Sylvain Le Corff

Introduction to computational statistics

Chapter 1

M-estimation Z-estimation, maximum likelihood

1.1 Method of moments

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

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

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

Example 1.1. Let $(X_1, \dots, X_n)$ be i.i.d. random variables with exponential distribution with parameter $\theta > 0$. Using $T_1 : x \mapsto x$ and $T_2 : x \mapsto x^2$ we have for all $\theta > 0$,

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

The moment estimator associated with T_1 is

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

The moment estimator associated with T_2 is

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

1.2 Z-estimation

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

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

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

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

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

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

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

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

Example 1.2. 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 $\widehat{\theta}_n = n^{-1} \sum_{i=1}^{n} X_i$.

1.3 Maximum likelihood

Definition 1.3. Let $(\mathsf{X}, \mathcal{X})$ be a measurable space equipped with a sigma-finite measure μ. Let $(f_\theta)_{\theta \in \Theta}$ be a family of probability densities with respect to μ and $(X_i)_{1 \leq i \leq n}$ be i.i.d. random variables with probability density f_{θ_*}, $\theta_* \in \Theta$. The likelihood of $(X_i)_{1 \leq i \leq n}$ is the function

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

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

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

Example 1.4. Let $(X_i)_{1 \leq i \leq n}$ be i.i.d. Bernoulli random variables with parameter $\theta_* \in (0, 1)$. For all $\theta \in (0, 1)$,

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

and

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

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

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

1.4 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 1.5. For all $1 \leq k \leq n$, let $x_k \in \mathbb{R}^d$ and consider $(\xi_k)_{1\leq k\leq n}$ i.i.d. random variables with distribution $\mathcal{N}(0,1)$ and the linear regression model:

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

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

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

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

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

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

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

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

1.5 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 1.6. *Consider the following assumptions.*

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

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

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

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

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

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

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

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

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

which concludes the proof. □

Remark 1.7. 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.

1.5.1 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 1.8.

- 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

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

Chapter 2

Expectation Maximization algorithm

2.1 Introduction

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

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

We aim at solving

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

When L is positive, the problem is often written:

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

Example 2.1. A very popular setting is when and $f_\theta : z \mapsto p_\theta(z, X)$ where p_θ is the joint probability density function of two random variables (Z, X). Assuming that the random variable X is observed and that Z is not observed, we consider the likelihood function

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

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

2.2 Algorithm

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

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

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

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

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

Lemma 2.2. *For all $\theta, \theta' \in \Theta$,*

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

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

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

As log is concave, by Jensen's inequality,

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

which concludes the proof. □

By Lemma 2.2, 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 1 and introduced in [?].

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

Algorithm 1: A generic EM algorithm

The most common setting in which the EM algorithm is used is the case of Example 2.1.

Example 2.3. In the setting of Example 2.1, we have

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

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

2.3 Example: mixture of Gaussian distributions

In this example, we assume that the joint distribution of (Z, X) belongs to a family of distributions parametrized by a vector θ with real components. For $k \in \{1, \ldots, M\}$, write $\pi_k = \mathbb{P}_\theta(Z = k)$. Assume that conditionally on the event $\{Z = k\}$, X has a Gaussian distribution with mean $\mu_k \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. The probability density of this conditional distribution

is written g_k^θ, where the parameter $\theta = (\{\pi_k\}_{1\leq k\leq K}, \{\mu_k\}_{1\leq k\leq K}, \Sigma)$ belongs to the set $\Theta = \mathbb{S}_K \times (\mathbb{R}^d)^K \times \mathbb{R}^{d\times d}$ with $\mathbb{S}_K = \{(\pi_1, \dots, \pi_K) \in [0,1]^K \ ; \ \sum_{k=1}^K \pi_k = 1\}$. For all $1 \leq k \leq K$, the explicit computation of $\mathbb{P}_\theta(Z = k|X)$ writes

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

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

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

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

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

The algorithm is initialized by choosing $\theta^{(0)}$ randomly. Then, for each iteration $p \geq 0$, the update is decomposed into two steps.

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

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

2. Update the parameter estimate by computing:

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

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

Chapter 3

Variatonal inference and autoencoders

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.

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,

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

where

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

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

$$\mathrm{ELBO}(q) \leq \log p(x)\,.$$

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

$$q_* \in \text{argmax}_{q\in\mathcal{D}} \text{ELBO}(q)\,.$$

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:

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

In this case, we can write, for all $q \in \mathcal{D}$, and all $j \in \{1,\dots,d\}$,

$$\begin{aligned}\text{ELBO}(q) &= \mathbb{E}_q\left[\log \frac{p(Z_1,\dots,Z_d,x)}{\prod_{j=1}^d q_j(Z_j)}\right], \\ &= \mathbb{E}_{q_j}\left[\mathbb{E}_{q_{-j}}\left[\log p(Z_j|Z_{-j},x)\right]\right] - \mathbb{E}_{q_j}\left[\log q_j(Z_j)\right] + \text{C}\end{aligned}$$

where C does not depend on q_j, $z_{-j} = (z_\ell)_{1\le\ell\le d;\ell\neq j}$ and $q_{-j}(z_{-j}) = \prod_{1\le\ell\le d;\ell\neq j} q_\ell(z_\ell)$.

Assume that we want to optimize $\text{ELBO}(q)$ on q_j only, the other densities $(q_\ell)_{\ell\neq j}$ being kept fixed. 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 desnity 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. Therefore,

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

Therefore, optimizing $\text{ELBO}(q)$ 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, $1 \le j \le d$ until a stopping criterion is met. In Algorithm 2, 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, mean-field family $\mathcal{D}$, initial variational distribution $\{q_j^{(0)}\}_{1\le j\le d}$, maximum number of iteration N
Result: A variational distribution for each coordinate of Z, $q_j^{(N)}$, $1 \le j \le d$.
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 $1 \le \ell \le d$, $\ell \neq j$ and $q_{-j}^{(k)} = \prod_{1\le\ell\le d,\ell\neq j} q_\ell^{(k)}$;
  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

Application to a mixture of Gaussian distributions

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 $c_i \in \{1 \leq i \leq K\}$ the group X_i belongs to, this variable is not observed. The weight of component k is written ω_k, i.e. conditionally on μ, $c = (c_i)_{1 \leq i \leq n}$ are independent with multinomial distribution with parameters $\{\omega_1, \dots, \omega_K\}$ (in particular, c is independent of μ).

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 $(\phi_i)_{1 \leqslant i \leqslant n}$: $q(c_i = k) = \phi_i(k)$ for $1 \leqslant k \leqslant K$.

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

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

where

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

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

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

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

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

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

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

Then,

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

and

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

The update writes:

$$\varphi_i(k) \propto p(c_i = k)\exp\left(m_k x_i - \frac{m_k^2 + s_k}{2}\right) .$$

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

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

The update writes therefore,

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

Variational Autoencoders

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

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

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

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

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

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

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

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

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

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

The empirical loss is then given by

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

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

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

The joint optimization of θ and φ is a complex problem both for practical and theoretical reasons and many research works have been devoted to this problem in the past few years. In most cases,

$\mathcal{L}_n(\theta, \varphi)$ cannot be computed explicitly since expectations under the variational distribution are not explicit. Therefore, $\mathcal{L}_n(\theta, \varphi)$ is replaced by a Monte Carlo estimate $\widehat{\mathcal{L}}_n(\theta, \varphi)$:

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

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

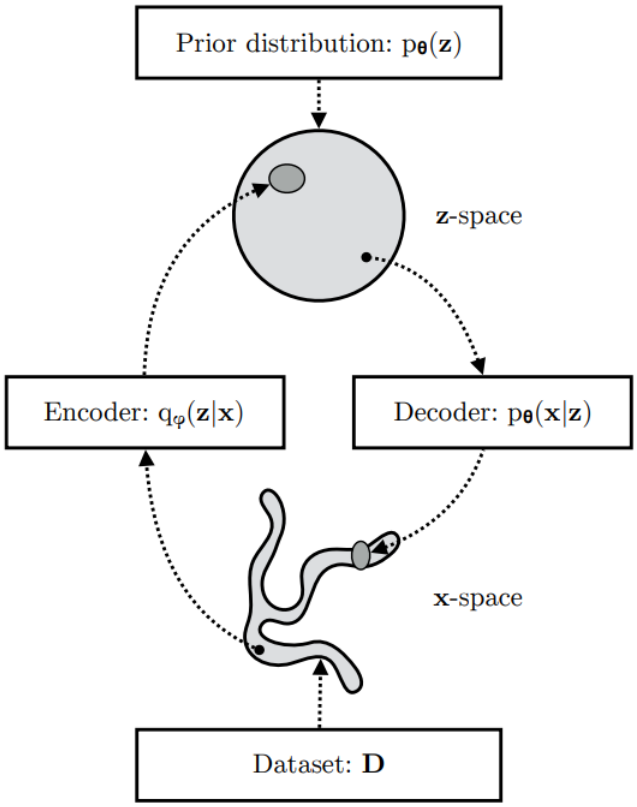

Fig. 3.1 An illustration of a VAE. From "An Introduction to Variational Autoencoders", Kingma et al., 2019.

Chapter 4

Markov chain Monte Carlo

4.1 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, the target distribution π can have several forms depending on the different contexts. The target distribution can be given for instance by a marginal density $p_\theta(y) = \int p_\theta(x, y)\lambda(\mathrm{d}x)$ or by a conditional/posterior distribution $p_\theta(x|y) = p_\theta(x)p_\theta(y|x)/\int p_\theta(x, y)\lambda(\mathrm{d}x)$ where $\theta \in \Theta$ and λ is a dominating measure.

Example 4.1. In Bayesian inference, prior belief is combined with data to obtain posterior distributions on which statistical inference is based. Except for some simple cases, Bayesian inference can be computationally intensive and may rely on computational techniques.

The basic idea in Bayesian analysis is that a parameter vector $\theta \in \Theta$ is unknown, so it is endowed with a *prior* distribution with probability density $\pi_0(\theta)$. We also introduce a model or likelihood, $\mathrm{p}(y_{1:n} \mid \theta)$, that is a probability density function for the data $y_{1:n}$ which depends on the parameter vector. Inference about θ is then based on the *posterior* distribution, which is obtained via Bayes's theorem,

$$\theta \mapsto \pi(\theta) = \frac{\pi_0(\theta)\,\mathrm{p}(y_{1:n} \mid \theta)}{\int \pi_0(\theta)\,\mathrm{p}(y_{1:n} \mid \theta)\lambda(\mathrm{d}\theta)} \propto \pi_0(\theta)\,\mathrm{p}(y_{1:n} \mid \theta).$$

In some simple cases, the prior and the likelihood are *conjugate* distributions that may be combined easily. For example, in n fixed repeated (i.i.d.) Bernoulli experiments with probability of success θ, a *Beta-Binomial* conjugate pair is taken. In this case the prior is $\mathrm{Beta}(a, b)$: $\pi_0(\theta) \propto \theta^a(1-\theta)^b \mathbb{1}_{(0,1)}(\theta)$; the values $a, b > -1$ are called hyperparameters. The likelihood in this example is $\mathrm{Binomial}(n, \theta)$: $\mathrm{p}(y \mid \theta) \propto \theta^y(1-\theta)^{n-y}$, from which we easily deduce that the posterior is also Beta, $\pi(\theta) \propto \theta^{y+a}(1-\theta)^{n+b-y}\mathbb{1}_{(0,1)}(\theta)$, and from which inference may easily be achieved.

In more complex experiments, the posterior distribution is often difficult to obtain by direct calculation, so alternatives have to be deployed to obtain samples approximately distributed as the posterior distribution.

Example 4.2. 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, is it 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 procedure 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).

4.2 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 $k \leq \ell$, $u_{k:\ell}$ means $(u_k, \ldots, u_\ell)$ and $u_{k:\infty}$ means $(u_{k+\ell})_{\ell \in \mathbb{N}}$.

Definition 4.3. 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 convention and notation.

- 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 $P^0 = I$ where I is the identity kernel: $(x, A) \mapsto \mathbf{1}_A(x)$, and set for $k \geq 0$, $P^{k+1} = P^k P$.

Definition 4.4. 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.

1. 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.
2. $\mathbb{P}(X_0 \in A) = \nu(A)$.

Note that in the 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 4.5. 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 applying P^k on both sides of $\pi P = \pi$ shows that $\pi P^{k+1} = \pi P^k$ 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 $(X_{k_1}, \ldots, X_{k_n})$ follows the same distribution as $(X_{k_1+p}, \ldots, X_{k_n+p})$.

Definition 4.6. 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 (with infinitesimal notation)

$$\pi(\mathrm{d}x)P(x, \mathrm{d}y) = \pi(\mathrm{d}y)P(y, \mathrm{d}x), \tag{4.1}$$

that is, for all measurable bounded or non-negative functions h on $\left(\mathsf{X}^2, \mathcal{X}^{\otimes 2}\right)$,

$$\iint_{\mathsf{X}^2} h(x, y)\pi(\mathrm{d}x)P(x, \mathrm{d}y) = \iint_{\mathsf{X}^2} h(x, y)\pi(\mathrm{d}y)P(y, \mathrm{d}x). \tag{4.2}$$

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).

Proposition 4.7. *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.

4.3 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. The reason for that is that an invariant probability measure will be a good candidate for the "limiting" distribution of $\{X_k : k \in \mathbb{N}\}$ (in some sense to be defined) and this in turn, will allow us to provide an approximation of $\pi(h) = \int_{\mathsf{X}} h(x)\pi(\mathrm{d}x)$ of the form $n^{-1}\sum_{k=0}^{n-1} h(X_i)$ for any measurable function h.

For simplicity we now assume that π has a density with respect to some dominating σ-finite measure λ and by abuse of notation, we also denote by π this density, that is we write $\pi(\mathrm{d}x) = \pi(x)\lambda(\mathrm{d}x)$ and we assume that this density π is positive.

Moreover, let Q be Markov kernel on $\mathsf{X} \times \mathcal{X}$ such that $Q(x, \mathrm{d}y) = q(x, y)\lambda(\mathrm{d}y)$, that is, for any $x \in \mathsf{X}$, $Q(x, \cdot)$ is also dominated by λ and denoting by $q(x, \cdot)$ this density, we assume for simplicity that $q(x, y)$ is positive for all $x, y \in \mathsf{X}$.

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

Input : n
Output: $X_0, \ldots, X_n$

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

end

Algorithm 2: 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 function α.

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

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

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

$$P^{MH}_{\langle\pi,Q\rangle}(x, \mathrm{d}y) = Q(x, \mathrm{d}y)\alpha(x, y) + \bar{\alpha}(x)\delta_x(\mathrm{d}y). \tag{4.3}$$

Lemma 4.8. *The Markov kernel $P^{MH}_{\langle\pi,Q\rangle}$ is π-reversible if and only if*

$$\pi(\mathrm{d}x)Q(x, \mathrm{d}y)\alpha(x, y) = \pi(\mathrm{d}y)Q(y, \mathrm{d}x)\alpha(y, x). \tag{4.4}$$

Equation (4.4) is often called the detailed balance condition.

Proof. First, note that

$$\pi(\mathrm{d}x)\bar{\alpha}(x)\delta_x(\mathrm{d}y) = \pi(\mathrm{d}y)\bar{\alpha}(y)\delta_y(\mathrm{d}x). \tag{4.5}$$

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 (4.3) with (4.5), we obtain that $P^{MH}_{\langle\pi,Q\rangle}$ is π-reversible if and only if the detailed balance condition (4.4) is satisfied. This completes the proof. □

We now provide an explicit expression of the acceptance probability α. The proof of Lemma 4.9 is straightforward.

Lemma 4.9. *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 (4.4).

Example 4.10 (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 $\eta_k \sim q(\cdot)$.

Gibbs sampler

The Gibbs sampler is a specific version of the Metropolis-Hastings algorithm in cases where the state $\mathsf{X} = \mathbb{R}^d$ and where for all $x \in \mathbb{R}^d$, x can be decomposed into $x = (x^{(1)}, \ldots, x^{(d)})$ so that for all $1 \le j \le d$, we know how to sample from $\pi(\cdot|x^{(-j)})$ where $x^{(-j)} = (x^{(\ell)})_{\ell \neq j}$. It is easy to write the proposal distribution associated with the Gibbs sampler and to show that the acceptance rate is 1.

Input : $X_k = (X_k^{(1)}, \ldots, X_k^{(q)})$
Output: X_{k+1}

Sample uniformly J in $\{1, \ldots, d\}$.
Sample $X_{k+1}^{(J)} \sim \pi(\cdot|X_k^{(-J)})$.
For all $1 \le j \le q$, $j \neq J$, set $X_{k+1}^{(-j)} = X_k^{(-j)}$.

Algorithm 3: One iteration of the (Random Scan) Gibbs sampler

4.4 Some alternatives to MCMC

In computational statistics, when it comes to approaching a target law in a very large space, classical techniques using Markov chains admitting "exactly" this target law for invariant distribution may suffer from a slow exploration of the state space. In a high dimensional framework, the candidate is often proposed in an uninformative region and it is likely that it is rejected. Some other approximation techniques do not even try to construct random variables with distribution close to π. We briefly discuss here approximation techniques.

4.4.1 Sequential Monte Carlo methods

In this section we briefly explain basic ideas of Sequential Monte Carlo methods. The rough idea of sequential Monte Carlo methods which target π is to find intermediate target distributions $\pi_1, \ldots, \pi_T = \pi$ and to construct, sequentially, Monte Carlo approximations of each π_k, $1 \leq k \leq T$.

This core ingredient of SMC methods is importance sampling. If π and g are densities with respect to the same dominating measure, and assuming that $g(x) = 0$ implies $\pi(x) = 0$, then, for any measurable function h, we can approximate $\pi(h)$ with $N^{-1}\sum_{k=1}^{N}\omega_k h(X_k)$ where $(X_k)_{k\in[1:N]} \overset{\text{i.i.d}}{\sim} g$ and $\omega_k = \pi(X_k)/g(X_k)$. Since π is typically known only up to a multiplicative factor, the quantity $N^{-1}\sum_{k=1}^{N}\omega_k h(X_k)$ is not explicit due to this multiplicative factor and we typically choose instead the auto-normalized estimator:

$$\pi_N(h) = \frac{N^{-1}\sum_{k=1}^{N}\omega_k h(X_k)}{N^{-1}\sum_{\ell=1}^{n}\omega_\ell} = \sum_{k=1}^{N}\left(\frac{\omega_k}{\sum_{\ell=1}^{N}\omega_\ell}\right)h(X_k) = \sum_{k=1}^{N}\bar{\omega}_k h(X_k)$$

where $\bar{\omega}_k = \omega_k/(\sum_{\ell=1}^{k}\omega_\ell)$. Now the right-hand-side can be calculated even if π is known only up to a multiplicative factor since $\bar{\omega}_k$ is a ratio where π is involved (both in the numerator and the denominator).

Thus, $\pi(h)$ is approximated using a population of "particles" $\{(X_k, \omega_k)\}_{k\in[1:N]}$ (we mean by particle a "support" point X_k and an associated weight ω_k). Note that a weight is usually unnormalized but when considering the approximation, we use the normalized weights: $\omega_k/(\sum_{\ell=1}^{k}\omega_\ell)$.

Of course, if all the X_k were iid from π directly, all the associated weights would be equal. So here, by allowing different weights, we are more flexible. Still, if the weights are too different, this is not satisfactory because only a few particles contain all the information. In that case, we prefer to resample inside the population.

4.4.1.1 Resampling step

Assume that $\{(X_k, \omega_k)\}_{k\in[1:N]}$ targets π_0. Define $\bar{\omega}_k = \omega_k/(\sum_{\ell=1}^{k}\omega_\ell)$. An example of resampling step is defined as follows. For $k \in [1:n]$, set independently $\tilde{X}_k = X_j$ with probability $\bar{\omega}_j$. Then $\{(\tilde{X}_k, 1)\}_{k\in[1:N]}$ still targets π_0. The target distribution is not changed but now, all the weights are equal. Informative particles (i.e. with high weights) are likely to be replicated after resampling while noninformative are likely to disappear (because they were not chosen). Support points are changed but still within the initial pool of support points.

4.4.1.2 Exploration step

Assume that $\{(X_k, \omega_k)\}_{k\in[1:N]}$ targets π_0 and that

$$\pi_1(y) = \int_{\mathsf{X}} \pi_0(\mathrm{d}x) q(x, y), \tag{4.6}$$

where q is the density of a Markov kernel. An example of exploration step is defined as follows. For $k \in [1:n]$, draw independently $\tilde{X}_k \sim r(X_k, \cdot)$, where $r(X_k, \cdot)$ is a density that can be easily simulated. Then, $\left\{(\tilde{X}_k, \omega_k q(X_k, \tilde{X}_k)/r(X_k, \tilde{X}_k)\right\}_{k\in[1:N]}$ targets π_1. Here, support points are moved and weights are updated by a multiplicative factor (except when $r = q$, in which case, support points are moved but the associated weights do not need to be updated since $q(X_k, \tilde{X}_k)/r(X_k, \tilde{X}_k) = 1$).

4.4.1.3 Reweighting step

Assume that $\{(X_k, \omega_k)\}_{k\in[1:N]}$ targets π_0 and that

$$\pi_1(x) = \frac{\pi_0(x) g(x)}{\int_{\mathsf{X}} \pi_0(\mathrm{d}u) g(u)}, \tag{4.7}$$

where g is a nonnegative function. Then, $\{(X_k, \omega_k g(X_k)\}_{k\in[1:N]}$ targets $\pi_1(h)$. Here, support points are unchanged but weights are updated.

Finally, when choosing the intermediate target distributions $\pi_1 \to \pi_2 \to, \ldots, \to \pi_T = \pi$, we must check that each step $\pi_i \to \pi_{i+1}$ corresponds either to (4.6) or (4.7) so that we can let evolve a population of particles through exploration, and reweighting steps. Resampling can always be performed when weights are too different and this is often measured when the Effective Sample Size (which is a real number between 1 and N), $\widehat{ESS} = 1/\sum_{k=1}^{N} (\bar{\omega}_k)^2$, falls below a certain arbitrary threshold.

4.5 Approximate Bayesian computation

Approximate Bayesian computation (ABC) is an alternative approach when computation of the posterior is challenging, either because the size of the data or the complexity of realistic models makes the calculation computationally intractable. In this setting, the parameter θ is endowed with a prior distribution π_0 and, the conditional density of the observation Y given θ is $y \mapsto p(y|\theta)$. ABC specifically provides a solution when the likelihood $y \mapsto p(y|\theta)$ cannot be evaluated. A generic description of the original ABC algorithm requires (i) the introduction of statistics $S(y) \in \mathbb{R}^m$ where m is usually sensibly smaller than the dimension of y and (ii) a distance d on $\mathbb{R}^m \times \mathbb{R}^m$. Note that if no statistics can be widely defined S can be the identity function. Then, the most direct ABC alforithm is described in Algorithm 5.

Data: Observation Y, threshold ε, N
Result: Samples approximately distributed according to $\pi(\theta) = p(\theta|Y)$.
for $i = 1 \to N$ **do**
  draw θ_i with prior distribution π_0;
  draw Y_i with distribution $p(\cdot|\theta_i)$;
end
Return all θ_i such that $\mathsf{d}(S(Y_i), S(Y)) < \varepsilon$;

Algorithm 4: ABC algorithm.

When S is the identity function, the random variables sampled by this algorithm have distribution $\pi_\varepsilon(\cdot|Y)$ where

$$\pi_\varepsilon(\theta|Y) \propto \int p(y|\theta)\pi_0(\theta)\mathbb{1}_{A_{\varepsilon,Y}}(y)\mathrm{d}y,$$

with $A_{\varepsilon,Y} = \{y; \mathsf{d}(y, Y) < \varepsilon\}$. The intuitive idea behind this algorithm is that if $\varepsilon \to 0$, $\pi_\varepsilon(\theta|Y) \to \pi(\theta|Y)$ and if $\varepsilon \to \infty$, $\pi_\varepsilon(\theta|Y) \to \pi_0(\theta)$. This initial version of the ABC approach raises many practical issues among which an appropriate calibration of ε, the choice of statistics S, and the widespread inefficiency of sampling candidates according to the prior distribution π. In practice, the threshold ε is usually determined as a quantile of the observed distance $(\mathsf{d}(S(Y_i), S(Y)))_{1 \leqslant i \leqslant N}$ which allows to introduce Algorithm 6.

Data: Observation Y, N, integer M_N
Result: Samples approximately distributed according to $\pi(\theta) = p(\theta|Y)$.
for $i = 1 \to N$ **do**
  draw θ_i with prior distribution π_0;
  draw Y_i with distribution $p(\cdot|\theta_i)$;
end
Return all θ_i such that $S(Y_i)$ is in the set of M_N nearest neighbors of $S(Y)$ with respect to distance d;

Algorithm 5: ABC algorithm with calibrated threshold.

These two algorithms generate independent samples but do not build upon the accepted samples to propose new candidates in a more efficient way than using the prior distribution. This can be performed by considering ABC within a MCMC algorithm.