Sylvain Le Corff

Introduction to computational statistics

Contents

1 Markov chain Monte Carlo 5
1.1 Introduction 5
1.2 Key elements on Markov chains 6
1.3 Metropolis-Hastings algorithm 7
1.4 Variants of MH algorithms 9
1.4.1 Metropolis–Adjusted Langevin Algorithm 9
1.4.2 Generalisation of MH Algorithms 10
1.4.3 Pseudo-marginal Monte Carlo methods 11
1.4.4 Hamiltonian Monte Carlo 12

2 Expectation Maximization algorithm 17
2.1 Introduction 17
2.2 Algorithm 17
2.3 Convergence properties 18
2.4 Application to latent data models 20
2.5 Example: mixture of Gaussian distributions 21
2.6 Monte Carlo EM in the curved exponential family setting 23

3 Sequential Monte Carlo Methods 25
3.1 Importance sampling 25
3.1.1 Convergence guarantees for the IS estimator 26
3.1.2 Convergence guarantees for the self-normalized IS estimator 27
3.2 Sequential importance sampling 28
3.2.1 Hidden Markov models 28
3.2.2 Sequential Monte Carlo methods for HMMs 30

4 Variational inference and autoencoders 33
4.1 Evidence Lower Bound 33
4.2 Coordinate ascent variational inference 34
4.3 Application to a mixture of Gaussian distributions 35
4.4 Variational Autoencoders 37
4.5 Gradient-based optimization 38
4.5.1 Pathwise Gradient 39
4.5.2 Score Function Gradient 40

A M-estimation Z-estimation, maximum likelihood 43
A.1 Method of moments 43
A.2 Z-estimation 44
A.3 Maximum likelihood 44
A.4 M-estimation 45

A.5 Consistency 45
References 47

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 $(\mathsf{X}^2, \mathcal{X}^{\otimes 2})$,

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

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 μ and by abuse of notation, we also denote by π this density, that is we write $\pi(\mathrm{d}x) = \pi(x)\mu(\mathrm{d}x)$ and we assume that this density π is positive. Moreover, let Q be a Markov kernel on $\mathsf{X} \times \mathcal{X}$ such that $Q(x, \mathrm{d}y) = q(x, y)\mu(\mathrm{d}y)$, that is, for any $x \in \mathsf{X}$, $Q(x, \cdot)$ is also dominated by μ 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

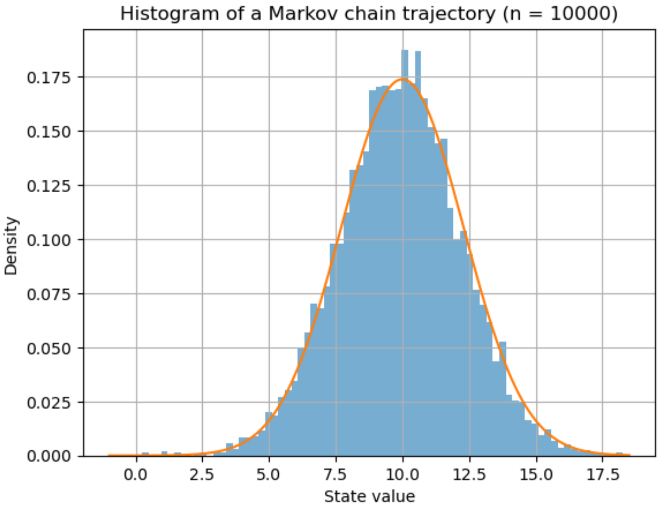


Fig. 1.1 Histogram of a Gaussian AR(1) process, $X_k = \mu + \phi X_{k-1} + \sigma Z_k$, where $(Z_k)_{k\in\mathbb{N}}$ is an iid sequence of standard Gaussian random variables, independent of X_0 with $|\phi| < 1$. The Gaussian distribution with mean $\mu/(1-\phi)$ and variance $\sigma^2/(1-\phi^2)$ is a stationary distribution of the Markov chain. Illustration of the histogram and of the stationary distribution with $\mu = 1$, $\phi = 0.9$ and $\sigma = 1$.

```
Input  : Initial distribution $\mu$, maximum number of iterations $n$.
Output: Markov chain $X_0, \dots, X_n$.
At iteration $k = 0$, draw $X_0 \sim \mu$.
for $k = 0$ to $n - 1$ do
    • Draw independently $Y_{k+1} \sim Q(X_k, \cdot)$ and $U_{k+1} \sim \mathrm{U}(0,1)$.
    • Set $X_{k+1} = \begin{cases} Y_{k+1} & \text{if } U_{k+1}, \leq \alpha(X_k, Y_{k+1}) \\ X_k & \text{otherwise.} \end{cases}$
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 $\mathcal{F}_k = \sigma(X_0, U_{1:k}, Y_{1:k})$ and note that $(X_k)_{k\in\mathbb{N}}$ is adapted to the filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$ (which is equivalent to $\sigma(X_{0:k}) \subset \mathcal{F}_k$). 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 $k \in \mathbb{N}$,

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

Therefore, $\{X_k : k \in \mathbb{N}\}$ 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 for all measurable bounded or non-negative functions h on $\left(\mathsf{X}^2, \mathcal{X}^{\otimes 2}\right)$,*

$$\int_{\mathsf{X}^2} h(x,y)\pi(\mathrm{d}x)Q(x,\mathrm{d}y)\alpha(x,y) = \int_{\mathsf{X}^2} h(x,y)\pi(\mathrm{d}y)Q(y,\mathrm{d}x)\alpha(y,x). \tag{1.4}$$

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 $k \geq 0$, given the current state X_k, a proposal Y_{k+1} is generated according to the Euler discretization of the overdamped Langevin dynamics,

$$Y_{k+1} = X_k + \frac{h^2}{2}\nabla \log \pi(X_k) + hZ_k,$$

where $Z_k \sim \mathcal{N}(0, I_d)$ 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

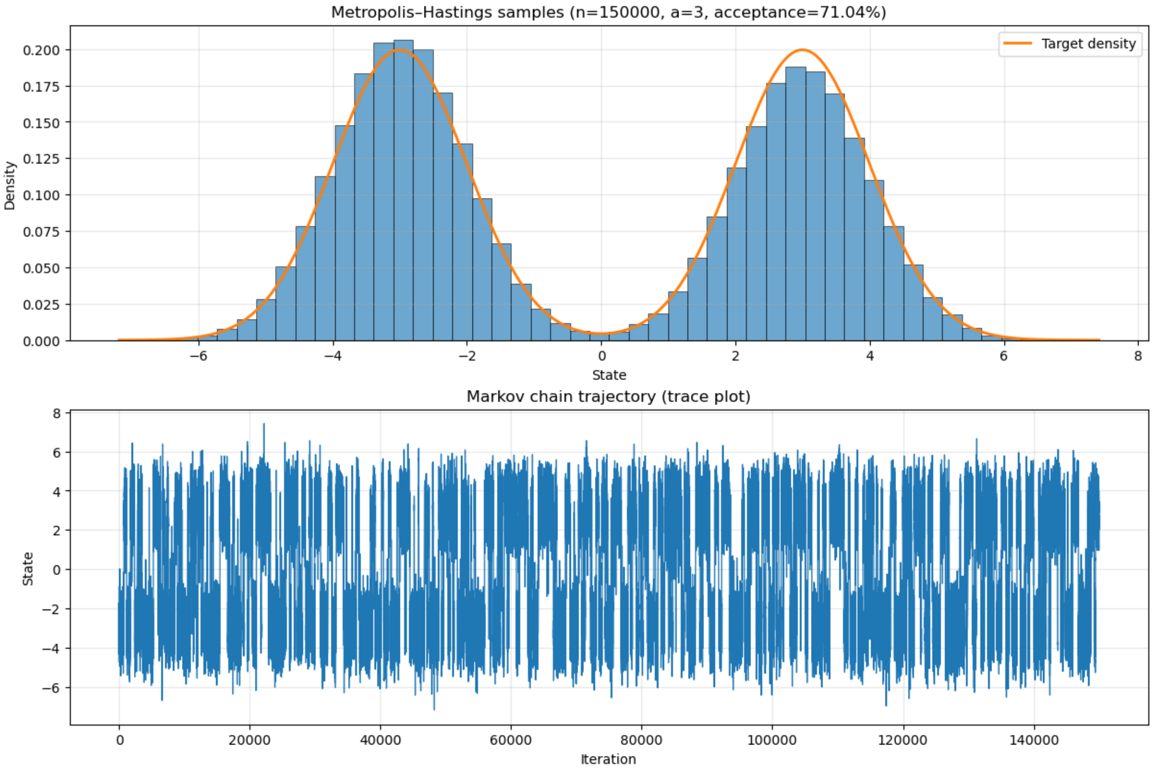


Fig. 1.2 Trajectory of a random walk MH sampler targetting a mixture of two Gaussian distributions, one centered at $a > 0$ and the other one centered at $-a$. At each iteration $k \geq 1$, the proposal distribution is a Gaussian centered at the current state X_k with unit variance. Example with $a = 3$

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

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 μ. 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 μ and write by abuse of notation, $\pi(\mathrm{d}x) = \pi(x)\mu(\mathrm{d}x)$ and $Q(x,\mathrm{d}y) = q(x,y)\mu(\mathrm{d}y)$. 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 $\mu$, total number of samples $n$.
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 $\bar{\pi}(\mathrm{d}\bar{x}) = \bar{\pi}(\mathrm{d}x\mathrm{d}w) = wR(x, \mathrm{d}w)\mu(\mathrm{d}x)$ (where we set $\bar{x} = (x, w)$). Note that $\bar{\pi}$ is indeed a probability measure on $\bar{\mathsf{X}} = \mathsf{X} \times \mathbb{R}_\star^+$, since

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

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

$$\mu_0(\mathrm{d}\bar{x}\mathrm{d}\bar{x}') = \mu_0(\mathrm{d}x\mathrm{d}w\mathrm{d}x'\mathrm{d}w') = wR(x, \mathrm{d}w)\mu(\mathrm{d}x)Q(x, \mathrm{d}x')R(x', \mathrm{d}w')$$
$$\mu_1(\mathrm{d}\bar{x}\mathrm{d}\bar{x}') = \mu_1(\mathrm{d}x\mathrm{d}w\mathrm{d}x'\mathrm{d}w') = w'R(x', \mathrm{d}w')\mu(\mathrm{d}x')Q(x', \mathrm{d}x)R(x, \mathrm{d}w).$$

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

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

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 $P^{MH}_{\langle \bar{\pi}, \bar{Q} \rangle}(\bar{x}, \mathrm{d}\bar{x}')$ is $\bar{\pi}$-reversible. Since Algorithm 2 corresponds to applying the Markov kernel $P^{MH}_{\langle \bar{\pi}, \bar{Q} \rangle}$, this completes the proof. Note that the extended target distribution Π has the marginal π with respect to the first component:

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

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 $\bar{\pi}$ 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).

1.4.4 Hamiltonian Monte Carlo

In Hamiltonian Monte Carlo (HMC), we extend the target density and construct a Markov chain on an extended space. We assume that the state space is $\mathbb{R}^d$ and that $\pi(q) \propto \mathrm{e}^{-U(q)}$ (in most of the HMC literature, the state variable x is replaced by q). A classical setting is to consider π as the marginal of the extended target

$$\bar{\pi}(q,p) \propto \exp\left\{-U(q) - p^\top p/2\right\}, \quad p,q \in \mathbb{R}^d. \tag{1.7}$$

In this straightforward setting, the extended target density can be written as the product of π and the standard Gaussian probability density of $\mathcal{N}(0, I_d)$. In what follows, we use the following terminology, very classical in the HMC literature.

- $q \in \mathbb{R}^d$ is the position and $U(q)$ is the called the *potential energy*.
- $p \in \mathbb{R}^d$ is the momentum and $K(p) = p^\top p/2$ is called the *kinetic energy*.
- $H(q,p) = U(q) + K(p)$ is called the *Hamiltonian*.

Several versions of HMC exist. We consider in this course the Leapfrog HMC which produces a Markov chain $(X_k)_{k\in\mathbb{N}} = (q_k, p_k)_{k\in\mathbb{N}}$ and where a transition of this algorithm can be decomposed into two different moves.

- The first transition sets $X^0_{k+1} = (q^0_{k+1}, p^0_{k+1})$ where the first component is freezed, i.e. $q^0_{k+1} = q_k$, while the second one is sampled as $p^0_{k+1} \sim \mathcal{N}(0, I_d)$. As the standard Gaussian distribution is the conditional law of p given q in (1.7)) this is a Gibbs move which means that the transition kernel associated with this move is $\bar{\pi}$-reversible.
- The second transition sets $X^L_{k+1} = (q^L_{k+1}, p^L_{k+1})$ deterministically from X^0_{k+1} using L steps, and then accepts or rejects the proposed candidate according to some well-chosen probability. Therefore, we need to establish how to choose the acceptance rate with deterministic moves to obtain a $\bar{\pi}$-reversible transition.

MH with deterministic moves

A natural question is therefore to understand if we can construct a MH algorithm with target $\bar{\pi}$ and where the proposal candidate is deterministic: $Q(x, \mathrm{d}y) = \delta_{\varphi(x)}(\mathrm{d}y)$, where $\varphi : \mathbb{R}^d \to \mathbb{R}^d$. Due to the Dirac mass, we are not in a standard framework but we can use the generalized version of MH algorithm as described in Section 1.4.2. Set

$$\mu_0(\mathrm{d}x\mathrm{d}y) = \bar{\pi}(\mathrm{d}x)\delta_{\varphi(x)}(\mathrm{d}y) \quad \text{and} \quad \mu_1(\mathrm{d}x\mathrm{d}y) = \bar{\pi}(\mathrm{d}y)\delta_{\varphi(y)}(\mathrm{d}x),$$

and write for any non-negative function h,

$$\begin{aligned}\int h(x,y)\mu_1(\mathrm{d}x\mathrm{d}y) &= \int \bar{\pi}(\mathrm{d}u)h(\underbrace{\varphi(u)}_{v}, \underbrace{u}_{\varphi^{-1}(v)}) \\ &= \int \bar{\pi}\circ\varphi^{-1}(\mathrm{d}v)h(v,\varphi^{-1}(v)) = \int \frac{\mathrm{d}\bar{\pi}\circ\varphi^{-1}}{\mathrm{d}\bar{\pi}}(v)\bar{\pi}(\mathrm{d}v)h(v,\varphi^{-1}(v)).\end{aligned}$$

If we want to let appear the integral of h with respect to $\mu_0(\mathrm{d}x\mathrm{d}y) = \bar{\pi}(\mathrm{d}x)\delta_{\varphi(x)}(\mathrm{d}y)$, we need to assume that $\varphi^{-1}(v) = \varphi(v)$ that is φ is an involution and in such a case:

$$\int h(x,y)\mu_1(\mathrm{d}x\mathrm{d}y) = \int h(x,y)\frac{\mathrm{d}\bar{\pi}\circ\varphi^{-1}}{\mathrm{d}\bar{\pi}}(x)\mu_0(\mathrm{d}x\mathrm{d}y)$$

and the acceptance probability is then

$$\alpha(x,y) = \frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_0}(x,y) \wedge 1 = \frac{\mathrm{d}\bar{\pi}\circ\varphi^{-1}}{\mathrm{d}\bar{\pi}}(x) \wedge 1.$$

A first point is that if we only use the involution, then after two steps we obtain the initial state. Therefore, this deterministic transition is often combined with another move that is not deterministic. For any involution, we can get a Metropolis Hastings with a theoretical expression of the acceptance probability as

$$\alpha(x,y) = \frac{\mathrm{d}\bar{\pi}\circ\varphi^{-1}}{\mathrm{d}\bar{\pi}}(x) \wedge 1$$

but in an ideal HMC we can choose φ so that that this is equal to 1. If $\bar{\pi}$ has a density with respect to the Lebesgue measure that we still denote $\bar{\pi}$, we get

$$\frac{\mathrm{d}\bar{\pi}\circ\varphi^{-1}}{\mathrm{d}\bar{\pi}}(x) = \frac{\bar{\pi}(\varphi^{-1}(x))}{\bar{\pi}(x)}\left|\frac{\partial\varphi^{-1}(x)}{\partial x}\right|,$$

where the second term is the Jacobian determinant of the mapping $\varphi^{-1} = \varphi$. To get 1 in the acceptance probability, we can impose that the two terms are equal to 1.

- The first term $\frac{\bar{\pi}(\varphi^{-1}(x))}{\bar{\pi}(x)}$ is one if the involution stays on the same level set (i.e. the moves according to φ does not change the value of $\bar{\pi}$).
- The second term $\left|\frac{\partial\varphi^{-1}(x)}{\partial x}\right|$ is one if the involution is volume-preserving. If for example the involution only keeps the volume then the Radon Nikodym simplifies to

$$\frac{\bar{\pi}(\varphi^{-1}(x))}{\bar{\pi}(x)} = \frac{\bar{\pi}(\varphi(x))}{\bar{\pi}(x)} \quad \text{since} \quad \varphi\circ\varphi = \mathrm{I}.$$

Hamiltonian dynamics

In order to propose a deterministic mapping, assume that the state depends on a real parameter t and we impose to stay on a level set of H, we get:

$$\frac{\mathrm{d}H(q_t,p_t)}{\mathrm{d}t} = 0 = \sum_{i=1}^{d}\frac{\partial H(q_t,p_t)}{\partial q_{t,i}}\frac{\mathrm{d}q_{t,i}}{\mathrm{d}t} + \frac{\partial H(q_t,p_t)}{\partial p_{t,i}}\frac{\mathrm{d}p_{t,i}}{\mathrm{d}t}.$$

This motivates the use the following dynamics: for all $1 \le i \le d$,

$$\begin{aligned}\frac{\partial H}{\partial q_{t,i}}(q_t,p_t) &= \frac{\partial U(q_t)}{\partial q_{t,i}} = -\frac{\mathrm{d}p_{t,i}}{\mathrm{d}t}\\ \frac{\partial H}{\partial p_{t,i}}(q_t,p_t) &= \frac{\partial K(p_t)}{\partial p_{t,i}} = p_{t,i} = \frac{\mathrm{d}q_{t,i}}{\mathrm{d}t}.\end{aligned} \tag{1.8}$$

It can be shown that this Hamiltonian dynamics moves along the same level sets and that it is volume-preserving. But unfortunately, it is not an involution. However, it is not possible to compure (q_t,p_t) and we need to use a discretization scheme to approximate solutions to the Hamiltonian dynamics.

The leapfrog integrator

Recall that for all $1 \le i \le d$,

$$\frac{\partial U(q_t)}{\partial q_{t,i}} = -\frac{\mathrm{d}p_{t,i}}{\mathrm{d}t}, \quad \text{and} \quad p_{t,i} = \frac{\mathrm{d}q_{t,i}}{\mathrm{d}t}.$$

A first idea of discretization would be to choose a small stepsize h and a number of steps L and compute, for $0 \le k \le L-1$,

$$\begin{aligned}p^{k+1} &= p^k - h\nabla U(q^k)\\ q^{k+1} &= q^k + hp^k.\end{aligned}$$

Unfortunately, this discretization is associated with a mapping that is not volume-preserving. That is the absolute value of the Jacobian determinant of the mapping is not equal to one. This is the reason why we use in practice the leapfrog discretization, defined by the following scheme: for $0 \leq k \leq L-1$,

$$\begin{aligned} p^{k+1/2} &= p^k - (h/2)\nabla U(q^k) \\ q^{k+1} &= q^k + hp^{k+1/2} \\ p^{k+1} &= p^{k+1/2} - (h/2)\nabla U(q^{k+1}). \end{aligned} \tag{1.9}$$

To see that it is volume-preserving, just note that a Leapfrog update can be decomposed into three mapping

$$(q^k, p^k) \xrightarrow{\varphi_1} (q^k, p^{k+1/2}) \xrightarrow{\varphi_2} (q^{k+1}, p^{k+1/2}) \xrightarrow{\varphi_1} (q^{k+1}, p^{k+1}), \tag{1.10}$$

where

$$\varphi_1(x,y) = (x, y - (h/2)\nabla U(x)) \quad \text{and} \quad \varphi_2(x,y) = (x + hy, y). \tag{1.11}$$

Each of these mappings is volume-preserving and so is the Leapfrog update. To sum-up, the Leapfrog update is an approximation of the Hamiltonian dynamics, it is a deterministic mapping that is volume-preserving which justifie the acceptance rate of the HMC algorithm.

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

$$L(\theta) = \int f_\theta(x)\mu(\mathrm{d}x).$$

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 $q_\theta : x \mapsto f_\theta(x)/L(\theta)$. 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:

$$Q(\theta, \theta') = \int \log f_\theta(x) q_{\theta'}(x)\mu(\mathrm{d}x) = \mathbb{E}_{\theta'}[\log f_\theta(X)],$$

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

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

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

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 $\theta,\theta' \in \Theta$,

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

As log is concave, by Jensen's inequality,

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

which concludes the proof. The inequality $\int \log(q_\theta(x))q_{\theta'}(x)\mu(\mathrm{d}x) \leq \int \log(q_{\theta'}(x))q_{\theta'}(x)\mu(\mathrm{d}x)$ 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 $\theta_0$.
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

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 Lemma 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. Assumptions H1 ensures that $\{\theta_k\}_{k\geq 0}$ is contained in a compact subset of Θ.

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.

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 there exists a set S such that:

(i) M is a closed point-to-set mapping over the complement of S;
(ii) for all $p \geq 0$, $L(\theta_{p+1}) > L(\theta_p)$ if $\theta_p \notin S$.

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

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

By assumption H2, we can define the set of stationary points of the log-likelihood:

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

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*

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

Proof. The proof is a consequence of Theorem 2.3. 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 $y \in M(\theta)$, $\alpha(y) \geq \alpha(\theta)$. Let $p \geq 0$ and $\theta_p \in \Theta \setminus \mathcal{S}$. 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

Let (X,Y) in $(\mathsf{X}\times\mathsf{Y},\mathcal{X}\otimes\mathcal{Y})$ be two random variables where Y is observed and X is not observed. The measurable space $(\mathsf{X}\times\mathsf{Y},\mathcal{X}\otimes\mathcal{Y})$ is endowed with the product measure $\mu\otimes\lambda$. A very popular setting for the EM algorithm is when $f_\theta : x\mapsto p_\theta(x,Y)$ where for all $\theta\in\Theta$, p_θ is a joint probability density function with respect to $\mu\otimes\lambda$ used to model the joint distribution of (X,Y). Assuming that the random variable Y is observed and that X is not observed, we consider the likelihood function

$$L(\theta) = \int f_\theta(x,Y)\mu(\mathrm{d}x),$$

which is a random variable depending on Y. This is the marginal density of Y when the parameter is θ. In this setting, $f_\theta(X,Y)/L(\theta)$ is the probability density of the conditional distribution of X given Y.

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

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

Therefore, the E-step of the EM algorithm amounts to computing the conditional expectation given Y of the complete data (joint) loglikelihood.

If $\{(X_i,Y_i)\}_{1\le i\le n}$ are i.i.d we write $Y=(Y_1,\ldots,Y_n)$ and $X=(X_1,\ldots,X_n)$. In this case, the intermediate quantity of the EM algorithm is

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

Curved exponential families

We may further restrict the analysis to the case where the complete-data likelihood p_θ belongs to the class of curved exponential family densities. This setting covers a wide range of practical models. We now introduce a set of assumptions that are satisfied in many scenarios. Let $\phi:\Theta\to\mathbb{R}$, $\psi:\Theta\to\mathbb{R}^q$, $h:\mathsf{X}\times\mathsf{Y}\to\mathbb{R}_+^*$ and $S:\mathsf{X}\times\mathsf{Y}\to\mathbb{R}^q$. Assume then that the complete-data likelihood may be written:

$$p_\theta(x,y) = h(x,y)\mathrm{exp}\left\{\phi(\theta)+\langle S(x,y);\psi(\theta)\rangle\right\}.$$

Therefore, the intermediate quantity of the EM algorithm is given by

$$Q(\theta,\theta') = \int \log p_\theta(x,Y)p_{\theta'}(x|Y)\mu(\mathrm{d}x) = \mathbb{E}_{\theta'}[\log h(X,Y)|Y] + \phi(\theta) + \mathbb{E}_{\theta'}[\langle S(X,Y);\psi(\theta)\rangle|Y].$$

If $p\ge 0$ and θ_p is the current parameter estimate, θ_{p+1} may be computed by solving

$$\theta_{p+1}\in\arg\max_{\theta\in\Theta}\ \left\{\phi(\theta)+\left\langle\mathbb{E}_{\theta_p}[S(X,Y)|Y];\psi(\theta)\right\rangle\right\}.$$

2.5 Example: mixture of Gaussian distributions

In this example, we assume that the joint distribution of (X, Y) belongs to a family of distributions parametrized by a vector θ with real components. For $k \in \{1, \ldots, M\}$, write $\pi_k = \mathbb{P}_\theta(X = k)$. Assume that conditionally on the event $\{X = k\}$, Z 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 $\mathbb{S}_K = \{(\pi_1, \ldots, \pi_K) \in [0,1]^K \ ; \ \sum_{k=1}^K \pi_k = 1\}$. For all $1 \le k \le K$, the explicit computation of $\mathbb{P}_\theta(X = k|Y)$ writes

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

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

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

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

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

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

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

and the update is decomposed into two steps.

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

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

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:

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

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

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

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

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

and

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

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

$$\pi_k^{(p+1)} = \frac{1}{n}\sum_{i=1}^{n}\omega_{i,k}^{(p)},$$

$$\mu_k^{(p+1)} = \frac{1}{\sum_{i=1}^{n}\omega_{i,k}^{(p)}}\sum_{i=1}^{n}\omega_{i,k}^{(p)}Y_i,$$

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

Remark 2.5. Note that the computation of $\mathbb{P}_{\theta^{(p)}}(X_i = k|Y_i)$ 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 X_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 $i \in \{1,\ldots,d\}$ 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 $\Sigma\ [\Delta_{j,i}]_{1\le i,j\le d} = (\det\Sigma)\ I_d$ so that $\Sigma^{-1} = \frac{[\Delta_{j,i}]_{1\le i,j\le d}^\top}{\det\Sigma}$, we finally get

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

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 in the curved exponential family setting

Assume that the complete-data log-likelihood belongs to a curved exponential family, i.e.,

$$\log p_\theta(X,Y) = \log h(x,y) + \phi(\theta) + \langle S(X,Y), \psi(\theta)\rangle,$$

where $S(X,Y) \in \mathbb{R}_+^*$, $S(X,Y) \in \mathbb{R}^q$, $\phi : \Theta \to \mathbb{R}$, $\psi : \Theta \to \mathbb{R}^q$. In this case,

$$Q(\theta, \theta^{(k)}) = \phi(\theta) + \mathbb{E}_{\theta^{(k)}}[\log h(X,Y) \mid Y] + \langle \mathbb{E}_{\theta^{(k)}}[S(X,Y) \mid Y],\ \psi(\theta)\rangle\,.$$

Hence, the E-step reduces to computing the conditional expectation of the sufficient statistics, and the M-step consists in solving

$$\theta^{(k+1)} = \arg\max_{\theta\in\Theta} \left\{\phi(\theta) + \left\langle \bar{S}(\theta^{(k)}),\ \psi(\theta)\right\rangle\right\}, \qquad \bar{S}(\theta^{(k)}) := \mathbb{E}_{\theta^{(k)}}[S(X,Y) \mid Y].$$

When $\bar{S}(\theta^{(k)})$ is not available in closed form, the Monte Carlo EM (MCEM) algorithm estimates it with a simulation-based approximation. At iteration k, let

$$X_{k,1}, \ldots, X_{k,M_k} \sim p_{\theta^{(k)}}(X|Y)$$

be (approximately) independent draws from the conditional distribution of X given Y under $\theta^{(k)}$. Define the Monte Carlo estimate

$$\widehat{S}^{(k)}(\theta^{(k)}) := \frac{1}{M_k}\sum_{m=1}^{M_k} S(X_{k,m}, Y)\,.$$

The corresponding update is obtained by the approximate M-step

$$\theta^{(k+1)} \in \arg\max_{\theta\in\Theta} \left\{\phi(\theta) + \left\langle \widehat{S}^{(k)}(\theta^{(k)}),\ \psi(\theta)\right\rangle\right\}.$$

Under standard regularity conditions, $\widehat{S}^{(k)}$ converges almost surely to $\mathbb{E}_{\theta^{(k)}}[S(X,Y) \mid Y]$ as $M_k \to \infty$, and convergence to a stationary point of the observed-data log-likelihood can be ensured when the number of Monte Carlo samples M_k increases sufficiently with k. Convergence rates of the MCEM algorithm can be obtained when nonasymptotic controls of the Monte Carlo estimation $\widehat{S}^{(k)}$ can be derived, see for instance [Fort and Moulines, 2003].

In many models, direct sampling from $p_{\theta^{(k)}}(X \mid Y)$ is not feasible but $p_{\theta^{(k)}}(X \mid Y)$ is known up to a normalizing constant. A common alternative is to approximate the conditional expectation in the E-step using a Markov chain Monte Carlo (MCMC) method targeting $p_{\theta^{(k)}}(X \mid Y)$. Specifically, at iteration k, let $\{X_{k,t}\}_{p\geq 1}$ be a Markov chain with invariant distribution $p_{\theta^{(k)}}(X \mid Y)$ (e.g., a Metropolis–Hastings or Gibbs sampler). After a burn-in period B_k, an ergodic average yields

$$\widehat{S}^{(k)}_{\mathrm{MCMC}}(\theta^{(k)}) := \frac{1}{M_k}\sum_{p=B_k+1}^{B_k+M_k} S(X_{k,p}, Y)\,,$$

which replaces $\bar{S}(\theta^{(k)})$ in the E-step. The resulting MCMC-EM update is

$$\theta^{(k+1)} \in \arg\max_{\theta\in\Theta} \left\{\phi(\theta) + \left\langle \widehat{S}^{(k)}_{\mathrm{MCMC}}(\theta^{(k)}),\ \psi(\theta)\right\rangle\right\}.$$

Compared to MCEM with independent samples, the estimator $\widehat{S}^{(k)}_{\mathrm{MCMC}}$ is typically correlated due to the Markov dependence, but consistency follows from the ergodic theorem under standard assumptions (irreducibility, aperiodicity, and geometric ergodicity). In practice, the accuracy of the E-step approximation can be controlled by the burn-in length B_k, the chain length M_k.

Chapter 3

Sequential Monte Carlo Methods

The use of Monte Carlo methods for non-linear filtering can be traced back to the pioneering work of [Handschin and Mayne, 1969]. These early approaches relied on sequential versions of importance sampling, where samples are generated from an instrumental (proposal) distribution and the target distribution is approximated by assigning appropriate importance weights to these samples. In the context of non-linear filtering, importance sampling can be implemented sequentially by carefully defining a sequence of proposal distributions. This makes it unnecessary to regenerate the entire set of samples whenever a new observation becomes available. This approach is known as Sequential Importance Sampling (SIS).

Although SIS was already known in the early 1970s, its application to non-linear filtering problems remained limited for many years. This was partly due to insufficient computational power at the time, but also to a more fundamental issue that was not fully understood until later: weight degeneracy, also referred to as sample impoverishment. As the number of iterations increases, most samples tend to receive very small normalized importance weights and therefore contribute negligibly to the approximation of the target distribution.

A major breakthrough came with the work of [Gordon et al., 1993], who proposed addressing this problem through resampling. Their approach rejuvenates the particle population by duplicating samples with high importance weights while eliminating those with low weights. This led to the development of the particle filter, which represented the first successful application of Sequential Monte Carlo (SMC) methods to non-linear filtering.

Since then, SMC methods have been widely adopted across many fields, including computer vision, signal processing, control, econometrics, robotics, and statistics. The purpose of this chapter is to review the fundamental components required to implement sequential Monte Carlo methods.

3.1 Importance sampling

Throughout this section, μ will denote a probability measure of interest on a measurable space $(\mathsf{X}, \mathcal{X})$, which we shall refer to as the target distribution. The aim is to approximate integrals of the form

$$\mu(f) = \int_{\mathsf{X}} f(x)\, \mu(\mathrm{d}x),$$

for real-valued measurable functions f. The Monte Carlo approach presented in the first part of the course consists in drawing an i.i.d. sample $(\xi_1, \ldots, \xi_N)$ from the probability measure μ and then evaluating the sample mean $\sum_{i=1}^{N} f(\xi_i)/N$. Of course, this technique is applicable only when it is possible (and reasonably simple) to sample from the target distribution μ.

Importance sampling is based on the idea that it may be simpler or computationally cheaper to sample from an instrumental distribution ν, and then to apply a change-of-measure. Assume that the target probability measure μ is absolutely continuous with respect to an instrumental probability measure ν from which sampling is easily feasible. Then, for any μ-integrable function f,

$$\mu(f) = \int f(x)\,\mu(\mathrm{d}x) = \int f(x)\,\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\,\nu(\mathrm{d}x).$$

In particular, if $(\xi_1, \ldots, \xi_N)$ is an i.i.d. sample from ν, we can introduce the following estimator of $\mu(f)$:

$$\tilde{\mu}_N^{\mathrm{IS}}(f) = \frac{1}{N}\sum_{i=1}^{N} f(\xi_i)\,\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\xi_i).$$

In many practical situations, the target probability measure μ is only known up to a normalizing constant. This situation arises in particular when applying importance sampling techniques to hidden Markov models and, more generally, in Bayesian statistics. In such cases, the Radon–Nikodym derivative $\mathrm{d}\mu/\mathrm{d}\nu$ is only available up to a multiplicative constant. Nevertheless, the importance sampling methodology can still be applied by resorting to the self-normalized importance sampling estimator, defined as

$$\hat{\mu}_N^{\mathrm{IS}}(f) = \frac{\sum_{i=1}^{N} f(\xi_i)\,\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\xi_i)}{\sum_{i=1}^{N}\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\xi_i)}.$$

3.1.1 Convergence guarantees for the IS estimator

Since the importance sampling estimator can be written as a sample mean of independent random variables, a wide range of theoretical results is available to quantify the accuracy of $\tilde{\mu}_{\nu,N}^{\mathrm{IS}}(f)$ as an estimator of $\mu(f)$. Some of these results are asymptotic in nature, such as the law of large numbers (LLN) and the central limit theorem (CLT). It is also possible to derive non-asymptotic guarantees, including Berry–Esseen type bounds, bounds on the moments of the error $\mathbb{E}\big|\tilde{\mu}_N^{\mathrm{IS}}(f) - \mu(f)\big|^p$ for $p > 0$, or bounds on tail probabilities.

Theorem 3.1. *Let f be a real-valued measurable function such that $\mu(|f|) < \infty$ and $|f|\mu \ll |f|\nu$, and let $(\xi_i)_{i\geq 1}$ be i.i.d. with distribution ν. Then,*

$$\lim_{N\to\infty} \tilde{\mu}_N^{\mathrm{IS}}(f) = \mu(f), \qquad \textit{almost surely.}$$

Assume in addition that

$$\int f^2(x)\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\right)^2 \nu(\mathrm{d}x) < \infty. \tag{3.1}$$

Then,

$$\sqrt{N}\big(\tilde{\mu}_N^{\mathrm{IS}}(f) - \mu(f)\big) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \mathrm{Var}_\nu\left(f\,\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\right), \qquad \textit{as } N \to \infty,$$

where

$$\mathrm{Var}_\nu\left(f\,\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) = \int \left(f(x)\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x) - \mu(f)\right)^2 \nu(\mathrm{d}x).$$

Although the importance sampling estimator is generic, its efficiency depends crucially on the interplay between the target distribution μ, the instrumental distribution ν, and the function f. In particular, for a given function f, it is possible to define an instrumental distribution ν that leads to a significantly lower variance than that obtained with the standard Monte Carlo approach, corresponding to the choice $\nu = \mu$.

We may also use inequalities on the control of moments of sums of independent random variables to provide non asymptotic bound for the importance sampling estimator.

Theorem 3.2 (Marcinkiewicz–Zygmund inequality). *Let $(X_1, \dots, X_n)$ be independent random variables and let $p \geq 2$. Then,*

$$\mathbb{E}\left[\left|\sum_{i=1}^{n}\big(X_i - \mathbb{E}[X_i]\big)\right|^p\right] \leq C_p\, n^{p/2-1} \sum_{i=1}^{n} \mathbb{E}[|X_i - \mathbb{E}[X_i]|^p]\,, \tag{3.2}$$

for some positive constant C_p depending only on p.

Controlling tail probabilities is often of utmost interest. This topic has also been extensively studied, and a canonical result in this area is Hoeffding's inequality.

Theorem 3.3 (Hoeffding inequality). *Let $(X_1, \dots, X_n)$ be independent bounded random variables such that $\mathbb{P}(a_i \leq X_i \leq b_i) = 1$ for all $i = 1, \dots, n$. Then, for any $t \geq 0$,*

$$\mathbb{P}\left(\sum_{i=1}^{n}\big(X_i - \mathbb{E}[X_i]\big) \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right),$$

and

$$\mathbb{P}\left(\sum_{i=1}^{n}\big(X_i - \mathbb{E}[X_i]\big) \leq -t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

These inequalities make it possible to derive non-asymptotic bounds on both moments and tail probabilities of importance sampling estimators.

Theorem 3.4. *Let $p \geq 2$ and $N \geq 1$. Then,*

$$\mathbb{E}\,\left[\left|\tilde{\mu}_N^{\mathrm{IS}}(f) - \mu(f)\right|^p\right] \;\leq\; C_p\, N^{-p/2}\,,$$

where $C_p < \infty$ depends only on p. Moreover, for any $t \geq 0$,

$$\mathbb{P}\Big(\left|\tilde{\mu}_N^{\mathrm{IS}}(f) - \mu(f)\right| \geq t\Big) \;\leq\; 2\exp\left(-\frac{Nt^2}{2\|f\, \mathrm{d}\mu/\mathrm{d}\nu\|_\infty^2}\right).$$

3.1.2 Convergence guarantees for the self-normalized IS estimator

Theorem 3.5. *Let f be a measurable function such that $\mu(|f|) < \infty$, and assume that $\mu \ll \nu$. Let $(\xi_i)_{i \geq 1}$ be i.i.d. with distribution ν. Then,*

$$\hat{\mu}_N^{\mathrm{IS}}(f) \xrightarrow{a.s.} \mu(f), \qquad \text{as } N \to \infty.$$

Assume in addition that

$$\int \left[1 + f(x)^2\right] \left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\right)^2 \nu(\mathrm{d}x) < \infty. \tag{3.3}$$

Then,

$$\sqrt{N}\,\left(\hat{\mu}_N^{\mathrm{IS}}(f) - \mu(f)\right) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}\big(0, \sigma^2(\nu, f)\big), \qquad \text{as } N \to \infty,$$

where

$$\sigma^2(\nu, f) = \int \left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\right)^2 \big(f(x) - \mu(f)\big)^2 \nu(\mathrm{d}x). \tag{3.4}$$

Proof. Note that

$$\hat{\mu}_N^{\mathrm{IS}}(f) = \frac{\sum_{i=1}^{N} f(\xi_i)\, \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\xi_i)}{\sum_{i=1}^{N} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\xi_i)} = \frac{N^{-1}\sum_{i=1}^{N} f(\xi_i)\, \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\xi_i)}{N^{-1}\sum_{i=1}^{N} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\xi_i)}.$$

Therefore, the first result is a direct consequence of

$$\frac{1}{N}\sum_{i=1}^{N} f(\xi_i)\,\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\xi_i) \xrightarrow{\text{a.s.}} \mu(f), \qquad \frac{1}{N}\sum_{i=1}^{N} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\xi_i) \xrightarrow{\text{a.s.}} 1.$$

For the central limit theorem, note that

$$\sqrt{N}\left(\hat{\mu}_N^{\mathrm{IS}}(f) - \mu(f)\right) = \frac{N^{-1/2}\sum_{i=1}^{N} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\xi_i)\big(f(\xi_i) - \mu(f)\big)}{N^{-1}\sum_{i=1}^{N} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\xi_i)}.$$

By thecentral limit theorem, the numerator converges in distribution to $\mathcal{N}(0, \sigma^2(\nu, f))$, while the denominator converges almost surely to 1. The proof is then completed with Slutsky's Lemma. □

Theorem 3.6. *Assume that* $\|\mathrm{d}\mu/\mathrm{d}\nu\|_{L^p(\nu)}^p < \infty$ *for some* $p \geq 2$. *Then, there exists a constant* $C_p < \infty$ *such that, for any* $N \geq 1$ *and any measurable function* f,

$$\mathbb{E}\left|\hat{\mu}_N^{\mathrm{IS}}(f) - \mu(f)\right|^p \leq C_p N^{-p/2}. \tag{3.5}$$

Moreover, there exist $c_1, c_2 > 0$ *such that for all* $t \geq 0$,

$$\mathbb{P}\Big(\left|\hat{\mu}_N^{\mathrm{IS}}(f) - \mu(f)\right| \geq t\Big) \leq c_1 \exp\left(-\frac{c_2 N t^2}{\|f\|_\infty^2}\right). \tag{3.6}$$

3.2 Sequential importance sampling

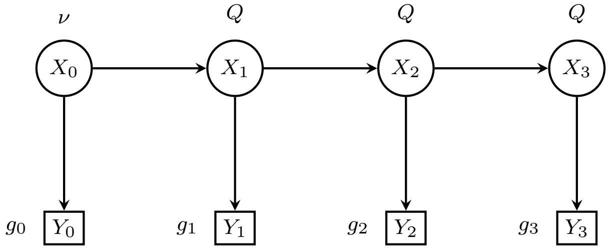
3.2.1 Hidden Markov models

A Hidden Markov model (HMM) is defined as a bivariate process $\{(X_k, Y_k)\}_{k\geq 0}$ such that:

- $\{X_k\}_{k\geq 0}$ is a Markov chain with transition kernel Q and initial distribution ν;
- Conditionally on the state process $\{X_k\}_{k\geq 0}$, the observations $\{Y_k\}_{k\geq 0}$ are independent, and for each n, the conditional distribution of Y_n depends only on X_n.

An alternative definition can be given as follows: a HMM is defined as a bivariate Markov chain that is only partially observed, whose transition kernel possesses a special structure: it is such that both the joint process $\{(X_k, Y_k)\}_{k\geq 0}$ and the marginal, hidden chain $\{X_k\}_{k\geq 0}$ are Markovian. A graphical representation of a HMM is given in Figure 3.2.1.

Fig. 3.1 A graphical model describing the joint distribution of a hidden Markov model.



Definition 3.7 (Hidden Markov Model). Let $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$ be two measurable spaces, and let Q and G denote, respectively, a Markov transition kernel on $(\mathsf{X}, \mathcal{X})$ and a transition kernel from $(\mathsf{X}, \mathcal{X})$

to $(\mathsf{Y}, \mathcal{Y})$. Define a Markov transition kernel T on the product space $(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$ as follows: for all bounded measurable functions f, and all $(x, y) \in \mathsf{X} \times \mathsf{Y}$,

$$Tf(x,y) := \int Q(x, \mathrm{d}x')\, G(x', \mathrm{d}y') f(x', y'). \tag{3.7}$$

Then, a Markov chain $\{(X_k, Y_k)\}_{k \geq 0}$ with transition kernel T is called a *hidden Markov model*.

We may assume that there exists a probability measure λ on $(\mathsf{Y}, \mathcal{Y})$ such that, for all $x \in \mathsf{X}$, $G(x, \cdot)$ is absolutely continuous with respect to λ, i.e., $G(x, \cdot) \ll \lambda(\cdot)$, with transition density function $g(x, \cdot)$. We also assume that there exists a probability measure μ on $(\mathsf{X}, \mathcal{X})$ such that for all $x \in \mathsf{X}$, $Q(x, \cdot) \ll \mu(\cdot)$ with transition density function $q(x, \cdot)$. The joint Markov transition kernel T is then dominated by the product measure $\mu \otimes \lambda$ and admits the transition density function

$$t((x,y),(x',y')) := q(x,x')\, g(x',y').$$

Proposition 3.8. *Let $\{(X_k, Y_k)\}_{k \geq 0}$ be a Markov chain on the product space $\mathsf{X} \times \mathsf{Y}$ with transition kernel T defined by (3.7) and initial distribution $\nu(\mathrm{d}x)G(x, \mathrm{d}y)$. Then, for any integer p, any ordered set of indices $k_1 < \cdots < k_p$, and any bounded measurable functions $f_1, \ldots, f_p$,*

$$\mathbb{E}\left[\prod_{i=1}^{p} f_i(Y_{k_i}) \,\middle|\, X_{k_1}, \ldots, X_{k_p}\right] = \prod_{i=1}^{p} \int_{\mathsf{Y}} f_i(y)\, G(X_{k_i}, \mathrm{d}y). \tag{3.8}$$

Proof. For all bounded measurable function h, we have

$$\begin{aligned}
\mathbb{E}\left[\prod_{i=1}^{p} f_i(Y_{k_i})\, h(X_{k_1}, \ldots, X_{k_p})\right] &= \int \nu(\mathrm{d}x_0) G(x_0, \mathrm{d}y_0) \prod_{i=1}^{k_p} Q(x_{i-1}, \mathrm{d}x_i) G(x_i, \mathrm{d}y_i) \\
&\quad \times \left[\prod_{i=1}^{p} f_i(y_{k_i})\right] h(x_{k_1}, \ldots, x_{k_p}) \\
&= \int \nu(\mathrm{d}x_0) \prod_{i=1}^{k_p} Q(x_{i-1}, \mathrm{d}x_i)\, h(x_{k_1}, \ldots, x_{k_p}) \\
&\quad \times \prod_{i \notin \{k_1, \ldots, k_p\}} \int_{\mathsf{Y}} G(x_i, \mathrm{d}y_i) \prod_{i \in \{k_1, \ldots, k_p\}} \int_{\mathsf{Y}} f_i(y_i) G(x_i, \mathrm{d}y_i),
\end{aligned}$$

which completes the proof. □

The joint probability of the unobservable states and the observations up to index n is given as follows. For any bounded measurable function f,

$$\begin{aligned}
\mathbb{E}\big[f(X_0, Y_0, \ldots, X_n, Y_n)\big] &= \int f(x_0, y_0, \ldots, x_n, y_n)\, \nu(\mathrm{d}x_0) g(x_0, y_0) \\
&\quad \times \prod_{k=1}^{n} Q(x_{k-1}, \mathrm{d}x_k) g(x_k, y_k)\, \lambda^{\otimes(n+1)}(\mathrm{d}y_0, \ldots, \mathrm{d}y_n),
\end{aligned} \tag{3.9}$$

where $\lambda^{\otimes(n+1)}$ denotes the $(n+1)$-fold product measure on $(\mathsf{Y}^{n+1}, \mathcal{Y}^{\otimes(n+1)})$. Marginalizing over the unobservable states $X_0, \ldots, X_n$ yields the distribution of the observations alone:

$$\mathbb{E}\big[f(Y_0, \ldots, Y_n)\big] = \int f(y_0, \ldots, y_n)\, L_n(y_0, \ldots, y_n)\, \lambda^{\otimes(n+1)}(\mathrm{d}y_0, \ldots, \mathrm{d}y_n), \tag{3.10}$$

where L_n is defined below and plays a central role in likelihood-based inference.

Definition 3.9 (Likelihood). The *likelihood* of the observations $(Y_0, \dots, Y_n)$ is the probability density function with respect to $\lambda^{\otimes(n+1)}$, given, for all $(y_0, \dots, y_n) \in \mathsf{Y}^{n+1}$, by

$$L_n(y_0, \dots, y_n) := \int \nu(\mathrm{d}x_0) g(x_0, y_0) \prod_{k=1}^{n} Q(x_{k-1}, \mathrm{d}x_k) g(x_k, y_k). \tag{3.11}$$

The *log-likelihood function* is defined, droping the dependency on $(y_0, \dots, y_n)$, as

$$\ell_n \ := \ \log L_n. \tag{3.12}$$

The likelihood function given in Definition 3.9 is not availbale in closed form. Therefore, in a setting where the law of the HMM depends on au unknown parameter $\theta \in \Theta$, and in the fully dominated case, the log-likelihood

$$\ell_n(\theta) := \log \int \nu_\theta(x_0) g_\theta(x_0, y_0) \prod_{k=1}^{n} Q_\theta(x_{k-1}, x_k) g_\theta(x_k, y_k) \mu^{\otimes(n+1)}(\mathrm{d}x_{0:n})$$

cannot be optimized easily. A natural solution is to estimate θ using the EM algorithm. Therefore, at iteration $p \geq 0$ with a current parameter estimate $\widehat{\theta}_p$ it is required to compute:

$$\begin{aligned} Q(\theta, \widehat{\theta}_p) &= \mathbb{E}_{\widehat{\theta}_p} \left[\log p_\theta(X_{0:n}, Y_{0:n}) | Y_{0:n} \right] \\ &= \mathbb{E}_{\widehat{\theta}_p} \left[\log \nu_\theta(X_0) g_\theta(X_0, Y_0) | Y_{0:n} \right] + \sum_{k=1}^{n} \mathbb{E}_{\widehat{\theta}_p} \left[\log q_\theta(X_{k-1}, X_k) g_\theta(X_k, Y_k) | Y_{0:n} \right]. \end{aligned}$$

This means that we need to sample (at least approximately) from the conditional distributions of (X_{k-1}, X_k) given $Y_{0:n}$ for a given parameter value $\widehat{\theta}_p$. This is in general not possible exactly and an appealing solution is to use Sequential Monte Carlo methods. In the next subsection we present these sequential methods, omitting the dependency on $\widehat{\theta}_p$ for simplicity.

3.2.2 Sequential Monte Carlo methods for HMMs

We now specialize the sampling techniques introduced above to hidden Markov models (HMMs). As in previous chapters, we consider the HMM where Q denotes the Markov transition kernel of the hidden chain, ν is the distribution of the initial state X_0, and $g(x, y)$, for $x \in \mathsf{X}$ and $y \in \mathsf{Y}$, denotes the conditional density of the observation given the state, with respect to the measure λ on $(\mathsf{Y}, \mathcal{Y})$. To simplify the notation, we will also use the shorthand

$$g_k(\cdot) := g(\cdot, Y_k).$$

In the following we consider the following distributions.

- For all $0 \leq k \leq n$, ϕ_k is the conditional distribution of X_k given $Y_{0:k}$, and is referred to as the filtering distribution at time k.
- For all $0 \leq k \leq \ell \leq n$, $\phi_{k:\ell|n}$ is the conditional distribution of $X_{k:\ell}$ given $Y_{0:n}$, and is referred to as the joint smoothing distribution of $X_{k:\ell}$ given $Y_{0:n}$.
- For all $0 \leq k \leq n$, $\phi_{k|n} = \phi_{k:k|n}$, is the marginal smoothing distribution at time k.

Note that for all bounded measurable functions f,

$$\phi_0(f) = \frac{\int f(x_0)\, g_0(x_0)\, \nu(\mathrm{d}x_0)}{\int g_0(x_0)\, \nu(\mathrm{d}x_0)}.$$

Therefore, $\phi_0(f)$ can be estimated using self-normalized importance sampling. Sample $(\xi_0^1, \dots, \xi_0^N)$ i.i.d. with probability density χ and associate each ξ_0^i with $\omega_0^i = g_0(\xi_0^i)/\chi(\xi_0^i)$ and compute

$$\phi_0^N(f) = \sum_{i=1}^N \frac{\omega_0^i}{\sum_{j=1}^N \omega_0^j} f(\xi_0^i).$$

In addition, for all $n \geq 0$ and all bounded measurable functions f,

$$\phi_{0:n|n}(f) = \frac{\int \nu(x_0) g_0(x_0) \prod_{k=1}^n q(x_{k-1}, x_k) g_k(x_k) f(x_{0:n}) \mu^{\otimes(n+1)}(\mathrm{d}x_{0:n})}{\int \nu(x_0) g_0(x_0) \prod_{k=1}^n q(x_{k-1}, x_k) g_k(x_k) \mu^{\otimes(n+1)}(\mathrm{d}x_{0:n})}.$$

Therefore, it is straightforward to note that for all $0 \leq k \leq n-1$,

$$\phi_{0:k+1|k+1}(x_{0:k+1}) \propto \phi_{0:k|k}(x_{0:k}) q(x_k, x_{k+1}) g_{k+1}(x_{k+1})$$

Therefore, for all bounded measurable functions f,

$$\phi_{k+1}(f) = \frac{\int \varphi_k(x_k) q(x_k, x_{k+1}) g_{k+1}(x_{k+1}) f(x_{k+1}) \mu(\mathrm{d}x_k) \mu(\mathrm{d}x_{k+1})}{\int \phi_k(x_k) q(x_k, x_{k+1}) g_{k+1}(x_{k+1}) \mu(\mathrm{d}x_k) \mu(\mathrm{d}x_{k+1})}.$$

As these integrals are intractable, if at time k we have access to a weighted sample $\{(\omega_k^i, \xi_k^i)\}_{1 \leq i \leq N}$ we may estimate φ_{k+1} by

$$\tilde{\phi}_{k+1}^N(x_{k+1}) \propto \sum_{i=1}^N \omega_k^i q(\xi_k^i, x_{k+1}) g_{k+1}(x_{k+1}).$$

This approximation suggests that an importance sampling estimator of $\phi_{k+1}(f)$ can be obtained as follows.

- For all $1 \leq i \leq N$ sample I_{k+1}^i in $\{1, \dots, N\}$ with probabilities proportional to $\{\omega_k^i\}_{1 \leq i \leq N}$.
- For all $1 \leq i \leq N$ sample $\xi_{k+1}^i \sim P(\xi_k^{I_{k+1}^i}, \cdot)$ where P is a proposal kernel.
- For all $1 \leq i \leq N$, compute

$$\omega_{k+1}^i = \frac{q(\xi_k^{I_{k+1}^i}, \xi_{k+1}^i) g_{k+1}(\xi_{k+1}^i)}{p(\xi_k^{I_{k+1}^i}, \xi_k^i)}.$$

Then, $\varphi_{k+1}(f)$ is estimated by

$$\phi_{k+1}^N(f) = \sum_{i=1}^N \frac{\omega_{k+1}^i}{\sum_{j=1}^N \omega_{k+1}^j} f(\xi_{k+1}^i).$$

Chapter 4

Variational inference and autoencoders

4.1 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:

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

4.2 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 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,

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

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
```

4.3 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 $\mu_{-k} = (\mu_j)_{1\le j\le K, j\neq k}$, $c_{-i} = (c_j)_{1\le j\le n, j\neq i}$, 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, $1 \le i \le n$

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

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

The update writes:

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

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

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 $\mu_k$ 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
            $\phi_i(k) \propto \omega_k \exp\left(m_k x_i - \frac{m_k^2 + s_k}{2}\right).$
        ;
    end
    for $k = 1 \to K$ do
        Set
            $m_k = \frac{\sum_{i=1}^{n}\phi_i(k)x_i}{1/\sigma^2 + \sum_{i=1}^{n}\phi_i(k)}$ and $s_k = \frac{1}{1/\sigma^2 + \sum_{i=1}^{n}\phi_i(k)}.$
        ;
    end
end
```

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

4.4 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 $\{(z,x) \mapsto q_\varphi(z|x)\}_{\varphi \in \varPhi}$. Then, we can write, for all $\varphi \in \varPhi, \theta \in \varTheta$, $x \in \mathsf{X}$,

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

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:

$$(\widehat{\theta}_n, \widehat{\varphi}_n) \in \mathrm{Argmax}_{\theta \in \varTheta, \varphi \in \varPhi} \mathcal{L}_n(\theta, \varphi). \tag{4.1}$$

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 $q_\varphi(\cdot|X_i)$.

4.5 Gradient-based optimization

In order to solve (4.1), we need to compute the gradient of the loss function. Under classical regularity assumptions:

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

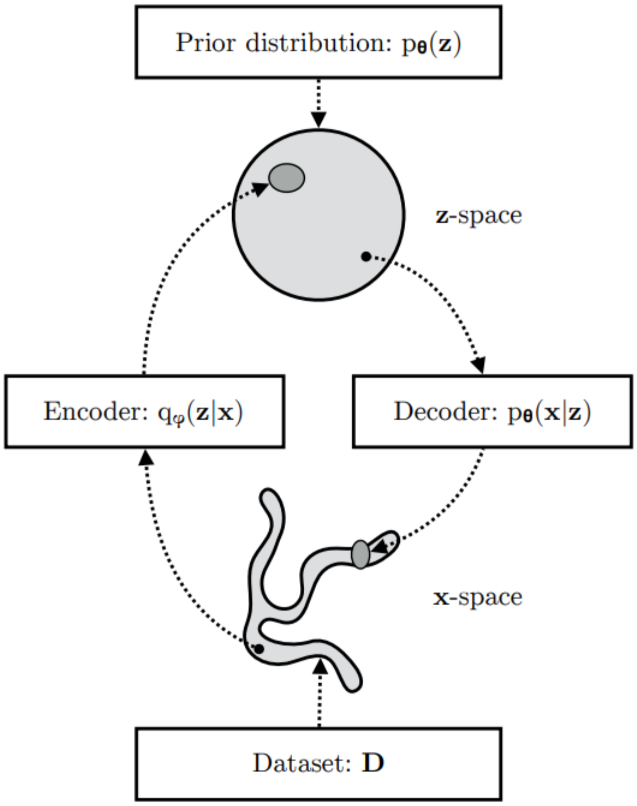


Fig. 4.1 An illustration of a VAE. From "An Introduction to Variational Autoencoders", Kingma et al., 2019.

This gradient can be estimated using samples from π_{data}. Given a batch of i.i.d. observations $(X_1, \ldots, X_n)_{1 \leqslant i \leqslant B}$ with distribution π_{data}, an estimator of $\nabla_\theta \mathcal{L}(\theta, \varphi)$ can be computed as follows:

$$S_\theta(\theta, \varphi; \{X_i\}_{i=1}^B) = -\frac{1}{B}\sum_{i=1}^B \frac{1}{M}\sum_{j=1}^M \nabla_\theta \log p_\theta(Z_{i,j}, X_i),$$

where for all $1 \leqslant i \leqslant B$, $(Z_{i,1}, \ldots, Z_{i,M})_{1 \leqslant j \leqslant M}$ are i.i.d. with distribution $q_\varphi(\cdot|X_i)$. Computing the gradient with respect to the variational parameter φ is more challenging since the inner expectation depends on q_φ. There are two common methods for computing this gradient.

4.5.1 Pathwise Gradient

The reparametrization trick involves expressing variational distribution using a deterministic transform $g(\varepsilon, \varphi)$, where ε is an auxiliary independent random variable drawn from a known probability density function p_ε. Using this trick, the ELBO can be expressed as:

$$\mathcal{L}(\theta, \varphi, x) = \mathbb{E}_{p_\varepsilon}\left[\log w_{\theta,\varphi}(x, g(\varepsilon, \varphi))\right],$$

where $w_{\theta,\varphi}(x, z) = p_\theta(x, z)/q_\varphi(z|x)$ the unnormalized importance weights and $\mathbb{E}_{p_\varepsilon}$ is the expectation under the law of ε when $\varepsilon \sim p_\varepsilon$. The pathwise gradient [Kingma and Welling, 2013, Rezende et al., 2014] of the ELBO is given by:

$$\nabla_\varphi \mathcal{L}(\theta, \varphi; x) = \mathbb{E}_{p_\varepsilon}\left[\nabla_z \log w_{\theta,\varphi}(x, z) \nabla_\varphi g(\varepsilon, \varphi)\right] - \mathbb{E}_{p_\varepsilon}\left[\nabla_\varphi \log q_\varphi(g(\varepsilon, \varphi)|x)\right].$$

The gradient estimator with respect to φ of the ELBO can be estimated using samples from the dataset. Let $(X_1, \ldots, X_B)_{1 \leqslant i \leqslant B}$ be i.i.d. with distribution π_{data}. Then, an estimator of $\nabla_\varphi \mathcal{L}(\theta, \varphi)$ can be computed as follows:

$$\begin{aligned} &S_\varphi(\theta, \varphi; \{X_i\}_{i=1}^B) \\ &\quad = -\frac{1}{B}\sum_{i=1}^{B} \frac{1}{M}\sum_{j=1}^{M} \left\{ \nabla_z \log \frac{p_\theta(X_i, g(\varepsilon_{i,j}, \varphi))}{q_\varphi(g(\varepsilon_{i,j}, \varphi)|x_i)} \nabla_\varphi g(\varepsilon_{i,j}, \varphi) - \nabla_\varphi \log q_\varphi(g(\varepsilon_{i,j}, \varphi)|X_i) \right\}, \end{aligned} \tag{4.2}$$

where, for all $1 \leq i \leq B$, $(\varepsilon_{i,1}, \ldots, \varepsilon_{i,M})$ are independent samples from p_ε.

Example 4.1. In a context of deep Gaussian models, $q_\varphi(\cdot|x)$ is a Gaussian probability density function with mean $\mu_\varphi(x) \in \mathbb{R}^d$ and variance $\mathrm{diag}(\sigma_\varphi^2(x)) \in \mathbb{R}^{d\times d}$ where $(\mu_\varphi(x), \sigma_\varphi^2(x))$ is the output of a neural network with input x. Then, if $z = \mu_\varphi(x) + \mathrm{diag}(\sigma_\varphi(x))\varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \mathrm{I}_d)$, $z \sim q_\varphi(\cdot|x)$. Writing $g(\varepsilon, \varphi) = \mu_\varphi(x) + \sigma_\varphi(x)\varepsilon$, the Jacobian of g with respect to ε is

$$J_\varepsilon[g](\varepsilon, \varphi) = \mathrm{diag}(\sigma_\varphi(x)).$$

Therefore, by standard change of variables, with $z = g(\varepsilon, \varphi)$,

$$\log q_\varphi(g(\varepsilon, \varphi)|x) = \log p_\varepsilon(\varepsilon) - \sum_{i=1}^{d} \log \sigma_{\varphi,i}(x).$$

4.5.2 Score Function Gradient

Alternatively, the score function gradient, also known as the Reinforce gradient [Glynn, 1990, Williams, 1992, Paisley et al., 2012], can be used. Unlike the reparameterization trick, this method does not necessitate reparameterization and is applicable to a wider range of variational distributions.

Proposition 4.2. *For all $\theta \in \Theta$, $\varphi \in \Phi$, we have:*

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

Proof. For all $x \in \mathsf{X}$, the score function gradient of the ELBO with respect to φ is given by:

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

Using the fact that $\mathbb{E}_{q_\varphi(\cdot|x)}\left[\nabla_\varphi \log q_\varphi(Z|x)\right] = 0$ under regularity conditions on $q_\varphi(z|x)$ yields

$$\begin{aligned} \nabla_\varphi \mathcal{L}(\theta, \phi; x) &= \mathbb{E}_{q_\varphi(\cdot|x)}\left[\nabla_\varphi \log q_\varphi(Z|x)(\log p_\theta(x, Z) - \log q_\varphi(Z|x))\right] \\ &= \mathbb{E}_{q_\varphi(\cdot|x)}\left[\log \frac{p_\theta(x, Z)}{q_\varphi(Z|x)} \nabla_\varphi \log q_\varphi(Z|x)\right]. \end{aligned}$$

□

The gradient estimator with respect to φ of the ELBO can be estimated using samples from the dataset. Let $(X_1, \ldots, X_B)_{1 \leqslant i \leqslant B}$ be i.i.d. with distribution π_{data}. Then, an estimator of $\nabla_\varphi \mathcal{L}(\theta, \varphi)$ can be computed as follows:

$$S_\varphi(\theta, \varphi; \{X_i\}_{i=1}^B) = -\frac{1}{B} \sum_{i=1}^B \frac{1}{M} \sum_{j=1}^M \log \frac{p_\theta(X_i, Z_{i,j})}{q_\varphi(Z_{i,j}|X_i)} \nabla_\varphi \log \frac{p_\theta(X_i, Z_{i,j})}{q_\varphi(Z_{i,j}|X_i)}, \tag{4.3}$$

where, for all $1 \leq i \leq B$, $(Z_{i,1}, \ldots, Z_{i,M})$ are independent samples from $q_\varphi(\cdot|X_i)$.

The pathwise gradient estimator often yields lower-variance estimates than the score function estimator [Miller et al., 2017, Buchholz et al., 2018], but its variance can sometimes exceed that of the score function estimator, especially when the score function correlates with other components of the pathwise estimator. Several methods have been proposed to further reduce variance, such as the Rao-Blackwellization estimator [Ranganath et al., 2014], Control Variates [Liévin et al., 2020], Stop Gradient estimator [Roeder et al., 2017], Quasi-Monte Carlo VAE [Buchholz et al., 2018], and Multi-Level Monte Carlo estimator [Fujisawa and Sato, 2021, He et al., 2022]. While our analysis focuses on the convergence rate of score function and pathwise gradient estimators, our convergence results also apply to most of these other methods.

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 $(X_1, \ldots, X_n)$ 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 $T_j : \mathsf{X} \to \mathbb{R}$, $1 \leq j \leq d$, such that $\mathbb{E}_\theta[|T_j(X_1)|] < \infty$. Then, write for all $1 \leq j \leq d$, $\theta \in \Theta$,

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

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

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

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

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

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

Example A.1. Let $(X_1, \ldots, X_n)$ 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 $\{T_j\}_{1\le j\le d}$ 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, $1 \le j \le d$, such that for all $\theta_* \in \Theta$, $1 \le j \le d$, $\mathbb{E}_{\theta_*}[|\psi_j(\theta_*, X_1)|] < \infty$. 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 $(X_i)_{1\le i\le n}$ be i.i.d. random variables with probability density f_{θ_*}, $\theta_* \in \Theta$. The likelihood of $(X_i)_{1\le i\le n}$ 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 $(X_i)_{1\le i\le n}$ 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:

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

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*

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

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_*},

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

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.

Buchholz et al., 2018. Buchholz, A., Wenzel, F., and Mandt, S. (2018). Quasi-Monte Carlo variational inference. In *International Conference on Machine Learning*, pages 668–677. PMLR.

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.

Fort and Moulines, 2003. Fort, G. and Moulines, E. (2003). Convergence of the monte carlo expectation maximization for curved exponential families. *The Annals of Statistics*, 31(4):1220–1259.

Fujisawa and Sato, 2021. Fujisawa, M. and Sato, I. (2021). Multilevel Monte Carlo variational inference. *Journal of Machine Learning Research*, 22(278):1–44.

Glynn, 1990. Glynn, P. W. (1990). Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM*, 33(10):75–84.

Gordon et al., 1993. Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). Novel approach to nonlinear and non-gaussian bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, 140(2):107–113.

Handschin and Mayne, 1969. Handschin, J. E. and Mayne, D. Q. (1969). Monte carlo techniques to estimate the conditional expectation in multi-stage non-linear filtering. *International Journal of Control*, 9(5):547–559.

He et al., 2022. He, Z., Xu, Z., and Wang, X. (2022). Unbiased mlmc-based variational bayes for likelihood-free inference. *SIAM Journal on Scientific Computing*, 44(4):A1884–A1910.

Kingma and Welling, 2013. Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Liévin et al., 2020. Liévin, V., Dittadi, A., Christensen, A., and Winther, O. (2020). Optimal variance control of the score-function gradient estimator for importance-weighted bounds. In *Advances in Neural Information Processing Systems*, volume 33, pages 16591–16602.

Miller et al., 2017. Miller, A., Foti, N., D'Amour, A., and Adams, R. P. (2017). Reducing reparameterization gradient variance. In *Advances in Neural Information Processing Systems*, volume 30.

Paisley et al., 2012. Paisley, J., Blei, D., and Jordan, M. (2012). Variational bayesian inference with stochastic search. In *International Conference on Machine Learning*, pages 1367–1374. PMLR.

Ranganath et al., 2014. Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *International Conference on Artificial Intelligence and Statistics*, pages 814–822. PMLR.

Rezende et al., 2014. Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286. PMLR.

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.

Roeder et al., 2017. Roeder, G., Wu, Y., and Duvenaud, D. K. (2017). Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In *Advances in Neural Information Processing Systems*, volume 30.

Williams, 1992. Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256.

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.