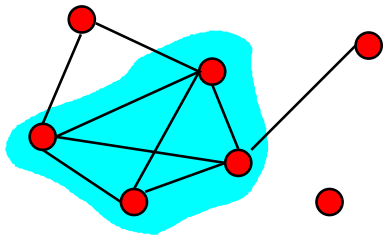


Figure 5: A maximal clique

DEFINITION V. .6. [PATH]
A *path* is a sequence of connected vertices that are globally distinct.

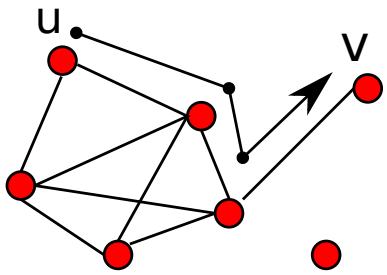


Figure 6: A path from u to v

DEFINITION V. .7. [CYCLE]
A *cycle* is a sequence of vertices $(v_0, \dots, v_k)$ such that:

- $v_0 = v_k$,
- $\forall j \in [\![0, k-1]\!], (v_j, v_{j+1}) \in E$,
- $\forall i, j \in [\![0, k]\!], v_i = v_j \implies \{i, j\} = \{0, k\}$.

DEFINITION V. .8. Let A, B, C be distinct subsets of V. C *separates* A and$B$ if all paths from A to B go through C.

DEFINITION V. .9. [CONNECTED COMPONENT]
A *connected component* is a subgraph induced by the equivalence class of the relation $u\mathcal{R}v$ if and only if exists a path from u to v.

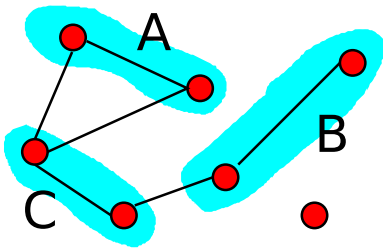


Figure 7: C separates A and B

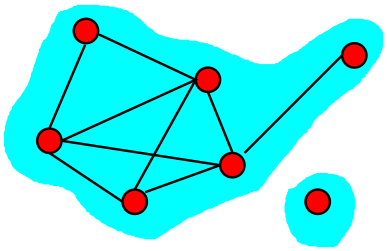


Figure 8: A graph with 2 connected components

In this course we will consider there is only one connected component. Otherwise we can deal with them independently.

V. B. Oriented graphs

DEFINITION V. .10. [PARENT, CHILDREN, ANCESTOR, DESCENDANT]

- v is a *parent* of u if $(v, u) \in E$,
- v is a *children* of u if $(u, v) \in E$,
- v is an *ancestor* of u if there exists a path from u to v,
- v is a *descendant* of u if there exists a path from u to v.

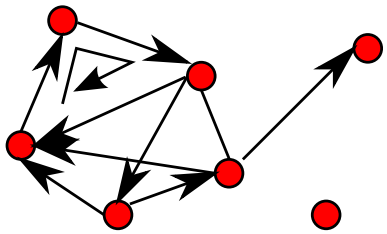


Figure 9: An oriented graph with a cycle

DEFINITION V. .11. [DIRECTED ACYCLIC GRAPH]
A *directed acyclic graph* (DAG) is a directed graph without any cycle.

DEFINITION V. .12. [TOPOLOGICAL ORDER]
Let $G = (V, E)$ a graph with $n = \mathrm{card}(V) < +\infty$. I is a *topological order* if

- I is a bijection from $[\![1, n]\!]$ to V,
- If u is a parent of v, then $I(u) < I(v)$.

PROPOSITION V. .1. *$G = (V, E)$ has a topological order if and only if G is a DAG.*

PROOF The direct implication is easy. Reciprocally, use a depth-first search. □

VI. Review on MARKOV chains

In this annex we assume that we work with random variables taking values in a set $\mathcal{X}$ with $|\mathcal{X}| = K < +\infty$. However K is typically very large since it corresponds to all the configurations that the set of variables of a graphical model can take.

Consider $(X_t)_{t\in\mathbb{N}}$ a sequence of random variables.

DEFINITION VI. .1. [TIME HOMOGENOUS MARKOV CHAIN]
$(X_t)_{t\in\mathbb{N}}$ is a time homogenous MARKOV chain if

$$\forall t \geq 0, \quad \forall (x, y) \in \mathcal{X}, p(X_{t+1} = y \,|\, X_t = x, X_{t-1}, \ldots, X_0) = p(X_{t+1} = y \,|\, X_t = x) = S(x, y)$$

S is called *transition matrix* of the MARKOV chain.

PROPOSITION VI. .1. *If $K < +\infty$, then S is a stochastic matrix:*

$$\forall x, y \in \mathcal{X}, \quad S(x, y) \geq 0 \quad \textit{and} \quad S\mathbf{1} = \mathbf{1}$$

DEFINITION VI. .2. [STATIONARY DISTRIBUTION]
The distribution π on $\mathcal{X}$ is stationary if

$$S^\top \pi = \pi \qquad \text{or equivalently} \qquad \forall y \in \mathcal{X}, \quad \pi(y) = \sum_{x\in\mathcal{X}} \pi(x) S(x, y)$$

If $p(X_{T_0}) = \pi$ with π a stationary distribution of S, then we have $p(X_t) = \pi$ for all $t \geq T_0$.

THEOREM VI. .2. [PERRON-FROBENIUS]
Every stochastic matrix S has at least *one stationary distribution.*

PROPOSITION VI. .3. *One has:*

$$\forall m \in \mathbb{N}, \forall x, y, \quad S^m(x, y) = p(X_{t+m} = y \,|\, X_t = x)$$

DEFINITION VI. .3. [IRREDUCIBLE MARKOV CHAIN]
A MARKOV chain is *irreducible* if

$$\forall x, y \in \mathcal{X}, \quad \exists m \in \mathbb{N}^* \,|\, S^m(x, y) > 0$$

DEFINITION VI. .4. [PERIOD OF A STATE]
The greatest common divider of the elements in the set $\{m > 0 \,|\, S^m(x, x) > 0\}$ is called the period of a state. When the period is equal to 1 the state is said to be aperiodic.

Definition VI. .5. [Aperiodic Markov chain]
If all the states of a Markov chain are aperiodic the chain is said to be aperiodic.

Definition VI. .6. [Regular Markov chain]
A Markov chain is regular if $S(x, y) > 0$ for all $x, y \in \mathcal{X}$.

Remark VI. .1. A regular Markov chain is clearly irreducible aperiodic. The converse is not true.

Proposition VI. .4. *If a* Markov *chain on a finite state space is irreducible and aperiodic, then its transition matrix has a unique stationary distribution π and for any initial distribution q_0 on X_0, if $q_t = p(X_t)$, then $q_t \underset{t\to+\infty}{\longrightarrow} \pi$.*

Remark VI. .2. If the state space is not finite, an additional assumption is needed on the Markov chain: it needs to be recurrent positive. We do not define this notion in this course.

We want to construct an irreducible aperiodic transition S whose stationary distribution is

$$\pi(x) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(x_c)$$

Definition VI. .7. [Detailed balance]
A Markov chain is *reversible* if exists a probability distribution π such that

$$\forall x, y \in \mathcal{X}, \quad \pi(x)S(x, y) = \pi(y)S(y, x)$$

This equation is called the *detailed balance* equation and can be reformulated as

$$p(X_{t+1} = y, X_t = x) = p(X_{t+1} = x, X_t = y)$$

Proposition VI. .5. *If π satisfies detailed balance, then π is a stationary distribution.*

Proof One has $\sum_{x\in\mathcal{X}} S(x, y)p(x) = \sum_x p(y)S(y, x) = p(y) \sum_{x\in\mathcal{X}} S(y, x) = p(y)$. □

VII. Schur complement

Let us consider the block matrix $M = \begin{pmatrix} A & L \\ R & U \end{pmatrix}$. Our goal is to explicit the blocks of its inverse in terms of the initial blocks A, L, U, R[4].

We can block diagonalize M by premultiplying it by D and postmultiplying by G, where:

$$D = \begin{pmatrix} I & 0 \\ -RA^{-1} & I \end{pmatrix} \qquad \text{and} \qquad G = \begin{pmatrix} I & -A^{-1}L \\ 0 & I \end{pmatrix}$$

Indeed:

$$DMG = \begin{pmatrix} A & 0 \\ 0 & U - RA^{-1}L \end{pmatrix}$$

[4] L stands for left, U for upper R for right

and we denote by Δ this block diagonal matrix.

Definition VII. .1. [Schur complement]
The Schur complement of M w.r.t. A is $[M_{/A}] = U - RA^{-1}L$.

By symmetry we obtain the Schur complement of M w.r.t. U as $[M_{/U}] = A - LU^{-1}R$.

From the previous calculations we obtain

Lemma VII. .1. [determinant lemma]

$$\det(M) = \det(A)\det([M_{/A}]) = \det(U)\det([M_{/U}])$$

We also have the following result:

Lemma VII. .2. [positivity lemma]
If M is symmetric then $M \succcurlyeq 0$ if and only if $A \succcurlyeq 0$ and $[M_{/A}] \succcurlyeq 0$.

Proof If M is symmetric then $G = D^\top$. Then

$$\begin{aligned} A \succcurlyeq 0 \text{ and } [M_{/A}] \succcurlyeq 0 &\iff \forall x, x^\top \Delta x \geq 0 \\ &\iff \forall x, (D^\top x)^\top M (D^\top x) \geq 0 \\ &\iff \forall y, y^\top M y \geq 0 \qquad \text{as } D \text{is nonsingular} \iff M \succcurlyeq 0 \end{aligned}$$

□

Proposition VII. .3. [Woodbury-Sherman-Morrison formula]
M is nonsingular if and only if A and $[M_{/A}]$ are. In this case, we have:

$$[M_{\mathbb{A}}]^{-1} = U^{-1} + U^{-1}R[M_{/U}]^{-1}LU^{-1}$$

Proof As $\Delta^{-1} = G^{-1}M^{-1}D^{-1}$, one has $M^{-1} = G\Delta^{-1}D$, from which we obtain:

$$M^{-1} = \begin{pmatrix} A^{-1} + A^{-1}L[M_{/A}]^{-1}RA^{-1} & A^{-1}L[M_{/A}]^{-1} \\ -[M_{/A}]^{-1}RA^{-1} & [M_{/A}]^{-1} \end{pmatrix}$$

Doing the same calculation with the decomposition associated to U, we obtain:

$$M^{-1} = \begin{pmatrix} [M_{/U}]^{-1} & -U^{-1}R[M_{/U}]^{-1} \\ -[M_{/U}]^{-1}LU^{-1} & U^{-1} + U^{-1}R[M_{/U}]^{-1}LU^{-1} \end{pmatrix}$$

□

A useful consequence of the Schur component is to prove rigorously the following inversion lemma:

Lemma VII. .4. [matrix inversion]
Let $X \in \mathbb{R}^{p \times n}$. Then for any λ:

$$(I + \lambda X^\top X)^{-1} = I - \lambda X (I + \lambda X X^\top)^{-1} X^\top$$

In practice, we often want to invert matrix such as $I + \lambda X^\top X$ where X is a design matrix[5] and we usually have $n \gg p$. In that case, the inversion lemma replaces the problem of inverting a square matrix of dimension n (complexity in $\mathcal{O}(n^3)$) by a less costly one of dimension p.

PROOF We can assume $\lambda \neq 0$. We consider $M = \begin{pmatrix} I & X \\ X^\top & -\frac{1}{\lambda} I \end{pmatrix} = \begin{pmatrix} A & L \\ R & U \end{pmatrix}$.

Then $[M_{/U}]^{-1} = (I + \lambda X^\top X)$. Using the WOODBURY-SHERMAN-MORRISON formula, it comes:

$$[M_{/U}]^{-1} = A^{-1} + A^{-1} L [M_{/A}]^{-1} R A^{-1}$$

which gives us the result. □

[5] n is the number of samples and p the number of features