Sylvain Le Corff

Introduction to computational statistics

Contents

Chapter 1

Markov chain Monte Carlo

1.1 Introduction

This chapter aims at designing algorithms to obtain samples from a complex distribution π defined on a measurable space $(\mathsf{X}, \mathcal{X})$. Such algorithms can be applied in many situations, and the target distribution can have several forms depending on the different contexts. In many areas of statistics and machine learning, we are interested in drawing samples from π although π is known only up to a normalizing constant. Such situations arise naturally in a wide variety of contexts, for example, in Bayesian inference, where π represents a posterior distribution over parameters given data, or in energy-based models, where π is a probability distribution defined through an energy function.

Direct sampling from π is often infeasible, either because the distribution has a complex, high-dimensional structure or because computing the normalizing constant is intractable. Markov Chain Monte Carlo (MCMC) methods provide a powerful framework for addressing this challenge. The central idea is to construct a Markov chain whose stationary distribution is precisely π, and then to use the long-run behavior of the chain to generate approximate samples from π.

Example 1.1. Energy-based models (EBM) are very flexible models which describe the target distribution using an unnormalized function, referred to as the energy function. These models are easier to design that models with a tractable likelihood such as autoregressive models, in particular in high-dimensional setting. As the energy function is not normalized, it can be easily parameterized with any nonlinear regression function. Using neural networks such as Multi-layer Perceptrons, or convolutional neural networks, it is straightforward to introduce energy function with specific structures depending nonlinearly on the input.

In a generic setting, the target random variable takes values in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and the target distribution (the target density with respect to the Lebesgue measure) is written:

$$x \mapsto \pi_\theta(x) \propto \exp\left(-\mathrm{E}_\theta(x)\right) = \frac{\exp\left(-\mathrm{E}_\theta(x)\right)}{\int \exp\left(-\mathrm{E}_\theta(u)\right) \mathrm{d}u},$$

where θ is an unknown parameter to estimate and E_θ is the energy function. The normalizing constant is often written Z_θ and referred to as the partition function:

$$Z_\theta = \int \exp\left(-\mathrm{E}_\theta(u)\right) \mathrm{d}u.$$

Since Z_θ is an intractable integral, evaluation and differentiation of $x \mapsto \log \pi_\theta(x)$ is not possible in usual settings. In order to estimate the unknwon parameter θ using i.i.d. data, an appealing approach is to use gradient-based maximization procedures of the likelihood function. This means that we need to compute:

$$x \mapsto \nabla_\theta \log \pi_\theta(x) = -\nabla_\theta \mathrm{E}_\theta(x) - \nabla_\theta \log Z_\theta.$$

The first term can be evaluated easily as $\mathrm{E}_\theta(x)$ is known. For the second term, we can write, under regularity assumptions on the model:

$$\begin{aligned}\nabla_\theta \log Z_\theta = Z_\theta^{-1} \int \nabla_\theta \exp\left(-\mathrm{E}_\theta(u)\right) \mathrm{d}u \\ = \int \left\{-\nabla_\theta \mathrm{E}_\theta(u)\right\} Z_\theta^{-1} \exp\left(-\mathrm{E}_\theta(u)\right) \mathrm{d}u = \int \left\{-\nabla_\theta \mathrm{E}_\theta(u)\right\} \pi_\theta(u) \mathrm{d}u.\end{aligned}$$

Therefore $\nabla_\theta \log Z_\theta = \mathbb{E}_{\pi_\theta}[-\nabla_\theta \mathrm{E}_\theta(X)]$ where $\mathbb{E}_\mu[f(X)]$ denotes the expectation of $f(X)$ when $X \sim \mu$. Therefore, it is possible to train an EBM by providing a Monte Carlo estimate of $\nabla_\theta \log Z_\theta$ which requires to obtain samples from π_θ. However, this is not straightforward as π_θ is known only up to a multiplicative normalizing constant (as in the Bayesian setting).

1.2 Key elements on Markov chains

Let $(\mathsf{X}, \mathcal{X})$ be a measurable space, i.e. $\mathcal{X}$ is a σ-algebra on X, and consider the following notations.

- $\mathsf{M}_+(\mathsf{X})$ is the set of non-negative measures on $(\mathsf{X}, \mathcal{X})$.
- $\mathsf{M}_1(\mathsf{X})$ is the set of probability measures on $(\mathsf{X}, \mathcal{X})$.
- $\mathsf{F}(\mathsf{X})$ is the set of real-valued measurable functions f on X and $\mathsf{F}_+(\mathsf{X})$ the set of non-negative measurable functions on X.
- If $k \leq \ell$, $(u_k, \ldots, u_\ell)$ is denoted by $u_{k:\ell}$ and $(u_{k+\ell})_{\ell \in \mathbb{N}}$ by $u_{k:\infty}$.

Definition 1.2. We say that $P : \mathsf{X} \times \mathcal{X} \to \mathbb{R}^+$ is a Markov kernel if, for all $(x, A) \in \mathsf{X} \times \mathcal{X}$,

- $\mathsf{X} \ni y \mapsto P(y, A)$ is $\mathcal{X}/\mathcal{B}(\mathbb{R}^+)$ measurable,
- $\mathcal{X} \ni B \mapsto P(x, B)$ is a probability measure on $(\mathsf{X}, \mathcal{X})$.

For all $(x, A) \in \mathsf{X} \times \mathcal{X}$, as a function of the first component only, $P(\cdot, A)$ is measurable and as a function of the second component only, $P(x, \cdot)$ is a probability measure. In particular, $P(x, \mathsf{X}) = 1$ for all $x \in \mathsf{X}$. Since $P(x, \cdot)$ is a measure, we also use the infinitesimal notation: $P(x, \mathrm{d}y)$. For example,

$$P(x, A) = \int_{\mathsf{X}} \mathbf{1}_A(y) P(x, \mathrm{d}y) = \int_A P(x, \mathrm{d}y).$$

For all $\mu \in \mathsf{M}_+(\mathsf{X})$, all Markov kernels P, Q on $\mathsf{X} \times \mathcal{X}$, and all measurable non-negative or bounded functions h on X, we use the following conventions and notations.

- μP is the (positive) measure: $\mathcal{X} \ni A \mapsto \mu P(A) = \int \mu(\mathrm{d}x) P(x, A)$,
- PQ is the Markov kernel: $(x, A) \mapsto \int_{\mathsf{X}} P(x, \mathrm{d}y) Q(y, A)$,
- Ph is the measurable function $x \mapsto \int_{\mathsf{X}} P(x, \mathrm{d}y) h(y)$.

It is easy to check that if μ is a probability measure, then μP is also a probability measure (since $\mu P(\mathsf{X}) = \int_{\mathsf{X}} \mu(\mathrm{d}x) P(x, \mathsf{X}) = \int_{\mathsf{X}} \mu(\mathrm{d}x) = 1$). With this notation, using Fubini's theorem,

$$\begin{aligned}\mu(P(Qh)) = (\mu P)(Qh) = (\mu(PQ))h \\ = \mu((PQ)h) = \int_{\mathsf{X}^3} \mu(\mathrm{d}x) P(x, \mathrm{d}y) Q(y, \mathrm{d}z) h(z).\end{aligned}$$

For a given Markov kernel P on $\mathsf{X} \times \mathcal{X}$, define $P^0 = \mathrm{Id}$ where Id is the identity kernel: $(x, A) \mapsto \mathbf{1}_A(x)$, and set for $k \geq 0$, $P^{k+1} = P^k P$.

Definition 1.3. Let $\{X_k : k \in \mathbb{N}\}$ be a sequence of random variables on the same probability space $(\Omega, \mathcal{G}, \mathbb{P})$ and taking values on X, we say that $\{X_k : k \in \mathbb{N}\}$ is a Markov chain with Markov kernel P and initial distribution $\nu \in \mathsf{M}_1(\mathsf{X})$ if and only if the two following statements hold.

i) For all $(k, A) \in \mathbb{N} \times \mathcal{X}$, $\mathbb{P}(X_{k+1} \in A | X_{0:k}) = P(X_k, A)$, $\mathbb{P}$-a.s.
ii) $\mathbb{P}(X_0 \in A) = \nu(A)$.

Note that, in this definition, we consider $\mathbb{P}(X_{k+1} \in A | X_{0:k})$, that is, the conditional probability is with respect to the sigma-field $\sigma(X_{0:k})$. We can replace $\sigma(X_{0:k})$ by $\mathcal{F}_k$ as soon as we know that $(X_k)_{k\geq 0}$ is $(\mathcal{F}_k)_{k\geq 0}$-adapted.

Definition 1.4. We say that $\pi \in \mathsf{M}_1(\mathsf{X})$ is an invariant probability measure for the Markov kernel P on $\mathsf{X} \times \mathcal{X}$ if $\pi P = \pi$.

If (X_k) is a Markov chain with Markov kernel P and assuming that $X_0 \sim \pi$, then for all $k \geq 1$, we have $X_k \sim \pi$ since $\pi P^{k+1} = \pi P^k$ and therefore, for all $k \in \mathbb{N}$, $\pi P^k = \pi$. It can be readily checked that if π is an *invariant probability measure* for P, then the sequence of random variables $\{X_k : k \in \mathbb{N}\}$ is a *strongly stationary sequence* in the sense that for all $n, p \in \mathbb{N}^*$, and all n-tuple $k_{1:n}$, the random vector $(X_{k_1}, \ldots, X_{k_n})$ has the same distribution as $(X_{k_1+p}, \ldots, X_{k_n+p})$.

Definition 1.5. Let $\pi \in \mathsf{M}_1(\mathsf{X})$ and P be a Markov kernel on $\mathsf{X} \times \mathcal{X}$. We say that P is π-reversible if and only if for all measurable bounded or non-negative functions h on $\left(\mathsf{X}^2, \mathcal{X}^{\otimes 2}\right)$,

$$\iint_{\mathsf{X}^2} h(x, y)\pi(\mathrm{d}x)P(x, \mathrm{d}y) = \iint_{\mathsf{X}^2} h(x, y)\pi(\mathrm{d}y)P(y, \mathrm{d}x). \tag{1.1}$$

A Markov kernel P is π-reversible if and only if the probability measure $\pi(\mathrm{d}x)P(x, \mathrm{d}y)$ is symmetric with respect to (x, y). We often write, with infinitesimal notation,

$$\pi(\mathrm{d}x)P(x, \mathrm{d}y) = \pi(\mathrm{d}y)P(y, \mathrm{d}x). \tag{1.2}$$

Proposition 1.6. *Let P be a Markov kernel on $\mathsf{X} \times \mathcal{X}$. Let $\pi \in \mathsf{M}_1(\mathsf{X})$ such that P is π-reversible, then the Markov kernel P is π-invariant.*

Proof. For any $A \in \mathcal{X}$,

$$\pi P(A) = \iint_{\mathsf{X}^2} \mathbf{1}_A(y)\pi(\mathrm{d}x)P(x, \mathrm{d}y) = \iint_{\mathsf{X}^2} \mathbf{1}_A(y)\pi(\mathrm{d}y)P(y, \mathrm{d}x) = \int_A \pi(\mathrm{d}y)P(y, \mathsf{X}) = \pi(A),$$

which completes the proof. □

Therefore, if we want to check easily that a kernel P is π-invariant, it is sufficient to check that it is π-reversible.

1.3 Metropolis-Hastings algorithm

In this section, we are given a probability measure $\pi \in \mathsf{M}_1(\mathsf{X})$ and the idea now is to construct a Markov chain $\{X_k : k \in \mathbb{N}\}$ admitting π as invariant probability measure, in which case we say that π is a target distribution. In other words, we try to find a Markov kernel P on $\mathsf{X} \times \mathcal{X}$ such that P is π-invariant.

For simplicity we now assume that π has a density with respect to some dominating σ-finite measure λ and by abuse of notation, we also denote by π this density, that is we write $\pi(\mathrm{d}x) = \pi(x)\lambda(\mathrm{d}x)$ and we assume that this density π is positive. Moreover, let Q be a Markov kernel on $\mathsf{X} \times \mathcal{X}$ such that $Q(x, \mathrm{d}y) = q(x, y)\lambda(\mathrm{d}y)$, that is, for any $x \in \mathsf{X}$, $Q(x, \cdot)$ is also dominated by λ and denoting by $q(x, \cdot)$ this density, we assume for simplicity that $q(x, y)$ is positive for all $x, y \in \mathsf{X}$.

For a given function $\alpha : \mathsf{X}^2 \to [0, 1]$, Algorithm 1 describes the Metropolis algorithm.

The Markov kernel Q allows to propose a candidate for the next value of the Markov chain $(X_k)_{k\in\mathbb{N}}$ and this candidate is accepted or rejected according to a probability that depends on the

Input : Initial distribution μ, n.
Output: $X_0, \dots, X_n$.

At $t = 0$, draw $X_0 \sim \mu$.
for $t \leftarrow 0$ **to** $n-1$ **do**
- Draw independently $Y_{t+1} \sim Q(X_t, \cdot)$ and $U_{t+1} \sim \mathrm{U}(0,1)$.
- Set $X_{t+1} = \begin{cases} Y_{t+1} & \text{if } U_{t+1}, \leq \alpha(X_t, Y_{t+1}) \\ X_t & \text{otherwise.} \end{cases}$

end

Algorithm 1: The Metropolis Algorithm

function α. We now choose conveniently α in such a way that $(X_k)_{k\in\mathbb{N}}$ is a Markov chain with invariant probability measure π. First, we write down the Markov kernel associated with $(X_k)_{k\in\mathbb{N}}$. Write $\mathcal{F}_t = \sigma(X_0, U_{1:t}, Y_{1:t})$ and note that $(X_t)_{t\in\mathbb{N}}$ is adapted to the filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$ (which is equivalent to $\sigma(X_{0:t}) \subset \mathcal{F}_t$). Then, setting $\bar{\alpha}(x) = 1 - \int_{\mathsf{X}} Q(x, \mathrm{d}y)\alpha(x, y)$, we have for any bounded or non-negative measurable function h on X and any $t \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{E}\left[h(X_{t+1})|\mathcal{F}_t\right] &= \mathbb{E}\left[\mathbf{1}_{\{U_{t+1}<\alpha(X_t,Y_{t+1})\}}h(Y_{t+1})|\mathcal{F}_t\right] + \mathbb{E}\left[\mathbf{1}_{\{U_{t+1}\geq\alpha(X_t,Y_{t+1})\}}|\mathcal{F}_t\right]h(X_t) \\
&= \int_{\mathsf{X}} Q(X_t, \mathrm{d}y)\alpha(X_t, y)h(y) + \bar{\alpha}(X_t)h(X_t) \\
&= \int_{\mathsf{X}} \left[Q(X_t, \mathrm{d}y)\alpha(X_t, y) + \bar{\alpha}(X_t)\delta_{X_t}(\mathrm{d}y)\right] h(y) = P^{MH}_{\langle\pi,Q\rangle}h(X_t).
\end{aligned}$$

Therefore, $\{X_t : t \in \mathbb{N}\}$ is a Markov chain with Markov kernel

$$P^{MH}_{\langle\pi,Q\rangle}(x, \mathrm{d}y) = Q(x, \mathrm{d}y)\alpha(x, y) + \bar{\alpha}(x)\delta_x(\mathrm{d}y). \tag{1.3}$$

Lemma 1.7. *The Markov kernel $P^{MH}_{\langle\pi,Q\rangle}$ is π-reversible if and only if*

$$\pi(\mathrm{d}x)Q(x, \mathrm{d}y)\alpha(x, y) = \pi(\mathrm{d}y)Q(y, \mathrm{d}x)\alpha(y, x). \tag{1.4}$$

Equation (1.4) is often called the detailed balance condition.

Proof. First, note that

$$\pi(\mathrm{d}x)\bar{\alpha}(x)\delta_x(\mathrm{d}y) = \pi(\mathrm{d}y)\bar{\alpha}(y)\delta_y(\mathrm{d}x). \tag{1.5}$$

Indeed, for any measurable function h on X^2, we have

$$\begin{aligned}
\iint_{\mathsf{X}^2} h(x, y)\pi(\mathrm{d}x)\bar{\alpha}(x)\delta_x(\mathrm{d}y) &= \int_{\mathsf{X}} h(x, x)\pi(\mathrm{d}x)\bar{\alpha}(x) \\
&= \int_{\mathsf{X}} h(y, y)\pi(\mathrm{d}y)\bar{\alpha}(y) = \iint_{\mathsf{X}^2} h(x, y)\pi(\mathrm{d}y)\bar{\alpha}(y)\delta_y(\mathrm{d}x).
\end{aligned}$$

Combining (1.3) with (1.5), we obtain that $P^{MH}_{\langle\pi,Q\rangle}$ is π-reversible if and only if the detailed balance condition (1.4) is satisfied. This completes the proof. □

We now provide an explicit expression of the acceptance probability α. The proof of Lemma 1.8 is straightforward.

Lemma 1.8. *Define*

$$\alpha^{MH}(x, y) = \min\left(\frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}, 1\right)$$

and

$$\alpha^{b}(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y) + \pi(y)q(y, x)}.$$

Then, α^{MH} and α^b satisfy the detailed balance condition (1.4).

Example 1.9 (The random walk MH sampler). If $\mathsf{X} = \mathbb{R}^p$ and if the proposal kernel is $Q(x, \mathrm{d}y) = q(y - x)\lambda(\mathrm{d}y)$ where q is a symmetric density with respect to λ on X, (by symmetric, we mean that $q(u) = q(-u)$ for all $u \in \mathsf{X}$) then at each time step in the MH algorithm, we draw a candidate $Y_{k+1} \sim q\,(y\ - X_k)\,\lambda(\mathrm{d}y)$. In such a case, the acceptance probability is $\alpha(x, y) = \min\left(\pi(y)/\pi(x), 1\right)$ and the associated algorithm is called the *(symmetric) Random Walk Metropolis-Hasting.* Another way of writing the proposal update is $Y_{k+1} = X_k + \eta_k$ where $\eta_k \sim q$.

1.4 Variants of MH algorithms

1.4.1 Metropolis–Adjusted Langevin Algorithm

The Metropolis–Adjusted Langevin Algorithm (MALA) combines a Langevin-type proposal with a Metropolis–Hastings correction. The algorithm was in particular analyzed in [Roberts and Tweedie, 1996]. For all $t \geq 0$, given the current state X_t, a proposal Y_{t+1} is generated according to the Euler discretization of the overdamped Langevin dynamics,

$$Y_{t+1} = X_t + \frac{h^2}{2}\nabla \log \pi(X_t) + hZ_t,$$

where $Z_t \sim \mathcal{N}(0, I_d)$ and $h > 0$ denotes the stepsize parameter. This defines a proposal kernel Q given by

$$Q(x, \cdot) = \mathcal{N}\left(x + \frac{h^2}{2}\nabla \log \pi(x),\, h^2 I_d\right).$$

MALA can be interpreted as a Metropolis-corrected Euler discretization of the overdamped Langevin diffusion

$$\mathrm{d}X_t = \frac{1}{2}\nabla \log \pi(X_t)\, dt + \mathrm{d}B_t,$$

which has π as its invariant distribution. The Metropolis–Hastings step corrects for the discretization bias introduced by the Euler scheme, ensuring exact invariance of π. The efficiency of MALA strongly depends on the choice of the stepsize h. In high dimension, for a broad class of target distributions (including product measures), optimal scaling is achieved when

$$h \asymp d^{-1/6},$$

which leads to a non-degenerate limiting acceptance rate, as established in [Roberts and Rosenthal, 1998]. Using larger stepsizes results in low acceptance probabilities, while overly small stepsizes lead to slow exploration of the state space. Compared to Random Walk Metropolis algorithms, MALA leverages gradient information through $\nabla \log \pi$, yielding improved mixing and reduced random-walk behavior, particularly in moderately high dimensions. This improvement comes at the expense of an increased computational cost per iteration due to gradient evaluations, resulting in a trade-off between statistical efficiency and numerical cost.

1.4.2 Generalisation of MH Algorithms

Let $\pi \in \mathsf{M}_1(\mathsf{X})$ and let Q be a Markov kernel on $\mathsf{X} \times \mathcal{X}$. In this chapter, we have presented the Metropolis-Hastings algorithm when π and $Q(x, \cdot)$ have both densities with respect to a common dominating measure λ. In this section, we do not make such an assumption so that the expression of α^{MH} given in Lemma 1.8 is not available anymore and should be adapted. Instead, we will need

the following assumption. Define

$$\mu_0(\mathrm{d}x\mathrm{d}y) = \pi(\mathrm{d}x)Q(x,\mathrm{d}y) \quad \text{and} \quad \mu_1(\mathrm{d}x\mathrm{d}y) = \pi(\mathrm{d}y)Q(y,\mathrm{d}x).$$

(B1) There exists a function $(x,y) \mapsto r(x,y)$ such that $r(x,y) > 0$, μ_0-a.s. and for all $h \in \mathsf{F}_+(\mathsf{X}^2)$,

$$\int h(x,y)\mu_1(\mathrm{d}x\mathrm{d}y) = \int h(x,y)r(x,y)\mu_0(\mathrm{d}x\mathrm{d}y) \tag{1.6}$$

This equation shows that the measure μ_1 is dominated by μ_0 with a μ_0-a.s. positive density:

$$r(x,y) = \frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_0}(x,y).$$

Then, by symmetry, we can easily show that $1/r(x,y) = r(y,x)$, μ_1-a.s. And finally the two measures, μ_0 and μ_1 are equivalent (one is dominated by the other and conversely). In this case, the generalised version of the Metropolis-Hastings kernel, where α^{MH} given in Lemma 1.8 is replaced by $\alpha(x,y) = r(x,y) \wedge 1$ is π-reversible.

Lemma 1.10. *Assume (B1). Then, setting $\alpha(x,y) = r(x,y) \wedge 1$, the MH kernel:*

$$P^{MH}_{\langle \pi, Q\rangle}(x,\mathrm{d}y) = Q(x,\mathrm{d}y)\alpha(x,y) + \bar{\alpha}(x)\delta_x(\mathrm{d}y), \quad \textit{where} \quad \bar{\alpha}(x) = 1 - \int_{\mathsf{X}} Q(x,\mathrm{d}y)\alpha(x,y),$$

is π-reversible.

Proof. Similarly to Lemma 1.7, we only need to check the detailed balance condition. Let $h \in \mathsf{F}_+(\mathsf{X})$, then,

$$\begin{aligned}
&\int_{\mathsf{X}^2} \pi(\mathrm{d}x)Q(x,\mathrm{d}y)\alpha(x,y)h(x,y) = \int_{\mathsf{X}^2} \mu_0(\mathrm{d}x\mathrm{d}y)(r(x,y)\wedge 1)h(x,y) \\
&= \int_{\mathsf{X}^2} \mu_0(\mathrm{d}x\mathrm{d}y)r(x,y)\left(1 \wedge \frac{1}{r(x,y)}\right)h(x,y) = \int_{\mathsf{X}^2} \mu_1(\mathrm{d}x\mathrm{d}y)\left(1 \wedge \underbrace{1/r(x,y)}_{r(y,x)}\right)h(x,y) \\
&= \int_{\mathsf{X}^2} \pi(\mathrm{d}y)Q(y,\mathrm{d}x)\alpha(y,x)h(x,y).
\end{aligned}$$

Thus, the detailed balance condition is verified and the proof is completed. □

1.4.3 Pseudo-marginal Monte Carlo methods

Assume that π and Q are dominated by a common dominating measure λ and write by abuse of notation, $\pi(\mathrm{d}x) = \pi(x)\lambda(\mathrm{d}x)$ and $Q(x,\mathrm{d}y) = q(x,y)\lambda(\mathrm{d}y)$. When considering a Metropolis-Hastings algorithm, we need an explicit expression of $\pi(x)$ for any $x \in \mathsf{X}$, up to a multiplicative constant. It may happen that we are not able to calculate $\pi(x)$ explicitly (even up to a multiplicative constant). Instead, assume that we are able to have an unbiased estimator of $\pi(x)$. To obtain such an unbiased estimator, we draw $W \sim R(x,\mathrm{d}w)$ where R is a Markov kernel from X to $\mathbb{R}^+_\star$, that is, a Markov kernel on $\mathsf{X} \times \mathcal{B}(\mathbb{R}^+_\star)$ such that $\int_{\mathbb{R}^+_\star} wR(x,\mathrm{d}w) = \pi(x)$ (the *unbiasedness* condition). The pseudo-marginal algorithm is described in Algorithm 2 below. We now justify the Pseudo-marginal Monte Carlo algorithm by showing that it is actually a generalized MH algorithm (as described in Lemma 1.10) by considering extended Markov chain, $(\bar{X}_k)_{k\in\mathbb{N}} = (X_k, W_k)_{k\in\mathbb{N}}$ on an extended space and with an extended target. Define the extended

Input : Initial distribution μ, n.
Output: $X_0, \dots, X_n$

At $t = 0$, draw $X_0 \sim \mu$ and $W_0 \sim R(X_0, \cdot)$.
for $t \leftarrow 0$ **to** $n-1$ **do**
- Draw $\tilde{X}_{t+1} \sim Q(X_t, \cdot)$ and then $\tilde{W}_{t+1} \sim R(\tilde{X}_{t+1}, \cdot)$.
- Set $(X_{t+1}, W_{t+1}) = \begin{cases} (\tilde{X}_{t+1}, \tilde{W}_{t+1}) & \text{with probability } \frac{\tilde{W}_{t+1} q(\tilde{X}_{t+1}, X_t)}{W_t q(X_t, \tilde{X}_{t+1})} \wedge 1, \\ (X_t, W_t) & \text{with probability } 1 - \frac{\tilde{W}_{t+1} q(\tilde{X}_{t+1}, X_t)}{W_t q(X_t, \tilde{X}_{t+1})} \wedge 1. \end{cases}

end

Algorithm 2: The Pseudo-Marginal MH Algorithm

target distribution $\Pi(\mathrm{d}\bar{x}) = \Pi(\mathrm{d}x\mathrm{d}w) = wR(x, \mathrm{d}w)\lambda(\mathrm{d}x)$ (where we set $\bar{x} = (x, w)$). Note that Π is indeed a probability measure on $\bar{\mathsf{X}} = \mathsf{X} \times \mathbb{R}_\star^+$, since

$$\iint_{\mathsf{X}\times\mathbb{R}_\star^+} \Pi(\mathrm{d}x\mathrm{d}w) = \int_{\mathsf{X}} \left(\int_{\mathbb{R}_\star^+} wR(x, \mathrm{d}w) \right) \lambda(\mathrm{d}x) = \int_{\mathsf{X}} \pi(x)\lambda(\mathrm{d}x) = 1.$$

Moreover, in Algorithm 2, the candidate $(\tilde{X}_{t+1}, W_{t+1})$ is proposed according to $\bar{Q}$ where the proposal kernel $\bar{Q}$ is defined by $\bar{Q}(\bar{x}, \mathrm{d}\bar{x}') = Q(x, \mathrm{d}x')R(x', \mathrm{d}w')$. In order to check (B1), we first set

$$\begin{aligned} \mu_0(\mathrm{d}\bar{x}\mathrm{d}\bar{x}') &= \mu_0(\mathrm{d}x\mathrm{d}w\mathrm{d}x'\mathrm{d}w') = wR(x, \mathrm{d}w)\lambda(\mathrm{d}x)Q(x, \mathrm{d}x')R(x', \mathrm{d}w') \\ \mu_1(\mathrm{d}\bar{x}\mathrm{d}\bar{x}') &= \mu_1(\mathrm{d}x\mathrm{d}w\mathrm{d}x'\mathrm{d}w') = w'R(x', \mathrm{d}w')\lambda(\mathrm{d}x')Q(x', \mathrm{d}x)R(x, \mathrm{d}w). \end{aligned}$$

Then, writing $Q(x, \mathrm{d}y) = q(x, y)\lambda(\mathrm{d}y)$, we obtain for all $h \in \mathsf{F}_+(\bar{\mathsf{X}}^2)$,

$$\begin{aligned} \int_{\bar{\mathsf{X}}^2} h(\bar{x}, \bar{x}')\mu_1(\mathrm{d}\bar{x}\mathrm{d}\bar{x}') &= \int_{\bar{\mathsf{X}}^2} h(\bar{x}, \bar{x}')w'q(x', x)[R(x, \mathrm{d}w)R(x', \mathrm{d}w')\lambda(\mathrm{d}x)\lambda(\mathrm{d}x')] \\ &= \int_{\bar{\mathsf{X}}^2} h(\bar{x}, \bar{x}')r(\bar{x}, \bar{x}')\mu_0(\mathrm{d}\bar{x}\mathrm{d}\bar{x}'), \end{aligned}$$

where

$$r(\bar{x}, \bar{x}') = \frac{w'q(x', x)}{wq(x, x')}.$$

Since r is positive, we can apply Lemma 1.10 with $\alpha(\bar{x}, \bar{x}') = r(\bar{x}, \bar{x}') \wedge 1$ and we finally get that $P^{MH}_{\langle \Pi, \bar{Q} \rangle}(\bar{x}, \mathrm{d}\bar{x}')$ is Π-reversible. Since Algorithm 2 corresponds to applying the Markov kernel $P^{MH}_{\langle \Pi, \bar{Q} \rangle}$, this completes the proof. Note that the extended target distribution Π has the marginal π with respect to the first component:

$$\Pi(A \times \mathbb{R}_\star^+) = \int_A \int_{\mathbb{R}_\star^+} wR(x, \mathrm{d}w)\lambda(\mathrm{d}x) = \int_A \pi(\mathrm{d}x) = \pi(A).$$

To sum up, $(\bar{X}_k)_{k\in\mathbb{N}} = (X_k, W_k)_{k\in\mathbb{N}}$ produced by Algorithm 2 is a generalized Metropolis-Hastings algorithm where the target distribution Π admits π as the marginal distribution on the first component. Note that $(X_k)_{k\in\mathbb{N}}$ is not a Markov chain anymore (but $(\bar{X}_k)_{k\in\mathbb{N}}$ is).

Chapter 2

Expectation Maximization algorithm

The Expectation Maximization (EM) algorithm [Dempster et al., 1977] is a general iterative method for maximum likelihood estimation in statistical models that involve latent (hidden) variables or missing data. When the likelihood function is difficult or impossible to optimize directly because part of the information is hidden, the EM algorithm provides a systematic way to alternate between estimating the missing information and optimizing the parameters. This algorithm is widely used in statistics, machine learning, and signal processing, especially for mixture models, clustering, density estimation, and probabilistic inference.

2.1 Introduction

Let $(\mathsf{Z}, \mathcal{Z})$ be a measurable space and λ be a measure on $(\mathsf{Z}, \mathcal{Z})$. Consider also a family $\{f_\theta\}_{\theta\in\Theta}$ of λ-integrable and positive functions. Define

$$L(\theta) = \int f_\theta(z)\lambda(\mathrm{d}z).$$

We aim at solving

$$\widehat{\theta} \in \mathrm{Argmax}_{\theta\in\Theta} L(\theta).$$

When L is positive, the problem is often written:

$$\widehat{\theta} \in \mathrm{Argmax}_{\theta\in\Theta}\, \ell(\theta) = \log L(\theta).$$

2.2 Algorithm

In the following, we write $q_\theta : z \mapsto f_\theta(z)/L(\theta)$. Solving the optimization problem is not possible in general frameworks. The Expectation Maximization (EM) algorithm computes sequentially $\{\theta_k\}_{k\geq 0}$ to estimate $\widehat{\theta}$. For all $\theta, \theta' \in \Theta$, we introduce the following quantity:

$$Q(\theta, \theta') = \int \log f_\theta(z) q_{\theta'}(z)\lambda(\mathrm{d}z) = \mathbb{E}_{\theta'}[\log f_\theta(Z)],$$

where $\mathbb{E}_\theta$ is a notation for the expectation under the density q_θ. Then, we can write for all $\theta, \theta' \in \Theta$

$$Q(\theta, \theta') = \int \log(L(\theta) q_\theta(z)) q_{\theta'}(z) \lambda(\mathrm{d}z) = \ell(\theta) - H(\theta, \theta'),$$

where $H(\theta, \theta') = -\int \log q_\theta(z) q_{\theta'}(z) \lambda(\mathrm{d}z)$.

Lemma 2.1. *For all $\theta, \theta' \in \Theta$,*

$$\ell(\theta) - \ell(\theta') \geq Q(\theta, \theta') - Q(\theta', \theta').$$

Proof. By definition, for all $\theta, \theta' \in \Theta$,

$$\begin{aligned} Q(\theta, \theta') - Q(\theta', \theta') &= \ell(\theta) - \ell(\theta') + H(\theta', \theta') - H(\theta, \theta') \\ &= \ell(\theta) - \ell(\theta') + \int \log\left(\frac{q_\theta(z)}{q_{\theta'}(z)}\right) q_{\theta'}(z) \lambda(\mathrm{d}z). \end{aligned}$$

As log is concave, by Jensen's inequality,

$$\int \log\left(\frac{q_\theta(z)}{q_{\theta'}(z)}\right) q_{\theta'}(z) \lambda(\mathrm{d}z) \leq \log \int \frac{q_\theta(z)}{q_{\theta'}(z)} q_{\theta'}(z) \lambda(\mathrm{d}z) = 0,$$

which concludes the proof. The inequality $\int \log(q_\theta(z)) q_{\theta'}(z) \lambda(\mathrm{d}z) \leq \int \log(q_{\theta'}(z)) q_{\theta'}(z) \lambda(\mathrm{d}z)$ is known as Gibbs' inequality. □

By Lemma 2.1, starting from a parameter estimate θ_k, $k \geq 0$, a direct solution to obtain a parameter θ such that $\ell(\theta) \geq \ell(\theta_k)$ is to choose θ such that $Q(\theta, \theta_k) \geq Q(\theta_k, \theta_k)$. This result motivates the Expectation Maximization (EM) algorithm given in Algorithm 3 and introduced in [Dempster et al., 1977].

Data: Initial parameter estimate θ_0
Result: A sequence of parameter estimate $\{\theta_k\}_{k \geq 0}$
for $k \geq 0$ **do**
  Compute the E-step: $\theta \mapsto Q(\theta, \theta_k)$;
  Compute the M-step: $\theta_{k+1} \in \mathrm{Argmax}_{\theta \in \Theta} Q(\theta, \theta_k)$;
end

Algorithm 3: A generic EM algorithm

The most common setting in which the EM algorithm is used is the case of Example **??**.

2.3 Convergence properties

The first theoretical guarantees for the EM algorithm were provided in [Wu, 1983]. Given $k \geq 0$ and $\theta_k \in \Theta$, the EM update is

$$\theta_{k+1} \in M(\theta_k), \quad \text{where} \quad M(\theta') = \arg\max_{\theta \in \Theta} Q(\theta, \theta').$$

Therefore, M is a map from points of Θ to subsets of Θ and is referred to as a point-to-set map on Θ. This map is said to be closed at $\eta_* \in \Theta$ if for all sequence $\{\eta_k\}_{k \geq 0}$ such that η_k converges to η_* as $k \to \infty$, if for all $k \geq 0$, $y_k \in M(\eta_k)$ and $\{y_k\}_{k \geq 0}$ converges to $y_* \in \Theta$ as $k \to \infty$, then $y_* \in M(\eta_*)$. Convergence of the EM algorithm relies on a more general convergence theorem for point-to-set map.

Theorem 2.2. *Assume that $\Theta \subset \mathbb{R}^m$ and let M be a point-to-set map defined on Θ. Consider a sequence $\{\theta_k\}_{k \geq 0}$ such that for all $k \geq 1$,*

$$\theta_{k+1} \in M(\theta_k).$$

and let $\mathcal{S} \subset \Theta$ denote the set of solution points. Assume that the sequence $\{\theta_k\}_{k\geq 0}$ is contained in a compact subset of Θ and that M is closed at all points $\theta \in \Theta \setminus \mathcal{S}$. Assume also that there exists $\alpha : \Theta \to \mathbb{R}$ a continuous function, such that for all $\theta \in \Theta$, if $\theta \in \Theta \setminus \mathcal{S}$ then for all $y \in M(\theta)$, $\alpha(y) > \alpha(\theta)$ and if $\theta \in \mathcal{S}$, then $\alpha(y) \geq \alpha(\theta)$ for all $y \in M(\theta)$. Then, every accumulation point of the sequence $\{\theta_k\}_{k\geq 0}$ belongs to the solution set $\mathcal{S}$.

Proof. The proof can be found in [Zangwill, 1969]. □

H1 The level set

$$\{\theta \in \Theta : \ell(\theta) \geq \ell(\theta_0)\}.$$

is compact.

By Proposition 2.1, the sequence $\{\ell(\theta_k)\}_{k\geq 0}$ is non-decreasing. Therefore, by H1 this sequence converges to some limit point ℓ_*. However, there is no guarantee that ℓ_* is the global maximum of ℓ.

H2 The functions $(\theta, \theta') \mapsto Q(\theta, \theta')$ and $\theta \mapsto \ell(\theta)$ are continuously differentiable, and differentiation under the integral sign is valid.

Then, convergence of the EM algorithm is a specific application of Theorem 2.2 as stated in Theorem 2.3

Theorem 2.3. *Assume that H1 holds. Let $\{\theta_p\}_{p\geq 0}$ be a sequence generated by the EM algorithm:*

$$\theta_{p+1} \in M(\theta_p),$$

and suppose that:

(i) M is a closed point-to-set mapping over the complement of S;
(ii) $L(\theta_{p+1}) > L(\theta_p)$ for all $\theta_p \notin S$.

Then, all limit points of $\{\theta_p\}_{p\geq 0}$ are stationary points (local maxima) of L, and

$$\lim_{p\to\infty} L(\theta_p) = L(\theta^*) \quad \text{for some } \theta^* \in S.$$

Theorem 2.4. *Assume that H1-2 hold. Then, all limit points of any EM sequence $\{\theta_p\}_{p\geq 0}$ are stationary points of L, and*

$$\lim_{p\to\infty} L(\theta_p) = L(\theta^*) \quad \text{for some stationary point } \theta^*.$$

Proof. The proof is a consequence of Theorem 2.3. By assumption H2, we can define the set of stationary points of the log-likelihood:

$$\mathcal{S} = \{\theta \in \Theta\,,\, \nabla\ell(\theta_*)\}\,.$$

We can also choose α as the log-likelihood function ℓ. Assumptions H1 ensures that $\{\theta_k\}_{k\geq 0}$ is contained in a compact subset of Θ. In addition, by H2, M is closed at all points $\theta \in \theta \in \Theta \setminus \mathcal{S}$, and for all limit point θ^* of $\{\theta_k\}_{k\geq 0}$, $\theta^* \in M(\theta^*)$. By Lemma 2.1, and the definition of M, for all $\theta \in \Theta$, if $\theta \in \mathcal{S}$ then for all $y \in M(\theta)$, $\alpha(y) \geq \alpha(\theta)$. It remains to prove that for all $\theta \in \Theta$, if $\theta \in \Theta \setminus \mathcal{S}$ then for all $y \in M(\theta)$, $\alpha(y) > \alpha(\theta)$. By definition of θ_p,

$$\nabla\ell(\theta_p) = \nabla_\theta Q(\theta, \theta_p)\big|_{\theta=\theta_p}.$$

Therefore, if $\theta_p \notin \mathcal{S}$, $\nabla\ell(\theta_p) \neq 0$ and θ_p is not a local maximum of $\theta \mapsto Q(\theta, \theta_p)$ which implies that $Q(\theta_{p+1}, \theta_p) > Q(\theta_p, \theta_p)$ and $L(\theta_{p+1}) > L(\theta_p)$. This concludes the frist part of the proof. Since θ^* maximizes $\theta \mapsto Q(\theta, \theta^*)$,

$$\nabla_\theta Q(\theta, \theta^*)\big|_{\theta=\theta^*} = 0.$$

Under the differentiability assumption,

$$\nabla_\theta Q(\theta, \theta^*)\big|_{\theta=\theta^*} = \nabla \ell(\theta^*),$$

which proves the result. □

2.4 Application to latent data models

A very popular setting is when and $f_\theta : z \mapsto p_\theta(z, X)$ where p_θ is the joint probability density function of two random variables (Z, X). Assuming that the random variable X is observed and that Z is not observed, we consider the likelihood function

$$L(\theta) = \int f_\theta(z, X)\lambda(\mathrm{d}z),$$

which is a random variable depending on X. This is the marginal density of X when the parameter is θ. In this setting, $f_\theta(Z, X)/L(\theta)$ is the probability density of the conditional distribution of Z given X. Solving $\widehat{\theta} \in \mathrm{Argmax}_{\theta\in\Theta} L(\theta)$ amounts to solving the maximum likelihood estimation problem. However, in this setting, as in many other settings, the integral is intractable and the optimization problem cannot be solved directly. We have

$$Q(\theta, \theta') = \int \log p_\theta(z, X) p_{\theta'}(z|X)\lambda(\mathrm{d}z) = \mathbb{E}_{\theta'}[\log p_\theta(Z, X)|X].$$

Therefore, the E-step of the EM algorithm amounts to computing the conditional expectation given X of the complete data (joint) loglikelihood. If $\{(Z_i, X_i)\}_{1\le i\le n}$ are i.i.d we write $X = (X_1, \dots, X_n)$ and $Z = (Z_1, \dots, Z_n)$. In this case, the intermediate quantity of the EM algorithm is

$$Q(\theta, \theta') = \sum_{i=1}^{n} \mathbb{E}_{\theta'}[\log p_\theta(Z_i, X_i)|X_i].$$

2.5 Example: mixture of Gaussian distributions

In this example, we assume that the joint distribution of (Z, X) belongs to a family of distributions parametrized by a vector θ with real components. For $k \in \{1, \dots, M\}$, write $\pi_k = \mathbb{P}_\theta(Z = k)$. Assume that conditionally on the event $\{Z = k\}$, X has a Gaussian distribution with mean $\mu_k \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. The probability density of this conditional distribution is written g_k^θ, where the parameter $\theta = (\{\pi_k\}_{1\le k\le K}, \{\mu_k\}_{1\le k\le K}, \Sigma)$ belongs to the set $\Theta = \mathbb{S}_K \times (\mathbb{R}^d)^K \times \mathbb{R}^{d\times d}$ with $\mathbb{S}_K = \{(\pi_1, \dots, \pi_K) \in [0, 1]^K \ ; \ \sum_{k=1}^K \pi_k = 1\}$. For all $1 \le k \le K$, the explicit computation of $\mathbb{P}_\theta(Z = k|X)$ writes

$$\mathbb{P}_\theta\left(Z = k|X\right) = \frac{\pi_k g_k^\theta(X)}{\sum_{\ell=1}^K \pi_\ell g_\ell^\theta(X)}.$$

Assume that $\{(X_i, Z_i)\}_{1\le i\le n}$ are i.i.d. with this distribution parameterized by $\theta \in \Theta$. Then, the complete-data loglikelihood writes

$$\log p_\theta(Z_{1:n}, X_{1:n}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{1}_{Z_i=k} \left\{\log(\pi_k) - \frac{1}{2}\log\det(2\pi\Sigma) - \frac{1}{2}(X_i - \mu_k)^\top \Sigma^{-1}(X_i - \mu_k)\right\}.$$

Therefore, the intermediate quantity of the EM algorithm is given, for all $\theta, \theta' \in \Theta$, by

$$\begin{aligned} Q(\theta, \theta') &= \mathbb{E}_{\theta'} \left[\log p_\theta(Z_{1:n}, X_{1:n}) | X_{1:n} \right], \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} \left\{ \mathbb{P}_{\theta'}(Z_i = k | X_i) \left(\log(\pi_k) - \frac{1}{2} \log \det(2\pi\Sigma) - \frac{1}{2} (X_i - \mu_k)^\top \Sigma^{-1} (X_i - \mu_k) \right) \right\}. \end{aligned}$$

The algorithm is initialized by choosing $\theta^{(0)}$ randomly. Then, for each iteration $p \geq 0$, the current parameter estimate is written

$$\theta^{(p)} = \left\{ \{\pi_k^{(p)}\}_{1 \leq k \leq K}, \{\mu_k^{(p)}\}_{1 \leq k \leq K}, \Sigma^{(p)} \right\},$$

and the update is decomposed into two steps.

1. Compute $\mathbb{P}_{\theta^{(p)}}(Z_i = k | X_i)$ for all $1 \leq i \leq n$, $1 \leq k \leq K$:

$$\mathbb{P}_{\theta^{(p)}}(Z_i = k | X_i) = \frac{\pi_k^{(p)} g_k^{\theta^{(p)}}(X)}{\sum_{\ell=1}^{K} \pi_\ell^{(p)} g_\ell^{\theta^{(p)}}(X)} = \omega_{i,k}^{(p)}.$$

2. Update the parameter estimate by computing:

$$\theta^{(p+1)} \in \mathrm{Argmax}_{\theta \in \Theta} Q(\theta, \theta^{(p)}).$$

The intermediate quantity is given, for all θ, by:

$$Q(\theta, \theta^{(p)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \left\{ \omega_{i,k}^{(p)} \left(\log(\pi_k) - \frac{1}{2} \log \det(2\pi\Sigma) - \frac{1}{2} (X_i - \mu_k)^\top \Sigma^{-1} (X_i - \mu_k) \right) \right\}.$$

By Lemma 2.6, and using that $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$, for all $1 \leq k \leq K-1$,

$$\frac{\partial Q(\theta, \theta^{(p)})}{\partial \pi_k} = \left(\sum_{i=1}^{n} \omega_{i,k}^{(p)} \right) \frac{1}{\pi_k} - \left(\sum_{i=1}^{n} \omega_{i,K}^{(p)} \right) \frac{1}{\pi_K}.$$

In addition, for all $1 \leq k \leq K$,

$$\frac{\partial Q(\theta, \theta^{(p)})}{\partial \mu_k} = \sum_{i=1}^{n} \omega_{i,k}^{(p)} \left(2\Sigma^{-1} X_i - 2\Sigma^{-1} \mu_k \right),$$

and

$$\frac{\partial Q(\theta, \theta^{(p)})}{\partial \Sigma^{-1}} = \frac{n}{2} \Sigma - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} \omega_{i,k}^{(p)} \left(X_i - \mu_k \right) \left(X_i - \mu_k \right)^\top.$$

The maximum likelihood estimator is defined as the only parameter $\widehat{\theta}^{(p+1)}$ such that all these equations are set to 0. We can note that there exists $c > 0$ such that for all $k \in \{1, \ldots, K\}$, $\pi_k = c \sum_{i=1}^{n} \omega_{i,k}^{(p)}$. Computing the sum for $k = 1$ to $k = K$ yields $c = n$. For $k \in \{1, \ldots, K\}$, it is given by

$$\pi_k^{(p+1)} = \frac{1}{n}\sum_{i=1}^{n} \omega_{i,k}^{(p)},$$
$$\mu_k^{(p+1)} = \frac{1}{\sum_{i=1}^{n} \omega_{i,k}^{(p)}} \sum_{i=1}^{n} \omega_{i,k}^{(p)} X_i,$$
$$\Sigma^{(p+1)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \omega_{i,k}^{(p)} \left(X_i - \mu_k^{(p+1)}\right)\left(X_i - \mu_k^{(p+1)}\right)^\top.$$

Remark 2.5. Note that the computation of $\mathbb{P}_{\theta^{(p)}}(Z_i = k|X_i)$ is explicit. The fact that the conditional expectation (and therefore $Q(\theta, \theta^{(p)})$) can be computed explicitly is a consequence of the fact that Z_i is a discrete random variable. In other cases, $Q(\theta, \theta^{(p)})$ is likely to be unavailable explicitly and is often replaced by Monte Carlo estimators. In the setting of Gaussian mixtures, the computation of $\theta^{(p+1)}$ is also explicit. The M-step is often replaced by simply choosing an estimator $\theta^{(p+1)}$ such that $Q(\theta^{(p+1)}, \theta^{(p)}) > Q(\theta^{(p)}, \theta^{(p)})$ which is tractable in many cases and yields the Generalized EM algorithm.

Lemma 2.6. *Let Σ be a symmetric and invertible matrix in $\mathbb{R}^{d\times d}$.*

(i) The derivative of the real valued function $\Sigma \mapsto \log\det(\Sigma)$ defined on $\mathbb{R}^{d\times d}$ is given by:

$$\partial_\Sigma\{\log\det(\Sigma)\} = \Sigma^{-1},$$

where, for all real valued function f defined on $\mathbb{R}^{d\times d}$, $\partial_\Sigma f(\Sigma)$ denotes the $\mathbb{R}^{d\times d}$ matrix such that for all $1 \leqslant i, j \leqslant d$, $\{\partial_\Sigma f(\Sigma)\}_{i,j}$ is the partial derivative of f with respect to $\Sigma_{i,j}$.

(ii) The derivative of the real valued function $x \mapsto x^\top \Sigma x$ defined on $\mathbb{R}^d$ is given by:

$$\partial_x\{x^\top \Sigma x\} = 2\Sigma x.$$

Proof. (i) Recall that for all $i \in \{1, \dots, d\}$ we have $\det(\Sigma) = \sum_{k=1}^{d} \Sigma_{i,k}\Delta_{i,k}$ where $\Delta_{i,j}$ is the (i,j)-cofactor associated with Σ. For any fixed i, j, the component $\Sigma_{i,j}$ does not appear in anywhere in the decomposition $\sum_{k=1}^{d} \Sigma_{i,k}\Delta_{i,k}$, except for the term $k = j$. This implies

$$\frac{\partial \log\det(\Sigma)}{\partial \Sigma_{i,j}} = \frac{1}{\det\Sigma}\frac{\partial\det(\Sigma)}{\partial\Sigma_{i,j}} = \frac{\Delta_{i,j}}{\det\Sigma}$$

Recalling the identity $\Sigma\,[\Delta_{j,i}]_{1\leq i,j\leq d} = (\det\Sigma)\, I_d$ so that $\Sigma^{-1} = \frac{[\Delta_{j,i}]_{1\leq i,j\leq d}^\top}{\det\Sigma}$, we finally get

$$\left[\frac{\partial \log\det(\Sigma)}{\partial\Sigma_{i,j}}\right]_{1\leq i,j\leq d} = (\Sigma^{-1})^\top = \Sigma^{-1}$$

where the last equality follows from the fact that Σ is symmetric.

(ii) Define $\varphi(x) = x^\top\Sigma x$. Then, by straightforward algebra, $\varphi(x+h) = \varphi(x) + 2h^\top\Sigma x + \varphi(h) = \varphi(x) + 2h^\top\Sigma x + o(\|h\|)$, which concludes the proof.

□

2.6 Monte Carlo EM

Chapter 3

Variational inference and autoencoders

3.1 Evidence Lower Bound

In this chapter, we consider models with latent (unobserved) data. Let (Z, X) be random variables in $\mathbb{R}^d \times \mathbb{R}^m$. We assume that the law of (Z, X) has a density $(z, x) \mapsto p(z, x)$ with respect to a reference measure. In this setting, we write

$$(z, x) \mapsto p(z, x) = p(z)p(x|z)\,,$$

where $z \mapsto p(z)$ is a prior density for Z and $x \mapsto p(x|z)$ is the conditional density (likelihood) of X given Z. We do not have access to the conditional density of Z given X, since this density is given by:

$$z \mapsto p(z|x) = \frac{p(z)p(x|z)}{p(x)} \propto p(z)p(x|z)\,,$$

where $p(x) = \int p(z)p(x|z)\mathrm{d}z$ is an intractable integral. The conditional law of latent variables given observations is of utmost importance in many machine learning approaches. For instance, in the E-step of the EM algorithm, the intermediate quantity requires to compute an expecation under such a distribution. In most situations, this distribution cannot be sampled from easily. Standard solutions to sample approximately from $p(z|x)$ include Markov Chain (or Sequential) Monte Carlo methods. In this chapter, we focus on variational approaches where $p(z|x)$ is replaced by a simpler distribution obtained by solving an optimization problem.

In variational inference, we introduce a variational family i.e. a family of densities to approximate $z \mapsto p(z|x)$. Let $\mathcal{D}$ be such a family, where the densities $q \in \mathcal{D}$ satisfy the two following assumptions.

- For all $q \in \mathcal{D}$, q is easy to evaluate.
- For all $q \in \mathcal{D}$, q is easy to sample from.

Then, for all x and all $q \in \mathcal{D}$, writing KL the Kullback-Leibler divergence between two probability distributions,

$$\begin{aligned}\mathrm{KL}\left(q \| p(\cdot|x)\right) = \int q(z) \log \frac{q(z)}{p(z|x)} \mathrm{d}z &= \mathbb{E}_q[\log q(Z)] - \mathbb{E}_q[\log p(Z|x)]\,, \\ &= \mathbb{E}_q[\log q(Z)] - \mathbb{E}_q[\log p(Z, x)] + \log p(x)\,, \\ &= -\mathcal{L}_x(q) + \log p(x)\,,\end{aligned}$$

where

$$\mathcal{L}_x(q) = \mathbb{E}_q\left[\log \frac{p(Z,x)}{q(Z)}\right].$$

Using Jensen's inequality, we obtain $\mathrm{KL}\left(q \| p(\cdot|x)\right) \geq 0$ so that

$$\mathcal{L}_x(q) \leq \log p(x)\,.$$

This inequality justifies the name Evidence Lower BOund for $\mathcal{L}_x = \mathbb{E}_q[\log(p(Z,x)/q(Z))]$. In variational inference, we then aim to approximate $p(\cdot|x)$ by q_* where:

$$q_* \in \mathrm{argmax}_{q\in\mathcal{D}}\, \mathcal{L}_x(q)\,.$$

3.2 Coordinate ascent variational inference

The most straightforward approach to solve the optimization problem is to consider a mean-field variational family i.e. to choose $\mathcal{D}$ such that:

$$\mathcal{D} = \left\{ z \mapsto q(z) = \prod_{j=1}^{d} q_j(z_j)\ ;\ q_j \text{ is a density} \right\}.$$

Even when considering a simple variational family such as $\mathcal{D}$, it is not possible to maximize the ELBO explicitly. Assume therefore that we want to optimize $\mathrm{ELBO}(q)$ on q_j only for some $1 \leq j \leq d$, the other densities $(q_\ell)_{\ell\neq j}$ being kept fixed. Write $z_{-j} = (z_\ell)_{1\leq\ell\leq d;\ell\neq j}$, and for all $q \in \mathcal{D}$, $q_{-j}(z_{-j}) = \prod_{1\leq\ell\leq d;\ell\neq j} q_\ell(z_\ell)$. Then,

$$\begin{aligned}
\mathcal{L}_x(q) &= \int \left\{\prod_{\ell=1}^{d} q_\ell(z_\ell)\right\} \left\{\log(p(z_{-j},x)p(z_j|z_{-j},x)) - \sum_{\ell=1}^{d}\log q_\ell(z_\ell)\right\} \mathrm{d}z_1 \ldots \mathrm{d}z_d\,,\\
&= \int q_j(z_j) \int \left\{\prod_{\ell=1,\ell\neq j}^{d} q_\ell(z_\ell)\right\} \{\log(p(z_{-j},x)p(z_j|z_{-j},x))\mathrm{d}z_{-j}\}\,\mathrm{d}z_j\\
&\qquad\qquad - \sum_{\ell=1}^{d}\int q_\ell(z_\ell)\log q_\ell(z_\ell)\mathrm{d}z_\ell\\
&= \mathbb{E}_{q_j}\left[\mathbb{E}_{q_{-j}}\left[\log p(Z_j|Z_{-j},x)\right]\right] - \mathbb{E}_{q_j}\left[\log q_j(Z_j)\right] + \mathrm{C}
\end{aligned}$$

where C does not depend on q_j. Consider the density $\tilde{q}_j$ such that

$$\tilde{q}_j(z_j) \propto \exp\left(\mathbb{E}_{q_{-j}}\left[\log p(z_j|Z_{-j},x)\right]\right)\,,$$

i.e. the density given by $\tilde{q}_j(z_j) = \exp(\mathbb{E}_{q_{-j}}[\log p(z_j|Z_{-j},x)])/C_j$ where C_j does not depend on z_j (C_j is the normalizing constant to obtain a density). Therefore,

$$\mathcal{L}_x(q) = -\mathbb{E}_{q_j}\left[\log\frac{q_j(Z_j)}{\tilde{q}_j(Z_j)}\right] + \mathrm{C} + \log C_j = -\mathrm{KL}(q_j\|\tilde{q}_j) + \mathrm{C} + \log C_j\,.$$

Therefore, optimizing $\mathcal{L}_x(q)$ on q_j only, the other densities $(q_\ell)_{\ell\neq j}$ being kept fixed, yields an optimum given by $\tilde{q}_j$. The algorithm Coordinate Ascent Variational Inference (CAVI) proposes therefore to sequentially update q_j, $1 \leq j \leq d$ until a stopping criterion is met. In Algorithm 4, we propose a version of the algorithm where the variational distribution of only one component of Z is updated at each iteration, of course many alternatives can be considered. A standard alternative is to update each variational distribution at each iteration.

Data: Observation x, initial variational distribution $\{q_j^{(0)}\}_{1\leq j\leq d}$, maximum number of iteration N
Result: A variational distribution for each coordinate of Z, $q_j^{(N)}$, $1 \leq j \leq d$.
for $k = 1 \to N$ **do**
  Draw $j \in \{1, \dots, d\}$ uniformly at random;
  Set $q_\ell^{(k)} = q_\ell^{(k-1)}$ for all $1 \leq \ell \leq d$, $\ell \neq j$ and $q_{-j}^{(k)} = \prod_{1\leq \ell\leq d, \ell\neq j} q_\ell^{(k)}$;
  Set

$$q_j^{(k)}(z_j) \propto \exp\left(\mathbb{E}_{q_{-j}^{(k)}}\left[\log p(z_j|Z_{-j}, x)\right]\right)$$

  ;
end

3.3 Application to a mixture of Gaussian distributions

This example can be found in [Blei et al., 2017]. Consider a mixture of K Gaussian distributions with means $\mu = (\mu_k)_{1\leqslant k\leqslant K}$ and variance 1. The variables $\mu = (\mu_k)_{1\leqslant k\leqslant K}$ are i.i.d. Gaussian with mean 0 and variance σ^2. For all $1 \leq i \leq n$, we denote by $c_i \in \{1, \dots, K\}$ the group the i-th observation belongs to. The variables μ and c are not observed. The random variables $c = (c_i)_{1\leq i\leq n}$ are independent of μ and are independent with multinomial distribution with parameters $\{\omega_1, \dots, \omega_K\}$, where $\sum_{k=1}^K \omega_k = 1$.

Conditionally on μ and c, the observations $(X_i)_{1\leqslant i\leqslant n}$ are independent and X_i has a Gaussian distribution with mean μ_{c_i} and variance 1. Marginalizing on c, conditionally on μ, the observations $(X_i)_{1\leqslant i\leqslant n}$ are i.i.d. and the conditional probability density of X_1 is:

$$x \mapsto p(x|\mu) = \sum_{k=1}^K \omega_k \varphi_{\mu_k,1}(x)\,,$$

where φ_{μ_k,η^2} the Gaussian probability density function with mean μ_k and variance η^2. The joint likelihood is therefore:

$$p(x_{1:n}) = \int p(x_{1:n}|\mu)p(\mu)\mathrm{d}\mu = \int \prod_{i=1}^n p(x_i|\mu)p(\mu)\mathrm{d}\mu = \int \prod_{i=1}^n \left(\sum_{k=1}^K \omega_k \varphi_{\mu_k,1}(x_i)\right) p(\mu)\mathrm{d}\mu.$$

Writing $z = (\mu, c)$, our objective is to estimate $p(\mu, c|x)$ where $c = (c_1, \cdots, c_n)$ are the components of the observations. Consider the following 'mean-field' approximation:

$$q(\mu, c) = \prod_{k=1}^K \varphi_{m_k,s_k}(\mu_k) \prod_{i=1}^n \mathrm{Cat}_{\phi_i}(c_i)\,,$$

which means that under q:

- μ and c are independent.
- $(\mu_k)_{1\leqslant k\leqslant K}$ are independent Gaussian random variables with means $(m_k)_{1\leqslant k\leqslant K}$ and variances $(s_k)_{1\leqslant k\leqslant K}$.
- $(c_i)_{1\leqslant i\leqslant n}$ are independent with multinomial distributions with parameters $(\phi_i)_{1\leqslant i\leqslant n}$.

Write $\mathcal{D}$ this family where the means $(m_k)_{1\leqslant k\leqslant K} \in \mathbb{R}^K$, and variances $(s_k)_{1\leqslant k\leqslant K} \in (\mathbb{R}_+^*)^K$ and the parameters $(\phi_i)_{1\leqslant i\leqslant n} \in \mathbb{S}_K^n$ where $\mathbb{S}_K$ is the simplex of dimension K. Then, we aim at solving the optimization problem:

$$q^* = \mathrm{Argmin}_{q\in\mathcal{D}}\, \mathrm{KL}\left(q \| p(\cdot|x)\right)\,.$$

Note that

$$\begin{aligned}\mathrm{KL}\left(q\|p(\cdot|x)\right) &= \mathbb{E}_q[\log q(\mu,c)] - \mathbb{E}_q[\log p(\mu,c|x)]\,,\\ &= \mathbb{E}_q[\log q(\mu,c)] - \mathbb{E}_q[\log p(\mu,c,x)] + \log p(x)\,,\\ &= -\mathcal{L}_x(q) + \log p(x)\,,\end{aligned}$$

where

$$\mathcal{L}_x(q) = -\mathbb{E}_q[\log q(\mu,c)] + \mathbb{E}_q[\log p(\mu,c,x)]\,.$$

CAVI algorithm computes iteratively $1 \leqslant k \leqslant K$,

$$q(\mu_k) \propto \exp\left(\mathbb{E}_{\tilde{q}_{\mu_k}}[\log p(\mu_k|x,c,\mu_{-k})]\right)$$

and for all $1 \leqslant i \leqslant n$,

$$q(c_i) \propto \exp\left(\mathbb{E}_{\tilde{q}_{c_i}}[\log p(c_i|x,c_{-i},\mu)]\right)\,,$$

where $\mu_{-k} = (\mu_j)_{1\leq j\leq K, j\neq k}$, $c_{-i} = (c_j)_{1\leq j\leq n, j\neq i}$, and $\mathbb{E}_{\tilde{q}_z}$ is the expectation under the variational distribution of all variables except z.

Update of the variaitonal distribution of c_i, $1 \leq i \leq n$

Note that

$$p(c_i|x,c_{-i},\mu) \propto p(c_i)p(x_i|c_i,\mu) \propto p(c_i)\prod_{k=1}^{K}\left(\varphi_{\mu_k,1}(x_i)\right)^{1_{c_i=k}}\,.$$

Then,

$$\mathbb{E}_{\tilde{q}_{c_i}}[\log p(c_i|x,c_{-i},\mu)] = \log p(c_i) + \sum_{k=1}^{K} 1_{c_i=k}\mathbb{E}_{\tilde{q}_{c_i}}[\log\varphi_{\mu_k,1}(x_i)]$$

and

$$\begin{aligned}\exp\left(\mathbb{E}_{\tilde{q}_{c_i}}[\log p(c_i|x,c_{-i},\mu)]\right) &\propto p(c_i)\exp\left(\sum_{k=1}^{K} 1_{c_i=k}\mathbb{E}_{\tilde{q}_{c_i}}[\log\varphi_{\mu_k,1}(x_i)]\right)\\ &\propto p(c_i)\exp\left(\sum_{k=1}^{K} 1_{c_i=k}\mathbb{E}_{\tilde{q}_{c_i}}[-(x_i-\mu_k)^2/2]\right)\,.\end{aligned}$$

The update writes:

$$\phi_i(k) \propto \omega_k \exp\left(m_k x_i - \frac{m_k^2 + s_k}{2}\right)\,.$$

Update of the variaitonal distribution of μ_k, $1 \leq k \leq K$

On the other hand,

$$p(\mu_k|x,c,\mu_{-k}) \propto p(\mu_k)\prod_{i=1}^{n} p(x_i|c_i,\mu)\,.$$

Then,

$$\mathbb{E}_{\tilde{q}_{\mu_k}}[\log p(\mu_k|x,c,\mu_{-k})] = \log p(\mu_k) + \sum_{i=1}^{n}\mathbb{E}_{\tilde{q}_{\mu_k}}[\log p(x_i|\mu,c_i)]$$

and

$$\begin{aligned}\exp\left(\mathbb{E}_{\tilde{q}_{\mu_k}}[\log p(\mu_k|x,c,\mu_{-k})]\right) &\propto p(\mu_k)\exp\left(\sum_{i=1}^{n}\sum_{\ell=1}^{K}\mathbb{E}_{\tilde{q}_{\mu_k}}[1_{c_i=\ell}\log\varphi_{\mu_\ell,1}(x_i)]\right)\\ &\propto p(\mu_k)\exp\left(\sum_{i=1}^{n}\phi_i(k)\mathbb{E}_{\tilde{q}_{\mu_k}}[\log\varphi_{\mu_k,1}(x_i)]\right)\\ &\propto \exp\left(-\frac{\mu_k^2}{2\sigma^2}-\frac{1}{2}\sum_{i=1}^{n}\phi_i(k)(x_i-\mu_k)^2\right),\\ &\propto \exp\left(-\frac{\mu_k^2}{2\sigma^2}+\sum_{i=1}^{n}\phi_i(k)x_i\mu_k-\frac{1}{2}\sum_{i=1}^{n}\phi_i(k)\mu_k^2\right).\end{aligned}$$

The update writes therefore,

$$m_k=\frac{\sum_{i=1}^{n}\phi_i(k)x_i}{1/\sigma^2+\sum_{i=1}^{n}\phi_i(k)}\quad\text{and}\quad s_k=\frac{1}{1/\sigma^2+\sum_{i=1}^{n}\phi_i(k)}.$$

Data: Observations $x=(x_1,\ldots,x_n)$, initial values of $\phi_i(k)$, m_k and s_k, $1\le k\le K$, $1\le i\le n$, maximum number of iteration N
Result: A variational distribution for each coordinate of μ_k and c_i, $1\le k\le K$, $1\le i\le n$.
for $p=1\to N$ **do**
  for $i=1\to n$ **do**
    Set
    $$\phi_i(k)\propto\omega_k\exp\left(m_kx_i-\frac{m_k^2+s_k}{2}\right).$$
    ;
  end
  for $k=1\to K$ **do**
    Set
    $$m_k=\frac{\sum_{i=1}^{n}\phi_i(k)x_i}{1/\sigma^2+\sum_{i=1}^{n}\phi_i(k)}\quad\text{and}\quad s_k=\frac{1}{1/\sigma^2+\sum_{i=1}^{n}\phi_i(k)}.$$
    ;
  end
end

Algorithm 4: A version of CAVI algorithm for a Bayesian mixture of Gaussian distributions.

3.4 Variational Autoencoders

Variational Auto-Encoders (VAE) are very popular approaches to introduce approximations of a target conditional distribution in the context of latent data models. Assume that $(X_1,\ldots,X_n)$ are i.i.d. random variables in X with unknown probability distribution function π_{data}. We consider a family of joint probability distributions $\{(z,x)\mapsto p_\theta(z,x)\}_{\theta\in\Theta}$ on $(\mathsf{Z}\times\mathsf{X},\mathcal{Z}\times\mathcal{X})$ where Z is a latent variable and X is the observation. In this setting, we often write, for all $\theta\in\Theta$, $x\in\mathsf{X}$, $z\in\mathsf{Z}$,

$$p_\theta(z,x)=p_\theta(z)p_\theta(x|z).$$

The latent variable generative model defines a joint density $(z,x)\mapsto p_\theta(x,z)$ on $(\mathsf{Z}\times\mathsf{X},\mathcal{Z}\times\mathcal{X})$ by specifying a prior $z\mapsto p_\theta(z)$ over the latent variable Z and a conditional density $x\mapsto p_\theta(x|z)$ also referred to as the decoder. The normalized loglikelihood is therefore given by

$$\ell_n(\theta)=\frac{1}{n}\sum_{i=1}^{n}\log p_\theta(X_i)=\frac{1}{n}\sum_{i=1}^{n}\log\int p_\theta(z)p_\theta(X_i|z)\mathrm{d}z,$$

and the conditional distribution $p_\theta(z|x) \propto p_\theta(z)p_\theta(x|z)$. In most cases, maximizing the average marginal log-likelihood of the data is not possible, as the marginal likelihood functions $p_\theta(X_i)$, $1 \leqslant i \leqslant n$, are not vailable explicitly as the integral for marginalizing the latent variable is intractable. Since a maximum likelihood estimator cannot be computed simply, VAEs introduce a variational approach wich aims at simultaneously providing a parameter estimate and an approximation of the conditional distribution of the latent variable given the observation. Consider a family of probability density functions $\{(z,x) \mapsto q_\varphi(z|x)\}_{\varphi \in \Phi}$. Then, we can write, for all $\varphi \in \Phi, \theta \in \Theta$, $x \in \mathsf{X}$,

$$\begin{aligned}\log p_\theta(x) = \int \log p_\theta(x) q_\varphi(z|x) \mathrm{d}z &= \mathbb{E}_{q_\varphi(\cdot|x)}\left[\log p_\theta(x)\right] \\ &= \mathbb{E}_{q_\varphi(\cdot|x)}\left[\log \frac{p_\theta(Z,x)}{p_\theta(Z|x)}\right] \\ &= \mathbb{E}_{q_\varphi(\cdot|x)}\left[\log \frac{q_\varphi(Z|x)}{p_\theta(Z|x)}\right] + \mathbb{E}_{q_\varphi(\cdot|x)}\left[\log \frac{p_\theta(Z,x)}{q_\varphi(Z|x)}\right].\end{aligned}$$

The first term of the right-hand-side is the Kullback-Leibler divergence between $q_\varphi(\cdot|x)$ and $p_\theta(\cdot|x)$, so that $\log p_\theta(x) \geq \mathcal{L}(\theta, \varphi, x)$, where

$$\mathcal{L}(\theta, \varphi, x) = \mathbb{E}_{q_\varphi(\cdot|x)}\left[\log \frac{p_\theta(Z,x)}{q_\varphi(Z|x)}\right]$$

is the Evidence Lower BOund (ELBO). This motivates the introduction of the following loss function:

$$\mathcal{L}(\theta, \varphi) = \mathbb{E}_{\pi_{\mathrm{data}}}[-\mathcal{L}(\theta, \varphi, X)] = \mathbb{E}_{\pi_{\mathrm{data}}}\left[\mathbb{E}_{q_\varphi(\cdot|X)}\left[\log \frac{q_\varphi(Z|X)}{p_\theta(Z,X)}\right]\right].$$

The empirical loss is then given by

$$\mathcal{L}_n(\theta, \varphi) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{q_\varphi(\cdot|X_i)}\left[\log \frac{q_\varphi(Z|X_i)}{p_\theta(Z,X_i)}\right],$$

where $(X_1, \ldots, X_n)$ are i.i.d. with distribution π_{data}, and we aim at solving the optimization problem:

$$(\widehat{\theta}_n, \widehat{\varphi}_n) \in \mathrm{Argmax}_{\theta \in \Theta, \varphi \in \Phi} \mathcal{L}_n(\theta, \varphi). \tag{3.1}$$

The joint optimization of θ and φ is a complex problem both for practical and theoretical reasons and many research works have been devoted to this problem in the past few years. In most cases, $\mathcal{L}_n(\theta, \varphi)$ cannot be computed explicitly since expectations under the variational distribution are not explicit. Therefore, $\mathcal{L}_n(\theta, \varphi)$ is replaced by a Monte Carlo estimate $\widehat{\mathcal{L}}_n(\theta, \varphi)$:

$$\widehat{\mathcal{L}}_n(\theta, \varphi) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{M}\sum_{j=1}^{M} \log \frac{q_\varphi(Z_{i,j}|X_i)}{p_\theta(Z_{i,j}, X_i)},$$

where for all $1 \leqslant i \leqslant n$, $(Z_{i,1}, \ldots, Z_{i,M})_{1 \leqslant j \leqslant M}$ are i.i.d. with distribution $q_\varphi(\cdot|X_i)$.

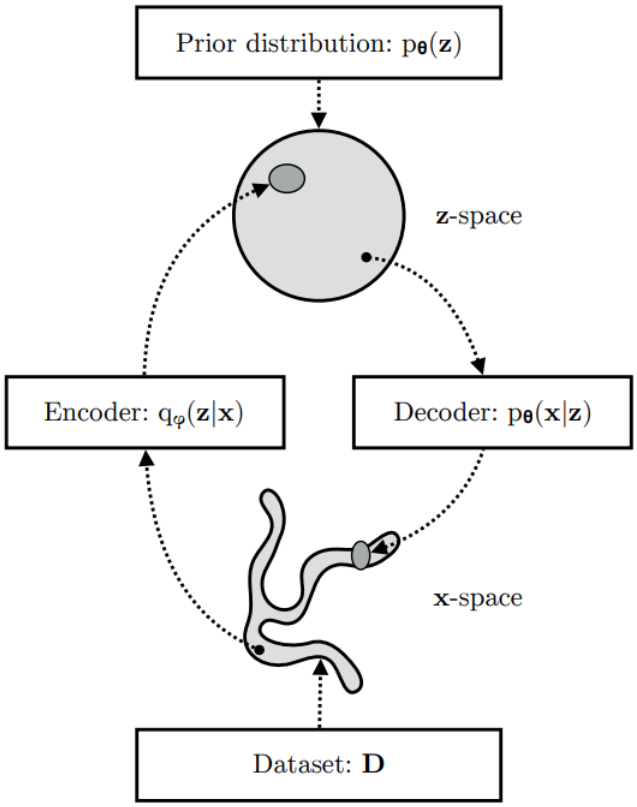

Fig. 3.1 An illustration of a VAE. From "An Introduction to Variational Autoencoders", Kingma et al., 2019.

Appendix A

M-estimation Z-estimation, maximum likelihood

A.1 Method of moments

Consider a measurable space $(\Omega, \mathcal{F})$ and i.i.d. random variables $(X_1, \dots, X_n)$ taking values in a measurable space $(\mathsf{X}, \mathcal{X})$. We assume that we have access to probabilities $(\mathbb{P}_\theta)_{\theta\in\Theta}$, where $\Theta \subset \mathbb{R}^d$. For all $\theta \in \Theta$, we write $\mathbb{E}_\theta$ the expectation under $\mathbb{P}_\theta$ and $\mathbb{V}_\theta$ the variance. The objective is to estimate the unknown parameter $\theta \in \Theta$. The method of moments consists in choosing d functions $T_j : \mathsf{X} \to \mathbb{R}$, $1 \le j \le d$, such that $\mathbb{E}_\theta[|T_j(X_1)|] < \infty$. Then, write for all $1 \le j \le d$, $\theta \in \Theta$,

$$e_j(\theta) = \mathbb{E}_\theta[T_j(X_1)].$$

As the quantities $e_j(\theta)$, $1 \le j \le d$, $\theta \in \Theta$, are usually unknown, they may be estimated by using empirical estimates. Assuming that for $1 \le j \le d$, $\mathbb{E}_\theta[|T_j(X_1)|^2] < \infty$, the Bienayme-Tchebychev inequality allows to quantify the empirical estimation error: for all $\varepsilon > 0$,

$$\mathbb{P}_\theta\left(\left|\frac{1}{n}\sum_{i=1}^n T_j(X_i) - e_j(\theta)\right| \ge \varepsilon\right) \le \frac{\mathbb{V}_\theta[T_j(X_1)]}{n\varepsilon^2}.$$

In order to estimate the unknown parameter θ we may consider the system of equations:

$$\forall j \in \{1, \dots, d\}, \quad e_j(\theta) = \frac{1}{n}\sum_{i=1}^n T_j(X_i).$$

Assuming that this system has a unique solution $\widehat{\theta}_n$, $\widehat{\theta}_n$ is referred to as the moment estimator associated with $\{T_j\}_{1\le j\le d}$.

Example A.1. Let $(X_1, \dots, X_n)$ be i.i.d. random variables with exponential distribution with parameter $\theta > 0$. Using $T_1 : x \mapsto x$ and $T_2 : x \mapsto x^2$ we have for all $\theta > 0$,

$$e_1(\theta) = \theta^{-1} \quad \text{and} \quad e_2(\theta) = 2\theta^{-2}.$$

The moment estimator associated with T_1 is

$$\widehat{\theta}_{n,1} = \frac{n}{\sum_{i=1}^n X_i}.$$

The moment estimator associated with T_2 is

$$\widehat{\theta}_{n,2} = \left(\frac{2n}{\sum_{i=1}^{n} X_i^2} \right)^{1/2}.$$

A.2 Z-estimation

The moment estimator associated with $\{T_j\}_{1 \leq j \leq d}$ is a solution to a system of equations of the form

$$\frac{1}{n} \sum_{i=1}^{n} \psi(\theta, X_i) = 0,$$

where for all $\theta \in \Theta$, $x \in \mathsf{X}$,

$$\psi(\theta, x) = \begin{pmatrix} T_1(x) - \mathbb{E}_\theta[T_1(X_1)] \\ \vdots \\ T_d(x) - \mathbb{E}_\theta[T_d(X_1)] \end{pmatrix}.$$

Consider now arbitrary functions ψ_j, $1 \leq j \leq d$, such that for all $\theta_* \in \Theta$, $1 \leq j \leq d$, $\mathbb{E}_{\theta_*}[|\psi_j(\theta_*, X_1)|] < \infty$. A Z-estimator associated with $\psi = (\psi_1, \ldots, \psi_d)^\top$ is any solution $\widehat{\theta}_n$ satisfying

$$\psi_n(\widehat{\theta}_n) = 0,$$

where for all $\theta \in \Theta$,

$$\psi_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \psi(\theta, X_i).$$

Example A.2. Let F be a distribution function on $\mathbb{R}$ such that for all $x \in \mathbb{R}$, $F(x) = 1 - F(-x)$. Let $(X_1, \ldots, X_n)$ be i.i.d with distribution function F_{θ_*} where for all $\theta \in \mathbb{R}$, $x \in \mathbb{R}$, $F_\theta(x) = F(x - \theta)$. In this setting,

$$\mathbb{E}_\theta[X_1] = \theta,$$

which suggests to choose $\psi(\theta, x) = x - \theta$. In this case, the Z-estimator associated with ψ is given by $\widehat{\theta}_n = n^{-1} \sum_{i=1}^{n} X_i$.

A.3 Maximum likelihood

Definition A.3. Let $(\mathsf{X}, \mathcal{X})$ be a measurable space equipped with a sigma-finite measure μ. Let $(f_\theta)_{\theta \in \Theta}$ be a family of probability densities with respect to μ and $(X_i)_{1 \leq i \leq n}$ be i.i.d. random variables with probability density f_{θ_*}, $\theta_* \in \Theta$. The likelihood of $(X_i)_{1 \leq i \leq n}$ is the function

$$L_n : \theta \mapsto \prod_{i=1}^{n} f_\theta(X_i).$$

A maximum likelihood estimator associated with L_n is any estimator solution to the following optimization problem

$$\widehat{\theta}_n \in \mathrm{Argmax}_{\theta \in \Theta} L_n(\theta).$$

Example A.4. Let $(X_i)_{1 \leq i \leq n}$ be i.i.d. Bernoulli random variables with parameter $\theta_* \in (0, 1)$. For all $\theta \in (0, 1)$,

$$L_n(\theta) = \prod_{i=1}^{n} \theta^{X_i} (1 - \theta)^{1 - X_i}$$

and

$$\ell_n(\theta) = \log L_n(\theta) = \left(\sum_{i=1}^n X_i\right)\log\theta + \left(\sum_{i=1}^n (1-X_i)\right)\log(1-\theta).$$

The function ℓ_n/n is stricly concave on $(0,1)$ with $\lim_{\theta\to 0}\ell_n(\theta)/n = -\infty$ and $\lim_{\theta\to 1}\ell_n(\theta)/n = -\infty$. This function has therefore a unique maximum given by

$$\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i.$$

A.4 M-estimation

Maximum likelihood estimators are defined as solutions to optimization problems. This is the case of many estimation procedures. Consider for instance a function $m : \Theta \times \mathsf{X} \to \mathbb{R}$, $(\theta, x) \mapsto m_\theta(x)$, such that for all $\theta, \theta_* \in \Theta$, $\mathbb{E}_{\theta_*}[|m_\theta(X_1)|] < \infty$ and consider also $M_n : \theta \mapsto n^{-1}\sum_{i=1}^n m_\theta(X_i)$. For all $\delta > 0$,

$$\mathbb{P}_{\theta_*}\left(|M_n(\theta) - M_{\theta_*}(\theta)| \geq \delta\right) \leq \frac{\mathbb{V}_{\theta_*}[m_\theta(X_1)]}{n\delta^2},$$

where

$$M_{\theta_*}(\theta) = \mathbb{E}_{\theta_*}[m_\theta(X_1)].$$

A M-estimator is any solution to the following optimization problem:

$$\widehat{\theta}_n \in \mathrm{Argmax}_{\theta\in\Theta} M_n(\theta).$$

Example A.5. For all $1 \leq k \leq n$, let $x_k \in \mathbb{R}^d$ and consider $(\xi_k)_{1\leq k\leq n}$ i.i.d. random variables with distribution $\mathcal{N}(0,1)$ and the linear regression model:

$$Y_k = \sum_{\ell=0}^p \beta_\ell \varphi_\ell(x_k) + \sigma\varepsilon_k,$$

where $\theta = (\sigma, \beta) \in \mathbb{R}_+^* \times \mathbb{R}^{p+1}$. The joint density of the observations is:

$$f_n : \theta \mapsto (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^n\left(Y_k - \sum_{\ell=0}^p \beta_\ell\varphi_\ell(x_k)\right)^2\right).$$

The maximum likelihood estimator of β coincides with the mean squared error estimator

$$\widehat{\beta}_n \in \mathrm{Argmin}_{\beta\in\mathbb{R}^{p+1}} \sum_{k=1}^n\left(Y_k - \sum_{\ell=0}^p \beta_\ell\varphi_\ell(x_k)\right)^2.$$

Consider the matrix Φ in $\mathbb{R}^{n\times(p+1)}$ such that for all $1 \leq i \leq n$, $1 \leq j \leq p+1$, $\Phi_{i,j} = \varphi_{j-1}(x_i)$. Then, $\widehat{\beta}_n$ is solution to

$$\Phi\Phi^\top\widehat{\beta}_n = \Phi Y,$$

where $Y = (Y_1, \ldots, Y_n)^\top$.

A.5 Consistency

When for all θ, θ_*, $\mathbb{E}_{\theta_*}[|m_\theta(X_1)|] < \infty$, by the law of large numbers, in $\mathbb{P}_{\theta_*}$-probability,

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m_\theta(X_i) \to_{n\to\infty} M_{\theta_*}(\theta) = \mathbb{E}_{\theta_*}[m_\theta(X_1)].$$

We also assume that θ_* is a maximum of M_{θ_*}.

Theorem A.6. *Consider the following assumptions.*

- *For all $\theta_* \in \Theta$, in $\mathbb{P}_{\theta_*}$-probability, $\sup_{\theta\in\Theta}|M_n(\theta) - M_{\theta_*}(\theta)| \to_{n\to\infty} 0$.*
- *For all $\theta_* \in \Theta$ and $\varepsilon > 0$,*

$$\sup_{\theta\in\Theta;|\theta-\theta_*|>\varepsilon} M_{\theta_*}(\theta) < M_{\theta_*}(\theta_*).$$

- *$(\widehat{\theta}_n)_{n\geq 0}$ is such that there exists $(\rho_n)_{n\geq 0}$ satisfying for all $\theta_* \in \Theta$, in $\mathbb{P}_{\theta_*}$-probability, $\rho_n \to_{n\to\infty} 0$ and*

$$\liminf_{n\to\infty} \mathbb{P}_{\theta_*}\left(M_n(\widehat{\theta}_n) \geq M_n(\theta_*) - \rho_n\right) = 1.$$

Then, for all $\theta_* \in \Theta$, in $\mathbb{P}_{\theta_*}$-probability, $\widehat{\theta}_n \to \theta_*$.

Proof. For all $\theta_* \in \Theta$, since θ_* is a maximum of M_{θ_*},

$$\begin{aligned} 0 \leq M_{\theta_*}(\theta_*) - M_{\theta_*}(\widehat{\theta}_n) &\leq M_{\theta_*}(\theta_*) - M_n(\theta_*) + M_n(\theta_*) - M_n(\widehat{\theta}_n) + M_n(\widehat{\theta}_n) - M_{\theta_*}(\widehat{\theta}_n) \\ &\leq 2\sup_{\theta\in\Theta}|M_n(\theta) - M_{\theta_*}(\theta)| + \rho_n \\ &\quad + \left\{M_n(\theta_*) - M_n(\widehat{\theta}_n) - \rho_n\right\} \mathbb{1}_{M_n(\theta_*)-\rho_n > M_n(\widehat{\theta}_n)}. \end{aligned}$$

Let $\varepsilon > 0$. There exists $\eta > 0$ such that $M_{\theta_*}(\theta) \leq M_{\theta_*}(\theta_*) - \eta$ for all $\theta \in \Theta$ such that $|\theta - \theta_*| \geq \varepsilon$. Therefore, $\{|\widehat{\theta}_n - \theta_*| \geq \varepsilon\} \subset \{M_{\theta_*}(\widehat{\theta}_n) \leq M_{\theta_*}(\theta_*) - \eta\}$. This yields

$$\mathbb{P}_{\theta_*}\left(|\widehat{\theta}_n - \theta_*| \geq \varepsilon\right) \leq \mathbb{P}_{\theta_*}\left(M_{\theta_*}(\widehat{\theta}_n) \leq M_{\theta_*}(\theta_*) - \eta\right) \leq \mathbb{P}_{\theta_*}\left(M_{\theta_*}(\theta_*) - M_{\theta_*}(\widehat{\theta}_n) > \eta\right),$$

which concludes the proof. □

Remark A.7. If Θ is compact in $\mathbb{R}^d$, M_{θ_*} is continuous, and for all $\theta \neq \theta_*$, $M_{\theta_*}(\theta) < M_{\theta_*}(\theta_*)$, the second assumption is satisfied.

Exponential models

Let $(X_1, \ldots, X_n)$ be i.i.d. random variables with density p_{θ_*} with respect to a reference measure μ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. The family $\{p_\theta\}_{\theta\in\Theta}$ is said to be in the exponential family if there exist $\eta : \theta \to \mathbb{R}^d$, $T : \mathsf{X} \to \mathbb{R}^d$, $h : \mathsf{X} \to \mathbb{R}_+$, $B : \Theta \to \mathbb{R}$ such that for all $x \in \mathsf{X}$,

$$p_\theta(x) = h(x)\exp\left(\langle\eta(\theta); T(x)\rangle - B(\theta)\right).$$

Example A.8.

- The density of a Poisson distribution with parameter $\theta > 0$ is given by

$$p_\theta : x \mapsto \frac{\theta^x}{x!}\mathrm{e}^{-\theta},$$

  so that $h(x) = (x!)^{-1}$, $T(x) = x$, $\eta(\theta) = \log\theta$, $B(\theta) = -\theta$.
- If $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$ and p_θ is the Gaussian probability density with mean μ and variance σ^2, $h(x) = 1$, $T(x) = (x, x^2)^\top$, $\eta(\theta) = (\mu/\sigma^2, -1/(2\sigma^2))^\top$, $B(\theta) = \log(2\pi\sigma^2)/2 + \mu/(2\sigma^2)$.

The canonical exponential family is given, for all $x \in \mathsf{X}$, by

$$p_\eta(x) = h(x)\exp\left(\langle\eta; T(x)\rangle - A(\eta)\right),$$

where

$$A(\eta) = \log\left(\int h(x) \exp\left(\langle \eta; T(x) \rangle\right) \mu(\mathrm{d}x)\right).$$

References

Blei et al., 2017. Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Dempster et al., 1977. Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Roberts and Rosenthal, 1998. Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B*, 60(1):255–268.

Roberts and Tweedie, 1996. Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363.

Wu, 1983. Wu, C. J. (1983). On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103.

Zangwill, 1969. Zangwill, W. I. (1969). Convergence conditions for nonlinear programming algorithms. *Management Science*, 16(1):1–13.