

Sylvain Le Corff

Sampling for generative models
Theory and practical applications

Notations and useful definitions

The following notations are used throughout these lecture notes. For $r, s \in \mathbb{N}$ such that $r \leq s$, we write $[r : s] = \{r, r+1, \dots, s\}$. For all real-valued sequences $\{a_k\}_{k \geq 0}$, we write $\liminf_n a_n = \lim_{n \rightarrow \infty} (\inf_{k \geq n} a_k)$ and similarly, $\limsup_n a_n = \lim_{n \rightarrow \infty} (\sup_{k \geq n} a_k)$. Moreover, $\lim_n a_n$ exists if and only if $\liminf_n a_n = \limsup_n a_n$. For all $a \in \mathbb{R}$, $a^+ = \max(a, 0)$ and $a^- = \max(-a, 0) = -\min(a, 0)$ and we have $|a| = a^+ + a^-$ and $a = a^+ - a^-$. For all set A , ∂A is the frontier of A , i.e. the set of points x such that in all neighborhoods of x , there are infinitely many points in A and infinitely many points in A^c . For all sequences of real-valued random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, $(X_n)_{n \geq 0}$ converges in distribution to X if one of the following equivalent statements hold.

- (a) For all bounded continuous functions h , $\lim_n \mathbb{E}[h(X_n)] = \mathbb{E}[h(X)]$.
- (b) For all $A \in \mathcal{B}(\mathbb{R})$ such that $\mathbb{P}(X \in \partial A) = 0$, $\lim_n \mathbb{P}(X_n \in A) = \mathbb{P}(X \in A)$.
- (c) For all $x \in \mathbb{R}$ such that $\mathbb{P}(X = x) = 0$, $\lim_n \mathbb{P}(X_n \leq x) = \mathbb{P}(X \leq x)$.
- (d) For all $u \in \mathbb{R}$, $\lim_n \mathbb{E}[e^{iuX_n}] = \mathbb{E}[e^{iuX}]$.

When there is no confusion, we say that X_n converges weakly to X . We say that $(X_n)_{n \geq 0}$ converges in probability to X if for all $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$

We say that $(X_n)_{n \geq 0}$ converges almost surely to X if

$$\mathbb{P}(\lim_{n \rightarrow \infty} X_n = X) = 1.$$

Note that almost sure convergence implies convergence in probability. The following properties are used many times in these lecture notes.

- (i) If $(X_n)_{n \geq 0}$ converges in distribution to X , then for all continuous functions f , $\{f(X_n)\}_{n \geq 0}$ converges in distribution to $f(X)$. Note that this property holds, when f is continuous and not necessarily bounded.
- (ii) By the Slutsky lemma, if $(X_n)_{n \geq 0}$ converges in probability to c where c is a constant and if $(Y_n)_{n \geq 0}$ converges in distribution to Y , then $\{(X_n, Y_n)\}_{n \geq 0}$ converges in distribution to (c, Y) . In particular, for all continuous functions f , $\{f(X_n, Y_n)\}_{n \geq 0}$ converges in distribution to $f(c, Y)$.
- (iii) $X \sim \mathcal{N}(0, 1)$ if and only if for all $u \geq 0$, $\mathbb{E}[e^{iuX}] = e^{-u^2/2}$. Moreover, $X \sim \mathcal{N}(\mu, \sigma^2)$ if and only if for all $u \geq 0$, $\mathbb{E}[e^{iuX}] = e^{-u^2\mathbb{V}[X]/2 + iu\mathbb{E}[X]}$ and in that case, $\sigma^2 = \mathbb{V}[X]$ and $\mu = \mathbb{E}[X]$.
- (iv) If $c \in \mathbb{R}$ is a constant, then $(X_n)_{n \geq 0}$ converges in probability to c if and only if $(X_n)_{n \geq 0}$ converges in distribution to c .

Contents

1	Target distributions and examples	5
1.1	Target distributions	6
1.1.1	Bayesian inference	6
1.1.2	Energy based models	7
1.1.3	Models with Latent variables	8
1.2	Some approximation methods	9
1.2.1	Variational inference	9
1.2.2	Variational auto-encoders	10
1.2.3	Sequential Monte Carlo methods	11
1.3	Approximate Bayesian computation	12
2	Variational autoencoders	15
2.1	Introduction	15
2.2	Gradient-based optimization	16
2.2.1	Pathwise Gradient	17
2.2.2	Score Function Gradient	18
2.3	Marginal likelihood and sampling approaches	19
2.4	Empirical Excess Risk	20
2.5	Classical extensions of VAEs	21
2.5.1	IWAEs	21
2.5.2	β -VAEs	22
3	Score-based diffusion models	23
3.1	Motivations of score-based approaches	23
3.2	Score-based generative models	24
3.3	Some theoretical guarantees	28
4	Basics in Markov chains	29
4.1	Main notation	29
4.2	Definitions	29
4.2.1	Additional notation	31
4.3	Canonical space	32
4.3.1	A simpler problem	32
4.3.2	The general case	32
4.3.3	The Markov property	33
5	Metropolis-Hastings algorithms	35
5.1	Invariant probability measures: existence	35
5.1.1	Metropolis-Hastings (MH) algorithms	36
5.2	Invariant probability measure: uniqueness	38

5.2.1	Application to Metropolis-Hastings algorithms.	40
6	Ergodicity and Law of Large numbers	41
6.1	Dynamical systems.....	41
6.2	Markov chains and ergodicity	42
7	Geometric ergodicity and Central Limit theorems	47
7.1	Total variation norm and coupling	47
7.2	Geometric ergodicity	50
7.3	The Poisson Equation	53
7.3.1	Definition	53
7.3.2	Poisson equation and martingales.....	54
7.3.3	Central Limit theorems	55
8	Variants of MH algorithms	57
8.1	Generalisation of MH Algorithms	57
8.2	Pseudo marginal Monte Carlo methods	58
8.3	Hamiltonian Monte Carlo	59
8.3.1	MH with deterministic moves	60
8.3.2	Hamiltonian dynamics	61
8.3.3	The leapfrog integrator.....	63
8.4	Data augmentation	64
8.4.1	Two-stage Gibbs sampler	67
9	Illustrations, exercises, extensions	69
9.1	Illustrations	69
9.1.1	Illustrations of HMC.....	69
9.2	Exercises	70
	References	73
Index	75

Chapter 1

Target distributions and examples

Contents

1.1	Target distributions	6
1.1.1	Bayesian inference	6
1.1.2	Energy based models	7
1.1.3	Models with Latent variables	8
1.2	Some approximation methods	9
1.2.1	Variational inference	9
1.2.2	Variational auto-encoders	10
1.2.3	Sequential Monte Carlo methods	11
1.3	Approximate Bayesian computation	12

In a general framework, generative modeling aims at designing algorithms to obtain samples from a complex distribution π defined on a measurable space $(\mathbf{X}, \mathcal{X})$. Such algorithms can be applied in many situations, we describe in this chapter some examples from various applications. The target distribution π may have several forms depending on the different contexts. Very common situations in which sampling from a complex distribution π include the following settings.

- The distribution π can be known up to a multiplicative constant (which is common in a Bayesian setting). Obtaining samples from π is far from straightforward in such cases, in particular in high dimensional problems.
- In other settings, we have access to a dataset $\mathcal{D} = \{X_1, \dots, X_n\}$ of i.i.d. data with distribution π and we introduce parametric models for the probability density associated with π , $\{p_\theta\}_{\theta \in \Theta}$, where Θ is a set of parameters. For all $\theta \in \Theta$, p_θ may depend on unobserved random variables and estimating θ using \mathcal{D} can be challenging (see for instance variational autoencoders).
- Designing simple parametric models $\{p_\theta\}_{\theta \in \Theta}$ may be too restrictive for complex random variables (images, videos, etc.). In these settings, we propose generative models to sample approximately from π without directly introducing such parametric models (see for instance score-based diffusion models).

For instance, in [Ng and Jordan, 2001], the authors discuss generative classifiers and propose to learn a model of a joint probability function of the inputs X and the outputs Y so that predictions can be made by using Bayes rule to compute the distribution of Y given X . Without being restricted to classification problems, we focus in these notes on generative models where the data distribution depends on other variables or parameters, and we propose to model the joint distribution of all random variables. In such a setting, we may assume that the random variables (X, Y) have an unknown distribution and we usually propose a parametric family of joint probability distributions $(x, y) \mapsto p_\theta(x, y)$, $\theta \in \Theta$, with respect to a dominating measure λ , in order to approximate this unknown distribution.

In machine/statistical learning most of the problems we aim to solve require then to sample random variables from a complex distribution. Depending on the context, this distribution can

be given for instance by a marginal density $p_\theta(y) = \int p_\theta(x, y) \lambda(dx)$ or by a conditional/posterior distribution $p_\theta(x|y) = p_\theta(x)p_\theta(y|x) / \int p_\theta(x, y) \lambda(dx)$.

1.1 Target distributions

We first introduce several contexts where target distributions appear naturally. In Bayesian inference, the target distribution is the posterior distribution of the parameter given all the data. In partially observed models, the target distribution is the distribution of the unobserved variables given the data. In these lecture notes, the target distribution we aim to sample from is written π (or π_{data} depending on the context), although it may depend on many parameters or input data as detailed in the examples of this chapter.

1.1.1 Bayesian inference

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 $\theta \mapsto \pi_0(\theta) \lambda(d\theta)$. We also introduce a model or likelihood, $p(y_{1:n} | \theta)$, that is a conditional 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,

$$\theta \mapsto \pi(\theta) = \frac{\pi_0(\theta) p(y_{1:n} | \theta)}{\int \pi_0(\theta) p(y_{1:n} | \theta) \lambda(d\theta)}. \quad (1.1)$$

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 $\theta \in (0, 1)$, a *Beta-Binomial* conjugate pair is taken. In this case the prior is $\text{Beta}(a, b)$: $\pi_0(\theta) \propto \theta^a (1 - \theta)^b$; the values $a, b > -1$ are called hyperparameters. The likelihood in this example is $\text{Binomial}(n, \theta)$: $p(y | \theta) \propto \theta^y (1 - \theta)^{n-y}$, from which we easily deduce that the posterior is also Beta, $\pi(\theta) \propto \theta^{y+a} (1 - \theta)^{n+b-y}$, 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. Note that in (1.1), the numerator is usually known and explicit while, due to the integral, the denominator is a multiplicative unknown constant.

Example 1.1. In Bayesian deep learning, uncertainty quantification may be obtained by analyzing the posterior distribution of the weights, as in [Blundell et al., 2015]. In this setting, using an input $x \in \mathbb{R}^p$, the observations Y , conditionally on the parameter θ , has a Gaussian distribution with mean $m_\theta(x)$ and covariance $\Sigma_\theta(x)$. The quantities $m_\theta(x)$ and $\Sigma_\theta(x)$ are outputs of a neural network (for instance a Multi-layer Perceptron) with input x and parameters θ . The prior distribution of θ is Gaussian with independent entries with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. Since $m_\theta(x)$ and $\Sigma_\theta(x)$ contain many nonlinearities the posterior distribution π of θ , i.e. the distribution of θ given Y and the input x , is unknown.

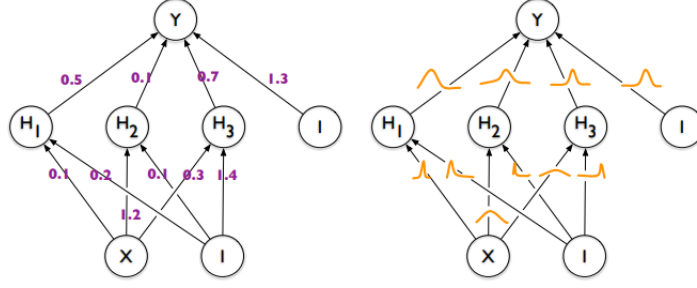


Fig. 1.1 Left: each weight (parameter) has a fixed value, as provided by classical backpropagation. Right: each weight is assigned a distribution, as provided by Bayes by Backprop. From [Blundell et al., 2015].

1.1.2 Energy based models

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 than 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. We can choose specific architectures tailored for every type of data, see [Song and Kingma, 2021] for a recent overview of EBM.

In a generic setting, the target random variable take 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:

$$x \mapsto \pi_\theta(x) \propto \exp(-E_\theta(x)) = \frac{\exp(-E_\theta(x))}{\int \exp(-E_\theta(u)) du},$$

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:

$$Z_\theta = \int \exp(-E_\theta(u)) du.$$

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 unknown 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:

$$x \mapsto \nabla_\theta \log \pi_\theta(x) = -\nabla_\theta E_\theta(x) - \nabla_\theta \log Z_\theta.$$

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

$$\begin{aligned} \nabla_\theta \log Z_\theta &= Z_\theta^{-1} \int \nabla_\theta \exp(-E_\theta(u)) du \\ &= \int \{-\nabla_\theta E_\theta(u)\} Z_\theta^{-1} \exp(-E_\theta(u)) du = \int \{-\nabla_\theta E_\theta(u)\} \pi_\theta(u) du. \end{aligned}$$

Therefore $\nabla_\theta \log Z_\theta = \mathbb{E}_{\pi_\theta}[-\nabla_\theta E_\theta(X)]$ 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).

1.1.3 Models with Latent variables

In some situations, observations are partial and unfortunately, do not contain some variables of interest. In such a case, given a generative process for the data, we might be interested in reconstructing the distribution of the missing variables given the data. This will be our target distribution.

Example 1.2 (Deep latent variable model). In [Kingma and Welling, 2013], the authors study a deep latent variable model for multivariate Bernoulli data. In this setting, $Y \in \{0, 1\}^D$ and conditionally on a variable $X \in \mathbb{R}^d$, (Y_1, \dots, Y_D) are independent with Bernoulli distribution with parameters $\mathbf{p}_\theta(X) = (p_{1,\theta}(X), \dots, p_{D,\theta}(X))$, where $\mathbf{p}_\theta(X)$ is the output of a Multi-layer Perceptron with input X and parameters θ (weights and biases). In this example, the input variable has a prior distribution $X \sim \mathcal{N}(0, I_d)$ and, for any value of θ , the conditional distribution of X given Y is not available explicitly.

Example 1.3 (Hidden Markov models). Such a model is defined by a bivariate Markov chain $(Z_k)_{k \in \mathbb{N}} = (X_k, Y_k)_{k \in \mathbb{N}}$ on $\mathbf{X} \times \mathbf{Y}$ where $(\mathbf{X}, \mathcal{X})$ and $(\mathbf{Y}, \mathcal{Y})$ are two measurable spaces, and where the transition is defined as follows.

$$\begin{aligned} \text{Conditionally on } Z_{0:k-1}, \quad X_k &\sim Q(X_{k-1}, \cdot), \\ \text{Conditionally on } (X_k, Z_{0:k-1}), \quad Y_k &\sim G(X_k, \cdot), \end{aligned}$$

with Q a Markov kernel on $\mathbf{X} \times \mathcal{X}$ and G a Markov kernel on $\mathbf{X} \times \mathcal{Y}$.

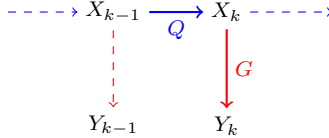


Fig. 1.2 A hidden Markov model.

In such a model, only $Y_{1:n}$ are observed and inference must be driven on the basis of $Y_{1:n}$ only. In a fully dominated model, we assume that $Q(x, dx') = q(x, x')\lambda(dx')$ and $G(x, dy) = g(x, y)\nu(dy)$ where λ and ν are σ -finite dominating measures on $(\mathbf{X}, \mathcal{X})$, and $(\mathbf{Y}, \mathcal{Y})$ respectively. In such a case, assuming that the initial distribution of X_0 is a Dirac mass at x_0 , we might be interested in the law of the missing variables $X_{1:n}$ given the observations $Y_{1:n}$, so that the target density is:

$$\pi(x_{1:n}) = \frac{\prod_{i=1}^n q(x_{i-1}, x_i) g(x_i, y_i)}{\int_{\mathbf{X}^n} \prod_{i=1}^n q(x_{i-1}, x_i) g(x_i, y_i) \lambda(dx_i)}.$$

Once again the denominator is an unknown multiplicative factor, whereas the numerator is explicit and we are therefore in a context where MCMC methods can be applied. Nevertheless, in this example, the number of hidden variables is n , which can be very large and the target distribution is thus associated with a very high dimensional space, \mathbf{X}^n , so that we have to be very careful when applying these methods.

1.2 Some approximation methods

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 refused, leading the Markov chain to remain stuck at the same place a certain amount of time.

Some other approximation techniques do not even try to construct random variables with distribution close to π . We briefly discuss in this chapter several approximation techniques: Variational Inference, Variational Auto-Encoders, Sequential Monte Carlo methods.

1.2.1 Variational inference

Variational Inference is a technique derived from the Machine Learning community whose principle is to approach the target density through various optimization techniques. The principle consists in giving up aiming exactly at the target law, but instead, we have at hand a sufficiently rich family of laws (that can be easily simulated) and the idea is then to select a member of this family close to the target in the sense of a certain divergence through optimization procedures. Variational inference has been successfully used in many applications and seems to be faster and more efficient to explore large spaces compared to classical Monte Carlo Markov chain methods. The proximity is often measured in terms of Kullback divergence or even more recently in terms of f -divergence. Although these techniques are extremely powerful and popular in practice, some statistical properties of these methods are still largely unknown.

Letting f_α be the convex function on $(0, +\infty)$ defined by $f_0(u) = -\log(u)$, $f_1(u) = u \log(u)$ and $f_\alpha(u) = [u^\alpha - 1]/(\alpha(\alpha - 1))$ for all $\alpha \in \mathbb{R} \setminus \{0, 1\}$, the α -divergence for all $\alpha \in \mathbb{R}$ is defined by

$$D_\alpha(\mathbb{P} \parallel \mathbb{Q}) = \int_{\mathbf{X}} f_\alpha \left(\frac{q(x)}{p(x)} \right) p(x) \lambda(dx), \quad (1.2)$$

where $\mathbb{P}(dx) = p(x) \lambda(dx)$ and $\mathbb{Q}(dx) = q(x) \lambda(dx)$. Of course, it is non-negative and null if $\mathbb{P} = \mathbb{Q}$ but there is no triangular inequality and therefore, it is not a distance.

In Variational Inference (also called VI), we want to approximate π by selecting the best candidate among a family of densities

$$\{x \mapsto p_\theta(x) : \theta \in \Theta\}$$

with respect to the α -divergence. In other words, we want to solve

$$\operatorname{argmin}_{\theta \in \Theta} D_\alpha(\pi \parallel p_\theta) = \operatorname{argmin}_{\theta \in \Theta} \int_{\mathbf{X}} f_\alpha \left(\frac{p_\theta(x)}{\pi(x)} \right) \pi(x) \lambda(dx).$$

Since π is known only up to multiplicative constant, we must check if the minimization of $\int_{\mathbf{X}} f_\alpha(p_\theta(x)/\pi(x)) \pi(x) \lambda(dx)$ can be equivalent to the minimization of another functional which will be more explicit. Write $\pi(x) = \tilde{\pi}(x)/C$ where $\tilde{\pi}(x)$ is explicit and $C = \int_{\mathbf{X}} \tilde{\pi}(x) \lambda(dx)$ is an "unexplicit" constant. Choosing for example $\alpha = 1$ yields

$$D_1(\pi \parallel p_\theta) = \int_{\mathbf{X}} \log \left(\frac{p_\theta(x)}{\pi(x)} \right) p_\theta(x) \lambda(dx) = \underbrace{\int_{\mathbf{X}} \log \left(\frac{\tilde{\pi}(x)}{p_\theta(x)} \right) p_\theta(x) \lambda(dx)}_{\mathcal{L}_\theta} + \log C.$$

In this litterature, $\mathcal{L}_\theta = \int_{\mathbf{X}} \log(\tilde{\pi}(x)/p_\theta(x)) p_\theta(x) \lambda(dx)$ is called the ELBO (Evidence Lower Bound) and maximizing the ELBO with respect to θ is equivalent to minimizing $D_1(\pi \parallel p_\theta)$ with

respect to θ . Note that since $D_1(\pi||p_\theta) = -\mathcal{L}_\theta + \log C$, we have

$$\mathcal{L}_\theta \leq \log C = \log \int_{\mathbf{X}} \tilde{\pi}(x) \lambda(\mathrm{d}x),$$

hence, the terminology "Evidence Lower Bound". In the case where $\alpha \neq 1$, some equivalent of the ELBO can also be found. In VI, we face an optimisation problem (which is in general a non-convex problem) and we have to resort to all the tools linked with optimisation problems: various stochastic gradient descents and their variants, etc.

1.2.2 Variational auto-encoders

Variational Auto-Encoders (VAEs) are very popular approaches to introduce approximations of a target distribution using latent variables. We consider a family of joint probability distributions $\{(z, x) \mapsto p_\theta(z, x)\}_{\theta \in \Theta}$ on $(\mathbf{Z} \times \mathbf{X}, \mathcal{Z} \times \mathcal{X})$ where Z is a latent variable and X is the observation. In this setting, we often write, for all $\theta \in \Theta$, $x \in \mathbf{X}$, $z \in \mathbf{Z}$,

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

The target distribution is then estimated by

$$p_\theta(x) = \int p_\theta(z)p_\theta(x|z)\mathrm{d}z.$$

Since the integral cannot be computed explicitly, $p_\theta(x)$ is not tractable and therefore the conditional distribution $p_\theta(z|x)$ is not available explicitly. A variational approach can be introduced in this setting by considering a family $\{(z, x) \mapsto q_\varphi(z|x)\}_{\varphi \in \Phi}$. Then, we can write, for all $\varphi \in \Phi$, $\theta \in \Theta$, $x \in \mathbf{X}$, $z \in \mathbf{Z}$,

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

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

$$\mathcal{L}(\theta, \varphi, x) = \mathbb{E}_{q_\varphi(\cdot|x)}\left[\log \frac{p_\theta(Z, x)}{q_\varphi(Z|x)}\right] \quad (1.3)$$

is the ELBO in this setting. In order to approximate $z \mapsto p_\theta(z|x)$, VAE propose to solve the optimization problem:

$$(\hat{\theta}, \hat{\varphi}) \in \operatorname{argmax}_{\theta \in \Theta, \varphi \in \Phi} \left\{ \frac{1}{n} \sum_{i=1}^n \mathcal{L}(\theta, \varphi, X_i) \right\}.$$

Of course, 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.

1.2.3 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 $\pi_1, \dots, \pi_T = \pi$ 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 $(X_k)_{k \in [1:N]} \stackrel{\text{i.i.d.}}{\sim} g$ 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.

1.2.3.1 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.

1.2.3.2 Exploration step

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

$$\pi_1(y) = \int_{\mathbf{X}} \pi_0(d\mathbf{x}) q(\mathbf{x}, y), \quad (1.4)$$

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, $\{(\tilde{X}_k, \omega_k q(X_k, \tilde{X}_k) / r(X_k, \tilde{X}_k))\}_{k \in [1:N]}$ 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 $q(X_k, \tilde{X}_k) / r(X_k, \tilde{X}_k) = 1$).

1.2.3.3 Reweighting step

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

$$\pi_1(x) = \frac{\pi_0(x)g(x)}{\int_{\mathcal{X}} \pi_0(du)g(u)}, \quad (1.5)$$

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 $\pi_1 \rightarrow \pi_2 \rightarrow \dots \rightarrow \pi_T = \pi$, we must check that each step $\pi_i \rightarrow \pi_{i+1}$ corresponds either to (1.4) or (1.5) 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.

1.3 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 algorithm is described in Algorithm 1.3.

Algorithm 1 ABC algorithm.

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

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

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

with $A_{\varepsilon,Y} = \{y; d(y, Y) < \varepsilon\}$. The intuitive idea behind this algorithm is that if $\varepsilon \rightarrow 0$, $\pi_\varepsilon(\theta|Y) \rightarrow \pi(\theta|Y)$ and if $\varepsilon \rightarrow \infty$, $\pi_\varepsilon(\theta|Y) \rightarrow \pi_0(\theta)$. 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 $(d(S(Y_i), S(Y)))_{1 \leq i \leq N}$ which allows to introduce Algorithm 1.3.

Algorithm 2 ABC algorithm with calibrated threshold.

Input: Observation Y , N , integer M_N .
Output: Samples approximately distributed according to $\pi(\theta) = p(\theta|Y)$.
for $i = 1 \rightarrow N$ **do**
 Draw θ_i with prior distribution π_0 .
 Draw Y_i with distribution $p(\cdot|\theta_i)$.
end for
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 .

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. See the exercises for a proof that Algorithm 1.3 targets the correct posterior distribution.

Algorithm 3 ABC within MCMC.

Input: Observation Y , N , threshold ε , output (θ_0, Y_0) from a "standard" ABC.
Output: Samples approximately distributed according to $\pi(\theta) = p(\theta|Y)$.
for $i = 1 \rightarrow N$ **do**
 Sample $\tilde{\theta} \sim q(\cdot|\theta_i)$.
 Sample $\tilde{Y} \sim p(\cdot|\tilde{\theta})$ and compute $S(\tilde{Y})$.
 Compute $\alpha = 1 \wedge \frac{\pi_0(\tilde{\theta})q(\theta_i|\tilde{\theta})}{\pi_0(\theta_i)q(\tilde{\theta}|\theta_i)} \mathbb{1}_{d(S(\tilde{Y}), S(Y)) < \varepsilon}$.
 Sample $U \sim U(0, 1)$.
 if $U < \alpha$ **then**
 Set $\theta_{i+1} = \tilde{\theta}$ and $Y_{i+1} = \tilde{Y}$.
 end if
end for
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 .

Chapter 2

Variational autoencoders

Contents

2.1	Introduction	15
2.2	Gradient-based optimization	16
2.2.1	Pathwise Gradient	17
2.2.2	Score Function Gradient	18
2.3	Marginal likelihood and sampling approaches	19
2.4	Empirical Excess Risk	20
2.5	Classical extensions of VAEs	21
2.5.1	IWAEs	21
2.5.2	β -VAEs	22

2.1 Introduction

Variational Auto-Encoders (VAE) are very popular approaches to introduce approximations of a target conditional distribution in the context of latent data models, see [Rezende et al., 2014, Kingma et al., 2019]. Assume that (X_1, \dots, X_n) are i.i.d. random variables in \mathbf{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 $(\mathbf{Z} \times \mathbf{X}, \mathcal{Z} \times \mathcal{X})$ where Z is a latent variable and X is the observation. In this setting, we often write, for all $\theta \in \Theta$, $x \in \mathbf{X}$, $z \in \mathbf{Z}$,

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

The latent variable generative model defines a joint density $(z, x) \mapsto p_\theta(x, z)$ on $(\mathbf{Z} \times \mathbf{X}, \mathcal{Z} \times \mathcal{X})$ 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

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

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 $p_\theta(X_i)$, $1 \leq i \leq n$, are not available 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 which 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 $\varphi \in \Phi, \theta \in \Theta, x \in \mathbf{X}$,

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

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, \dots, X_n) are i.i.d. with distribution π_{data} , and we aim at solving the optimization problem:

$$(\hat{\theta}_n, \hat{\varphi}_n) \in \text{Argmax}_{\theta \in \Theta, \varphi \in \Phi} \mathcal{L}_n(\theta, \varphi). \quad (2.1)$$

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 $\hat{\mathcal{L}}_n(\theta, \varphi)$:

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

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

2.2 Gradient-based optimization

In order to solve (2.1), we need to compute the gradient of the loss function. Under classical regularity assumptions:

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

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

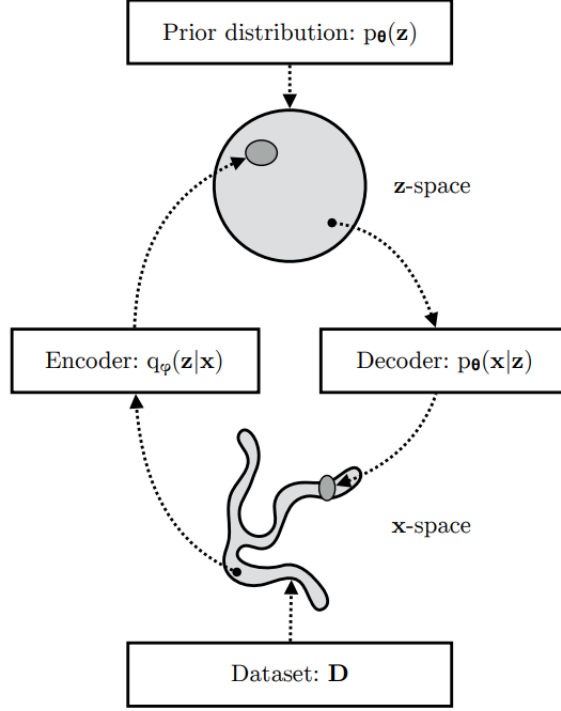


Fig. 2.1 An illustration of a VAE. From [Kingma et al., 2019].

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

where for all $1 \leq i \leq B$, $(Z_{i,1}, \dots, Z_{i,M})_{1 \leq j \leq M}$ 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.

2.2.1 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:

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

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:

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

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

$$\begin{aligned}
S_\varphi(\theta, \varphi; \{X_i\}_{i=1}^B) \\
= -\frac{1}{B} \sum_{i=1}^B \frac{1}{M} \sum_{j=1}^M \left\{ \nabla_z \log \frac{p_\theta(X_i, g(\varepsilon_{i,j}, \varphi))}{q_\varphi(g(\varepsilon_{i,j}, \varphi)|x_i)} \nabla_\varphi g(\varepsilon_{i,j}, \varphi) - \nabla_\varphi \log q_\varphi(g(\varepsilon_{i,j}, \varphi)|X_i) \right\}, \quad (2.2)
\end{aligned}$$

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

Example 2.1. 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 $\text{diag}(\sigma_\varphi^2(x)) \in \mathbb{R}^{d \times d}$ where $(\mu_\varphi(x), \sigma_\varphi^2(x))$ is the output of a neural network with input x . Then, if $z = \mu_\varphi(x) + \text{diag}(\sigma_\varphi(x))\varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \text{Id})$, $z \sim q_\varphi(\cdot|x)$. Writing $g(\varepsilon, \varphi) = \mu_\varphi(x) + \sigma_\varphi(x)\varepsilon$, the Jacobian of g with respect to ε is

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

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

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

2.2.2 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 2.2. *For all $\theta \in \Theta$, $\varphi \in \Phi$, we have:*

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

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

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

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

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

□

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

$$S_\varphi(\theta, \varphi; \{X_i\}_{i=1}^B) = -\frac{1}{B} \sum_{i=1}^B \frac{1}{M} \sum_{j=1}^M \log \frac{p_\theta(X_i, Z_{i,j})}{q_\varphi(Z_{i,j}|X_i)} \nabla_\varphi \log \frac{p_\theta(X_i, Z_{i,j})}{q_\varphi(Z_{i,j}|X_i)}, \quad (2.3)$$

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

Algorithm 4 displays an example of stochastic gradient descent to estimate θ and φ using Adam.

Algorithm 4 Adam Algorithm for ELBO Maximization

Input: Initial estimates θ_0, φ_0 , maximum number of iterations n , step sizes $\{\gamma_k\}_{k \geq 1}$, momentum parameters $\beta_1, \beta_2 \in [0, 1)$, regularization parameter $\delta \geq 0$ and batch size B .
Set $m_0 = 0$ and $v_0 = 0$.
for $k = 0$ to $k = n - 1$ **do**
 Sample a mini-batch $\{X_i\}_{i=1}^B$.
 Compute the stochastic gradient $g_{k+1} = (S_\theta(\theta_k, \varphi_k; \{X_i\}_{i=1}^B), S_\varphi(\theta_k, \varphi_k; \{X_i\}_{i=1}^B))^\top$.
 Set $m_{k+1} = \beta_1 m_k + (1 - \beta_1) g_{k+1}$.
 Set $v_{k+1} = \beta_2 v_k + (1 - \beta_2) g_{k+1} \odot g_{k+1}$.
 Set $(\theta_{k+1}, \varphi_{k+1}) = (\theta_k, \varphi_k) - \gamma_{k+1} [\delta I_d + v_k]^{-1/2} m_k$.
end for

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.

2.3 Marginal likelihood and sampling approaches

Marginal likelihood. For all parameters (θ, φ) , we can estimate the marginal loglikelihood using importance sampling approaches. Note that

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

Therefore, if (Z_1, \dots, Z_M) are i.i.d. with distribution $q_\varphi(\cdot|x)$, $\log p_\theta(x)$ may be estimated by

$$\ell_{\theta, M}(x) = \log \left(\frac{1}{M} \sum_{i=1}^M \frac{p_\theta(Z_i, x)}{q_\varphi(Z_i|x)} \right).$$

Choice of the prior. A common way to sample data using a VAE is to use the fact that

$$p_\theta(x) = \int p_\theta(z) p_\theta(x|z) dz,$$

and therefore, sampling x from p_θ amounts to sampling $z \sim p_\theta(z)$ and sampling $x \sim p_\theta(x|z)$. However, the prior density $p_\theta(z)$ is often chosen to be very simple, e.g. a Gaussian distribution, which often leads to poor sampling performance. An alternative suggested in [Tomczak and Welling, 2018] is to choose a prior that optimizes the ELBO which is given by

$$p_\theta(z) = \frac{1}{n} \sum_{i=1}^n q_\varphi(z|X_i).$$

This prior is referred to as the Variational Mixture of Posteriors prior. Sampling from this prior amounts to choose an observation X_i uniformly at random and sampling $z \sim q_\varphi(z|X_i)$.

2.4 Empirical Excess Risk

Despite the empirical success of VAE, general theoretical results on the statistical properties of VAEs are very recent although this is crucial to guide hyperparameters choice and the design of the prior, encoder and decoder distributions. In this section, we present some results provided in [Tang and Yang, 2021] to analyze the excess risk for learning densities using VAEs and in particular to choose good prior distributions. The key insight of [Tang and Yang, 2021] is to describe the VAE estimator as a M-estimator:

$$(\hat{\theta}_n, \hat{\varphi}_n) \in \text{Argmin}_{\theta \in \Theta, \varphi \in \Phi} \frac{1}{n} \sum_{i=1}^n m(\theta, \varphi, X_i),$$

where, assuming that the data distribution has a density π_{data} ,

$$m(\theta, \varphi, X_i) = \log \frac{\pi_{\text{data}}(x)}{p_\theta(x)} + \text{KL}(q_\varphi(\cdot|x) \| p_\theta(\cdot|x)),$$

with $p_\theta(z|x) \propto p_\theta(z)p_\theta(x|z)$ is the distribution of the latent variable given the observation. Let $\alpha > 0$ and $\psi_\alpha : x \mapsto e^{x^\alpha} - 1$ and for all random variable X define the Orlicz norm of X by

$$\|X\|_{\psi_\alpha} = \inf \left\{ \lambda > 0 ; \mathbb{E} \left[\psi_\alpha \left(\frac{|X|}{\lambda} \right) \right] \leq 1 \right\}.$$

Note that this norm is related to the tail of the distribution as for all $t > 0$,

$$\mathbb{P}(|X| \geq t) \leq 2 \exp \left\{ - \left(\frac{t}{\|X\|_{\psi_\alpha}} \right)^\alpha \right\}.$$

Assumption H1 is a tail condition on the loss function. It is shown in [Tang and Yang, 2021] that it can be satisfied by Deep Gaussian models.

H1 For a random variable X with distribution π_{data} there exist $\alpha, D > 0$ such that

$$\left\| \sup_{\theta \in \Theta, \varphi \in \Phi} \left\{ \left| \log \frac{\pi_{\text{data}}(X)}{p_\theta(X)} \right| + \text{KL}(q_\varphi(\cdot|X) \| p_\theta(\cdot|X)) \right\} \right\|_{\psi_\alpha} \leq D.$$

H2 Assume that there exist $a_0, a_1 > 0$ and a real-valued function b such that for all $x \in \mathbf{X}$, $\theta, \theta' \in \Theta$, $\varphi, \varphi' \in \Phi$,

$$\|\theta\|_\infty + \|\varphi\|_\infty \leq a_0$$

and

$$|m(\theta, \varphi, x) - m(\theta', \varphi', x)| \leq b(x) \|(\theta, \varphi) - (\theta', \varphi')\|_2,$$

with $\mathbb{E}_{\pi_{\text{data}}} [b(X)] \leq a_1$.

Theorem 2.3. *Assume that H1-2 hold. Then, there exist $c_0, c_1, c_2, d > 0$ such that with probability at least $1 - c_0 \exp\{-c_1(d \log n)^{1 \wedge \alpha}\}$,*

$$\mathbb{E}_{\pi_{\text{data}}} \left[m(\hat{\theta}_n, \hat{\varphi}_n, X) \right] \leq \inf_{\gamma > 0} \left\{ (1 + \gamma) \min_{\theta \in \Theta, \varphi \in \Phi} \mathbb{E}_{\pi_{\text{data}}} [m(\theta, \varphi, X)] + c_2(1 + \gamma^{-1}) \frac{dD}{n} \log(nd) \log^{1/\alpha} n \right\}.$$

The oracle inequality given in Theorem 2.3 can be used to control the total variation distance between the target distribution and the generative model using a Variational Mixture of Posteriors prior.

Proposition 2.4. *Assume that H1-2 hold. Assume also that for all x , the support \mathcal{Z} of $q_\varphi(\cdot|x)$ is bounded and that there exists $a_3 > 0$ such that for all x , all $z, z' \in \mathcal{Z}$, $\varphi, \varphi' \in \Phi$,*

$$|q_\varphi(z|x) - q_{\varphi'}(z'|x)| \leq a_3 (\|\varphi - \varphi'\|_2 + \|z - z'\|_2).$$

Then, there exist $c_0, c_1, c_2, d > 0$ such that with probability at least $1 - c_0 \exp\{-c_1(d \log n)^{1/\alpha}\}$,

$$\begin{aligned} D_{\text{tv}}^2 \left(\pi_{\text{data}}, \int \left(\frac{1}{n} \sum_{i=1}^n q_{\hat{\varphi}}(z|X_i) \right) p_{\hat{\theta}}(\cdot|z) dz \right) \\ \leq c_2 \min_{\theta \in \Theta, \varphi \in \Phi} \mathbb{E}_{\pi_{\text{data}}} [m(\theta, \varphi, X)] + c_3 \frac{dD}{n} \log(nd) \log^{1/\alpha} n. \end{aligned}$$

2.5 Classical extensions of VAEs

2.5.1 IWAEs

The Importance Weighted Autoencoder (IWAE) [Burda et al., 2015] is a variant of VAE that uses importance weighting to obtain a tighter evidence lower bound. The IWAE objective function is defined as:

$$\mathcal{L}_K^{\text{IWAE}}(\theta, \varphi) = \mathbb{E}_{\pi_{\text{data}}} \left[\mathbb{E}_{q_\varphi^{\otimes K}(\cdot|X)} \left[\log \frac{1}{K} \sum_{\ell=1}^K \frac{p_\theta(X, Z^{(\ell)})}{q_\varphi(Z^{(\ell)}|X)} \right] \right], \quad (2.4)$$

where $K \geq 1$ is the number of samples drawn from the variational posterior distribution. Define

$$\mathcal{L}_K^{\text{IWAE}}(\theta, \varphi, x) = \mathbb{E}_{q_\varphi^{\otimes K}(\cdot|x)} \left[\log \frac{1}{K} \sum_{\ell=1}^K \frac{p_\theta(x, Z^{(\ell)})}{q_\varphi(Z^{(\ell)}|x)} \right].$$

The motivation of (2.4) is given by Proposition 2.5.

Proposition 2.5. *For all $K \geq 1$, $\theta \in \Theta$, $\varphi \in \Phi$, $x \in \mathcal{X}$,*

$$\mathcal{L}_K^{\text{IWAE}}(\theta, \varphi, x) \leq \mathcal{L}_{K+1}^{\text{IWAE}}(\theta, \varphi, x) \leq \log p_\theta(x).$$

Proof. By Jensen's inequality,

$$\mathcal{L}_K^{\text{IWAE}}(\theta, \varphi, x) \leq \log \mathbb{E}_{q_\varphi^{\otimes K}(\cdot|x)} \left[\frac{1}{K} \sum_{\ell=1}^K \frac{p_\theta(x, Z^{(\ell)})}{q_\varphi(Z^{(\ell)}|x)} \right] \leq \log p_\theta(x).$$

By definition,

$$\mathcal{L}_{K+1}^{\text{IWAE}}(\theta, \varphi, x) = \mathbb{E}_{q_\varphi^{\otimes K+1}(\cdot|x)} \left[\log \frac{1}{K+1} \sum_{\ell=1}^{K+1} \frac{p_\theta(x, Z^{(\ell)})}{q_\varphi(Z^{(\ell)}|x)} \right].$$

□

2.5.2 β -VAEs

β -VAEs are state-of-the-art models used for unsupervised disentangled representation learning. They are designed to adjust the balance between the reconstruction term and latent space regularization by introducing a parameter $\beta > 0$. Note that the ELBO defined in (1.3) can be written

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

where p_θ in the Kullback-Leibler divergence is the prior distribution on the latent space. The ELBO of β -VAE is then define as:

$$\mathcal{L}_{\beta\text{-VAE}}(\theta, \phi, x) = \mathbb{E}_{q_\phi(\cdot|x)} [\log p_\theta(x|z)] - \beta \text{KL}(q_\phi(\cdot|x) \| p_\theta),$$

where the Lagrangian multiplier $\beta > 0$ is considered as a hyperparameter.

Chapter 3

Score-based diffusion models

Contents

3.1	Motivations of score-based approaches	23
3.2	Score-based generative models	24
3.3	Some theoretical guarantees	28

3.1 Motivations of score-based approaches

Assume that (X_1, \dots, X_n) are i.i.d. random variables in \mathbf{X} with unknown probability distribution function π_{data} . In previous chapters, we considered parametric probability density functions p_θ to model π_{data} . Such approaches may be cumbersome as the normalizing constant (Z_θ in the case of Energy-based models for instance) that makes p_θ a probability density function is not available. Sequential or Markov chain Monte Carlo methods provide solution to sample approximately from a probability density known up to a multiplicative constant but they can face slow convergence rates and require a lot of tuning in high dimensional settings. It is also crucial to note that the complexity of real data prohibits a thorough depiction of the distribution π_{data} through standard non-parametric density estimation strategies.

Modeling the score function $x \mapsto \nabla_x \log \pi_{\text{data}}(x)$ directly allows to overcome this difficulty. Score-based models can be trained without considering normalizing constants. Score-matching methods aim at estimating $x \mapsto \nabla_x \log \pi_{\text{data}}(x)$ using only the samples (X_1, \dots, X_n) and stochastic optimization procedures. In the approaches proposed by [Hyvärinen and Dayan, 2005, Vincent, 2011] the optimization procedures minimize

$$\theta \mapsto \mathcal{L}_{\text{uc}}(\theta) = \mathbb{E}_{\pi_{\text{data}}} \left[\|s_\theta(X) - \nabla_x \log \pi_{\text{data}}(X)\|_2^2 \right],$$

where s_θ is the parametric model for $\nabla_x \log \pi_{\text{data}}$. Of course, this loss function cannot be minimized as $\nabla_x \log \pi_{\text{data}}$ is unknown. However, [Hyvärinen and Dayan, 2005, Theorem 1] established that for all θ ,

$$\mathcal{L}_{\text{uc}}(\theta) = \mathbb{E}_{\pi_{\text{data}}} \left[\sum_{j=1}^d \left(\partial_j s_\theta(X) + \frac{1}{2} s_{j,\theta}(X)^2 \right) \right] + C,$$

where $s_{j,\theta}(X)$ is the j -th component of $s_\theta(X)$ and C is a constant that does not depend on θ . Therefore, the score estimator can be trained using stochastic optimization and an empirical estimate of the expectation under π_{data} computed with the samples (X_1, \dots, X_n) .

Langevin dynamics offers a very good motivation to learn the score of the target distribution instead of π_{data} itself. Starting with X_0 sampled from some proposal distribution, Langevin dynamics iteratively samples $(X_k)_{k \geq 1}$ with, for all $k \geq 1$,

$$X_k = X_{k-1} + \gamma_k \nabla_x \log \pi_{\text{data}}(X_{k-1}) + \sqrt{2\gamma_k} \varepsilon_k,$$

where $(\gamma_k)_{k \geq 1}$ is a sequence of positive step-sizes and $(\varepsilon_k)_{k \geq 1}$ are i.i.d. standard Gaussian random variables independent of X_0 . In [Durmus and Moulines, 2017], the authors obtained nonasymptotic bounds for the convergence to the target distribution in total variation distance for both constant and decreasing step-sizes. Of course, $\nabla_x \log \pi_{\text{data}}$ is not available and using Langevin dynamics with approximate score functions introduce additional errors which have to be controlled. In addition, the main limitation of this approach is that it requires to learn simultaneously the score estimate $s_\theta(X)$ and its partial derivatives $\partial_j s_\theta(X)$, $1 \leq j \leq d$, which can be computationally very sensitive.

In [Song and Ermon, 2019], the authors propose to noise data with multiple scales of noise $0 < \sigma_1 < \dots < \sigma_L$. The data distribution is then perturbed with Gaussian noise by considering, for all $1 \leq i \leq L$,

$$p_{\sigma_i} : x \mapsto \int \pi_{\text{data}}(z) \varphi_{z, \sigma_i^2 I_d}(x) dz,$$

where $\varphi_{\mu, \Sigma}$ is the Gaussian probability density function with mean μ and variance Σ . For all $1 \leq i \leq L$, p_{σ_i} is the probability density function of $X + \sigma_i Z$ where $X \sim \pi_{\text{data}}$ and $Z \sim \mathcal{N}(0, I_d)$ are independent. Then, we aim at estimating the score function associated with p_{σ_i} , $1 \leq i \leq L$. The authors of [Song and Ermon, 2019] introduced a weighted sum of mean squared errors:

$$\theta \mapsto \mathcal{L}_{\text{wuc}}(\theta) = \sum_{i=1}^L \lambda_i \mathbb{E}_{p_{\sigma_i}} \left[\|s_\theta(X, \sigma_i) - \nabla_x \log p_{\sigma_i}(X)\|_2^2 \right],$$

where $\lambda_i > 0$ and $s_\theta(X, \sigma_i)$ is the parametric estimation of $\nabla_x \log p_{\sigma_i}(X)$, $1 \leq i \leq L$. After training all conditional scores, samples approximately distributed according to π_{data} can be obtained with annealed Langevin dynamics. For all $i = L, \dots, 1$, $1 \leq k \leq n$, write

$$X_k^{(i)} = X_{k-1}^{(i)} + \gamma_{i,k} s_\theta(X_{k-1}^{(i)}, \sigma_i) + \sqrt{2\gamma_{i,k}} \varepsilon_{i,k},$$

where $X_0^{(L)}$ is sampled from an arbitrary distribution, $X_0^{(i)} = X_n^{(i+1)}$ for $L-1 \leq i \leq 1$, $(\gamma_{i,k})_{1 \leq k \leq n, 1 \leq i \leq L}$ is a sequence of positive step-sizes and $(\varepsilon_{i,k})_{1 \leq k \leq n, 1 \leq i \leq L}$ are i.i.d. standard Gaussian random variables.

Although adding noise to learn perturbed score functions and annealed Langevin dynamics are appealing to sample approximately from π_{data} , this approach is computationally intensive and depends on many sensitive hyperparameters: choice of L and standard deviations $0 < \sigma_1 < \dots < \sigma_L$, length of the Langevin dynamics n . Score-based generative models based on diffusion propose to introduce a continuous-time noising dynamics to efficiently learn the generative process without having to choose predefined noise levels.

3.2 Score-based generative models

Score-based Generative Models (SGMs) are probabilistic models designed to address this challenge using two main phases. The first phase, also referred to as the forward phase, involves progressively perturbing the empirical distribution by adding noise to the training data until its distribution approximately reaches an easy-to-sample distribution π_∞ . The second phase involves learning to reverse this noising dynamics by sequentially removing the noise, which is referred to as the sampling phase or backward phase. Reversing the dynamics during the backward phase would require in principle knowledge of the score function, i.e., the gradient of the logarithm of the

density at each time step of the diffusion. To circumvent this issue, the score function is learned by training a deep neural network. When applying these learned reverse dynamics to samples from π_∞ , we obtain a generative distribution that approximates π_{data} .

Forward process

Denote as $\beta : [0, T] \mapsto \mathbb{R}_{>0}$ the noise schedule, assumed to be continuous and non decreasing. Although originally developed using a finite number of noising steps [Sohl-Dickstein et al., 2015, Song and Ermon, 2019, Ho et al., 2020], most recent approaches consider time-continuous noise perturbations through the use of stochastic differential equations (SDEs) [Song et al., 2021]. Consider, therefore, a forward process given by

$$d\vec{X}_t = -\frac{\beta(t)}{2\sigma^2} \vec{X}_t dt + \sqrt{\beta(t)} dB_t, \quad \vec{X}_0 \sim \pi_{\text{data}}. \quad (3.1)$$

We denote by p_t the density of \vec{X}_t at time $t \in (0, T]$. Note that, up to the time change $t \mapsto \int_0^t \beta(s)/2ds$, this process corresponds to the standard Ornstein-Uhlenbeck (OU) process, solution to

$$d\vec{X}_t = -\frac{1}{\sigma^2} \vec{X}_t dt + \sqrt{2} dB_t, \quad \vec{X}_0 \sim \pi_{\text{data}},$$

see, e.g., [Karatzas and Shreve, 2012, Chapter 3]. Due to the linear nature of the drift with respect to $(X_t)_t$, it is well-known that an exact simulation can be performed for this process. Indeed, the marginal distribution of (3.1) at time t writes as

$$\vec{X}_t = m_t X_0 + \sigma_t Z, \quad (3.2)$$

with $Z \sim \mathcal{N}(0, I_d)$ independent of X_0 , $X_0 \sim \pi_{\text{data}}$,

$$m_t = \exp \left\{ -\int_0^t \beta(s) ds / (2\sigma^2) \right\} \quad \text{and} \quad \sigma_t^2 = \sigma^2 \left(1 - \exp \left\{ -\int_0^t \beta(s) / \sigma^2 ds \right\} \right). \quad (3.3)$$

Therefore, sampling from the forward process only necessitates access to samples from π_0 and $\mathcal{N}(0, I_d)$. Lemma 3.1 establishes that the stationary distribution π_∞ of the forward process is the Gaussian distribution with mean 0 and variance $\sigma^2 I_d$.

Lemma 3.1. *Assume that β is continuous, positive, non decreasing and such that $\int_0^\infty \beta(t) dt = \infty$. Let $(\vec{X}_t)_{t \geq 0}$ be a weak solution to the forward process (3.1). Then, the stationary distribution of $(\vec{X}_t)_{t \geq 0}$ is Gaussian with mean 0 and variance $\sigma^2 I_d$.*

Proof. Consider the process

$$\bar{X}_t = \exp \left(\frac{1}{2\sigma^2} \int_0^t \beta(s) ds \right) \vec{X}_t.$$

Itô's formula yields

$$\vec{X}_t = \exp \left(-\frac{1}{2\sigma^2} \int_0^t \beta(s) ds \right) \left(\vec{X}_0 + \int_0^t \sqrt{\beta(s)} \exp \left(\int_0^s \beta(u) / (2\sigma^2) du \right) dB_s \right).$$

Note that

$$\lim_{t \rightarrow \infty} \exp \left(-\frac{1}{2\sigma^2} \int_0^t \beta(s) ds \right) \vec{X}_0 = 0.$$

Then, the second term in the r.h.s. of (3.2) is Gaussian with mean 0 and variance $\sigma_t^2 \mathbf{I}_d$, where

$$\sigma_t^2 = \exp\left(-\frac{1}{\sigma^2} \int_0^t \beta(s) ds\right) \int_0^t \beta(s) e^{\int_0^s \beta(u)/\sigma^2 du} ds = \sigma^2 \left(1 - \exp\left(-\frac{1}{\sigma^2} \int_0^t \beta(s) ds\right)\right).$$

By assumption on β , $\lim_{t \rightarrow \infty} \sigma_t^2 = \sigma^2$, which concludes the proof. \square

In the literature, when $t \mapsto \beta(t)$ is constant equal to 2 (meaning that there is no time change), this diffusion process is referred to as the Variance-Preserving SDE [Conforti et al., 2023], leading to the so-called Denoising Diffusion Probabilistic Models [Ho et al., 2020]. Several continuous-time models could be used to noise data sampled from π_{data} . The choice of (3.1) is deeply related to two main properties.

- It is straightforward and computationally easy to sample from p_t i.e. to obtain a perturbed data point at all time steps by (3.2).
- The stationary distribution of (3.1), i.e. the asymptotic distribution of perturbed data when $t \rightarrow \infty$, is known and easy to sample from.

Backward process

Let $T > 0$ be the time horizon. The backward process on $[0, T]$ associated with (3.1) is given by $\overleftarrow{X}_0 \sim p_T$ and

$$\overleftarrow{X}_t = \eta(t, \overleftarrow{X}_t) dt + \sqrt{\bar{\beta}(t)} dB_t, \quad (3.4)$$

where

$$\bar{\beta}(t) := \beta(T - t) \quad \text{and} \quad \eta(t, \overleftarrow{X}_t) := \frac{\bar{\beta}(t)}{2\sigma^2} \overleftarrow{X}_t + \bar{\beta}(t) \nabla \log p_{T-t}(\overleftarrow{X}_t).$$

This SDE is crucial in score-based generative models since if $(\overleftarrow{X}_t)_{0 \leq t \leq T}$ is a weak solution to (3.4), then $(\overrightarrow{X}_t)_{0 \leq t \leq T}$ and $(\overleftarrow{X}_t)_{0 \leq t \leq T}$ have the same distribution. This means that if $\overleftarrow{X}_0 \sim p_T$, i.e. \overleftarrow{X}_0 is distributed as a noised data point at time T and if \overleftarrow{X}_T is obtained by sampling from (3.4), then $\overleftarrow{X}_T \sim \pi_{\text{data}}$. In order to use this property to sample approximately from π_{data} we need to consider three steps.

1. For all $0 \leq t \leq T$, $\nabla \log p_{T-t}$ depends on the score of the data distribution and is unknown. This quantity has to be replaced by an estimator in (3.4).
2. Even if the score is replaced by an estimator, we do not know how to sample from (3.4). In standard approaches, SBGMs usually use discretization techniques to obtain samples.
3. In practice, the backward process is initialized with the stationary distribution, i.e. $\overleftarrow{X}_0 \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$, instead of sampled from p_T .

These three steps allow to introduce a training procedure to learn the score functions and a sampling procedure to obtain random variables approximately distributed according to π_{data} . However, they all introduce approximation errors that need to be controlled if we aim at obtaining convergence guarantees.

Score estimation

Simulating the backward process means knowing how to operate the score. However, the score function $x \mapsto \nabla \log p_t(x)$ cannot be evaluated directly, because it depends on the unknown data distribution. In [Hyvärinen and Dayan, 2005], the authors proposed to estimate the score function associated with a distribution by minimizing the expected L^2 -squared distance between the true score function and the proposed approximation. In the context of diffusion models, this is typically

done with the use of a deep neural network architecture $s_\theta : [0, T] \times \mathbb{R}^d \mapsto \mathbb{R}^d$ parameterized by $\theta \in \Theta$, and trained to minimize:

$$\mathcal{L}_{\text{explicit}}(\theta) = \mathbb{E} \left[\left\| s_\theta \left(\tau, \vec{X}_\tau \right) - \nabla \log p_\tau \left(\vec{X}_\tau \right) \right\|_2^2 \right], \quad (3.5)$$

with $\tau \sim \mathcal{U}(0, T)$ independent of the forward process $(\vec{X}_t)_{t \geq 0}$. However, this estimation problem still suffers from the fact that the regression target is not explicitly known. A tractable optimization problem sharing the same optima can be defined though, through the marginalization over π_{data} of p_τ [Song et al., 2021]:

$$\mathcal{L}_{\text{score}}(\theta) = \mathbb{E} \left[\left\| s_\theta \left(\tau, \vec{X}_\tau \right) - \nabla \log p_\tau \left(\vec{X}_\tau | X_0 \right) \right\|_2^2 \right], \quad (3.6)$$

where τ is uniformly distributed on $[0, T]$, and independent of $X_0 \sim \pi_{\text{data}}$ and $\vec{X}_\tau \sim p_\tau(\cdot | X_0)$. This loss function is appealing as it only requires to know the transition kernel of the forward process. In (3.1), this is a Gaussian kernel with explicit mean and variance. By (3.2), we obtain

$$\mathcal{L}_{\text{score}}(\theta) = \mathbb{E} \left[\left\| s_\theta \left(\tau, \vec{X}_\tau \right) + \frac{\vec{X}_\tau - m_\tau X_0}{\sigma_\tau^2} \right\|_2^2 \right], \quad (3.7)$$

where $t \rightarrow m_t$ and $t \rightarrow \sigma_t^2$ are defined in (3.3).

Discretization and sampling

Once the score function is learned, it remains that, in most cases, the backward dynamics no longer enjoys a linear drift, which makes its exact simulation challenging. To address this issue, one solution is to discretize the continuous dynamics of the backward process. In this way, [Song et al., 2021] propose an Euler-Maruyama (EM) discretization scheme in which both the drift and the diffusion coefficients are discretized recursively. The Euler Exponential Integrator can also be used, see for instance [Conforti et al., 2023]. Consider $N \geq 1$ and a time discretization $t_0 = 0 < t_1 < \dots < t_N = T$. The Euler-Maruyama process $(\hat{X}_t^\theta)_{t \in [0, T]}$ is such that, for $t \in [t_k, t_{k+1}]$,

$$d\hat{X}_t^\theta = \bar{\beta}(t_k) \left(\frac{1}{2\sigma^2} \hat{X}_{t_k}^\theta + s_\theta \left(T - t_k, \hat{X}_{t_k}^\theta \right) \right) dt + \sqrt{\bar{\beta}(t_k)} dB_t, \quad \hat{X}_0^\theta \sim \pi_\infty. \quad (3.8)$$

In this setting, $(\hat{X}_t^\theta)_{t \in \{t_0, \dots, t_N\}}$ can be sampled exactly. Starting with $\hat{X}_0^\theta \sim \pi_\infty$, write for $0 \leq k \leq N-1$,

$$\hat{X}_{t_{k+1}}^\theta = \hat{X}_{t_k}^\theta + \bar{\beta}(t_k) \left(\frac{1}{2\sigma^2} \hat{X}_{t_k}^\theta + s_\theta \left(T - t_k, \hat{X}_{t_k}^\theta \right) \right) (t_{k+1} - t_k) + \sqrt{\bar{\beta}(t_k)(t_{k+1} - t_k)} Z_{k+1},$$

where $\{Z_k\}_{1 \leq k \leq N}$ are i.i.d. with standard Gaussian distribution and independent of \hat{X}_0^θ . Deonte by $\hat{\pi}_N^{(\beta, \theta)}$ the distribution of $\hat{X}_{t_N}^\theta$, i.e. the distribution of the random variable obtained after N steps of the discretized backward process. After training of θ , this distribution is the generative model which aims at sampling approximately from π_{data} .

3.3 Some theoretical guarantees

Many theoretical results have been proposed in the literature to control a distance or pseudo-distance between $\hat{\pi}_N^{(\beta, \theta)}$ and π_{data} . Most results focus on controls for the Kullback-Leibler divergence, the total variation and 2-Wasserstein distances, see for instance [De Bortoli et al., 2021, Conforti et al., 2023, Chen et al., 2023]. Consider the following additional notations. The modified marginal distribution is the marginal distribution of the forward process divided by the density of its stationary distribution defined, for all $x \in \mathbb{R}^d$ and all $0 \leq t \leq T$, by

$$\tilde{p}_t(x) := p_t(x)/\varphi_{\sigma^2}(x),$$

and $\tilde{s}_\theta(t, x) = s_\theta(t, x) + x/\sigma^2$ where φ_{σ^2} denote the density function of π_∞ , a Gaussian distribution with mean 0 and variance $\sigma^2 \text{I}_d$. The modified score function is $\nabla \log \tilde{p}_t(x) = \nabla \log p_t(x) + x/\sigma^2$.

Theorem 3.2. *Define $h := \sup_{k \in \{1, \dots, N\}} (t_k - t_{k-1})$. Then,*

$$\text{KL} \left(\pi_{\text{data}} \middle| \middle| \hat{\pi}_N^{(\beta, \theta)} \right) \leq \mathcal{E}_1^{\text{KL}}(\beta) + \mathcal{E}_2^{\text{KL}}(\theta, \beta) + \mathcal{E}_3^{\text{KL}}(\beta),$$

where

$$\begin{aligned} \mathcal{E}_1^{\text{KL}}(\beta) &= \text{KL}(\pi_{\text{data}} | \pi_\infty) \exp \left\{ -\frac{1}{\sigma^2} \int_0^T \beta(s) ds \right\}, \\ \mathcal{E}_2^{\text{KL}}(\theta, \beta) &= \sum_{k=0}^{N-1} \mathbb{E} \left[\left\| \nabla \log \tilde{p}_{T-t_k}(\vec{X}_{T-t_k}) - \tilde{s}_\theta(T-t_k, \vec{X}_{T-t_k}) \right\|^2 \right] \int_{T-t_{k+1}}^{T-t_k} \beta(t) dt, \\ \mathcal{E}_3^{\text{KL}}(\beta) &= 2h\beta(T)\mathcal{I}(\pi_{\text{data}} | \pi_\infty). \end{aligned}$$

The upper bound of Theorem 3.2 involves three different types of errors. The term $\mathcal{E}_1^{\text{KL}}$ represents the mixing time of the OU forward process, arising from the practical limitation of considering the forward process up to a finite time T . The second term $\mathcal{E}_2^{\text{KL}}$ corresponds to the approximation error, which stems from the use of a deep neural network to estimate the score function. Finally, $\mathcal{E}_3^{\text{KL}}$ is the discretization error of the EI discretization scheme. This last term vanishes as the discretization grid is refined (i.e., $h \rightarrow 0$).

Chapter 4

Basics in Markov chains

Contents

4.1	Main notation	29
4.2	Definitions	29
4.2.1	Additional notation	31
4.3	Canonical space	32
4.3.1	A simpler problem	32
4.3.2	The general case	32
4.3.3	The Markov property	33

4.1 Main notation

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

- $M_+(X)$ is the set of non-negative measures on (X, \mathcal{X}) .
- $M_1(X)$ is the set of probability measures on (X, \mathcal{X}) .
- $F(X)$ is the set of real-valued measurable functions f on X and $F_+(X)$ the set of non-negative measurable functions on X .
- If $k \leq \ell$, $u_{k:\ell}$ means (u_k, \dots, u_ℓ) and $u_{k:\infty}$ means $(u_{k+\ell})_{\ell \in \mathbb{N}}$.

4.2 Definitions

Definition 4.1. We say that $P : X \times \mathcal{X} \rightarrow \mathbb{R}^+$ is a Markov kernel, if for all $(x, A) \in X \times \mathcal{X}$,

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

For all $(x, A) \in X \times \mathcal{X}$, 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, $P(x, X) = 1$ for all $x \in X$. Since $P(x, \cdot)$ is a measure, we also use the infinitesimal notation: $P(x, dy)$. For example,

$$P(x, A) = \int_X \mathbf{1}_A(y) P(x, dy) = \int_A P(x, dy).$$

In almost all the course, a Markov kernel P allows to move a point x from a measurable space (X, \mathcal{X}) to another point on the same measurable space, that is, P is defined on $X \times \mathcal{X}$ but we

can more generally define a Markov kernel from a measurable space (X, \mathcal{X}) to another measurable space (Y, \mathcal{Y}) . In such case, P will be a Markov kernel on $X \times Y$.

Definition 4.2. 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 $\nu \in \mathcal{M}_1(X)$ if and only if

- (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 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 actually 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.

If $\{\mathcal{F}_k : k \in \mathbb{N}\}$ is a sequence of embedded sigma-fields on X (that is, $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all $k \in \mathbb{N}$), then $\{\mathcal{F}_k : k \in \mathbb{N}\}$ is called a filtration on X and we say that $(X_k)_{k \geq 0}$ is $(\mathcal{F}_k)_{k \geq 0}$ -adapted if X_k is $\mathcal{G}/\mathcal{F}_k$ -measurable for all $k \in \mathbb{N}$. Of course, the most natural filtration for $\{X_k : k \in \mathbb{N}\}$ is indeed $\mathcal{F}_k = \sigma(X_{0:k})$ and unsurprisingly, we call it the natural filtration. But other possibilities exist, where \mathcal{F}_k is enlarged to include some other variables alongside with $X_{0:k}$. For example, let $\{Y_k : k \in \mathbb{N}\}$ be any other sequence of random variables on $(\Omega, \mathcal{G}, \mathbb{P})$ and taking values in X (we do not assume anything on the relation between $\{X_k : k \in \mathbb{N}\}$ and $\{Y_k : k \in \mathbb{N}\}$). A typical example corresponds to $X_{k+1} = g(X_{0:k}, Y_{0:k})$, but we do not even need to assume that for the moment. Set $\mathcal{F}_k = \sigma(X_{0:k}, Y_{0:k})$ and assume that

$$\mathbb{P}(X_{k+1} \in A | \mathcal{F}_k) = P(X_k, A), \quad \mathbb{P} - \text{a.s.} \quad (4.1)$$

Since X_ℓ is \mathcal{F}_ℓ -measurable and $\mathcal{F}_\ell \subset \mathcal{F}_k$ for $\ell \leq k$, we deduce that $\sigma(X_{0:k}) \subset \mathcal{F}_k$. This allows to apply the tower property, which yields

$$\begin{aligned} \mathbb{P}(X_{k+1} \in A | X_{0:k}) &= \mathbb{E}[\mathbb{P}(X_{k+1} \in A | \mathcal{F}_k) | X_{0:k}] \\ &= \mathbb{E}[P(X_k, A) | X_{0:k}] = P(X_k, A), \quad \mathbb{P} - \text{a.s.} \end{aligned}$$

and therefore if we assume (4.1), then as soon as $(X_k)_{k \geq 0}$ is $(\mathcal{F}_k)_{k \geq 0}$ -adapted, we can conclude that $\{X_k : k \in \mathbb{N}\}$ is a Markov chain with Markov kernel P . Why is it useful? Well, Sometimes, we define iteratively X_{k+1} using other variables rather than $X_{0:k}$ only and therefore, considering $\mathbb{P}(X_{k+1} \in A | \mathcal{F}_k)$ is easier to deal with. Let us see it in action with a very simple example.

Example 4.3. Let $\{\epsilon_k : k \geq 1\}$ be i.i.d. random variables on \mathbb{R}^p with density f with respect to the Lebesgue measure on \mathbb{R}^p , and let $X_0 \sim \mu$. We assume that X_0 is independent of $\{\epsilon_k : k \in \mathbb{N}\}$. Define

$$X_{k+1} = aX_k + b\epsilon_{k+1}, \quad k \in \mathbb{N}.$$

Set $\mathcal{F}_0 = \sigma(X_0)$ and for $k \geq 1$, $\mathcal{F}_k = \sigma(X_0, \epsilon_{1:k})$. Since X_ℓ is a deterministic function of X_0 and $\epsilon_{1:\ell}$, we deduce that $(X_k)_{k \geq 0}$ is $(\mathcal{F}_k)_{k \geq 0}$ -adapted. Therefore, we only need to check (4.1). Now, for any non-negative or bounded measurable function h on X ,

$$\mathbb{E}[h(X_{k+1}) | \mathcal{F}_k] = \int_{-\infty}^{\infty} h(aX_k + b\epsilon) f(\epsilon) d\epsilon = \int_{-\infty}^{\infty} h(y) f\left(\frac{y - aX_k}{b}\right) \frac{1}{b^p} dy,$$

where the last equality follows from an adequate change of variable. Therefore, $\{X_k : k \in \mathbb{N}\}$ is a Markov chain with Markov kernel

$$(x, A) \mapsto P(x, A) = \int_A f\left(\frac{y - ax}{b}\right) \frac{1}{b^p} dy.$$

In this example, we can check that \mathcal{F}_k is actually the natural filtration of $\{X_k : k \in \mathbb{N}\}$ but we even do not need to check this property for getting that $\{X_k : k \in \mathbb{N}\}$ is a Markov chain.

4.2.1 Additional notation

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

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

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

$$\begin{aligned} \mu(P(Qh)) &= (\mu P)(Qh) = (\mu(PQ))h \\ &= \mu((PQ)h) = \int_{\mathbf{X}^3} \mu(dx) P(x, dy) Q(y, dz) h(z). \end{aligned}$$

Therefore, all these parenthesis can be discarded and we can write μPQh without any ambiguity. To sum up, measures act on the left side of a Markov kernel whereas functions acts on the right side. To make sure you have mastered all the notation, check your understanding with the following equalities $\delta_x P(A) = P(x, A) = P\mathbf{1}_A(x)$.

To finish up with notation, we now define the iterates of a Markov kernel P , which will come in very handy thereafter: for a given Markov kernel P on $\mathbf{X} \times \mathcal{X}$, define $P^0 = I$ where I is the identity kernel: $(x, A) \mapsto \mathbf{1}_A(x)$, and set for $k \geq 0$, $P^{k+1} = P^k P$.

Lemma 4.4. *Let $\{X_k : k \in \mathbb{N}\}$ be a Markov chain on the same probability space $(\Omega, \mathcal{G}, \mathbb{P})$ and taking values on \mathbf{X} , with Markov kernel P and with initial distribution $\nu \in \mathbf{M}_1(\mathbf{X})$. Then, for any $n \in \mathbb{N}$, the law of $X_{0:n}$ is $\nu(dx_0) \prod_{i=0}^{n-1} P(x_i, dx_{i+1})$ (with the convention that $\prod_{i=0}^{-1} = 1$).*

Proof. Recall that for all $(n, A) \in \mathbb{N} \times \mathcal{X}$, $\mathbb{P}(X_{n+1} \in A | X_{0:n}) = P(X_n, A)$, \mathbb{P} -a.s. or equivalently for all non-negative measurable functions h_{n+1} on \mathbf{X} , $\mathbb{E}[h_{n+1}(X_{n+1}) | X_{0:n}] = Ph_{n+1}(X_n)$ \mathbb{P} -a.s. We now show by induction that for all $n \in \mathbb{N}$,

$$(H_n) \text{ the law of } X_{0:n} \text{ is } \nu(dx_0) \prod_{i=0}^{n-1} P(x_i, dx_{i+1}).$$

We first note that (H_0) is true since by assumption, $X_0 \sim \nu$. Assume now that (H_n) holds for some $n \in \mathbb{N}$. Then, for all non-negative measurable functions h_0, \dots, h_{n+1} on \mathbf{X} , the tower property yields

$$\mathbb{E} \left[\prod_{i=0}^{n+1} h_i(X_i) \right] = \mathbb{E} \left[\left(\prod_{i=0}^n h_i(X_i) \right) \mathbb{E}[h_{n+1}(X_{n+1}) | X_{0:n}] \right] = \mathbb{E} \left[\left(\prod_{i=0}^n h_i(X_i) \right) Ph_{n+1}(X_n) \right],$$

and since the inner term in the rhs only depends on $X_{0:n}$, we can apply (H_n) and thus,

$$\begin{aligned} \mathbb{E} \left[\prod_{i=0}^{n+1} h_i(X_i) \right] &= \int_{\mathbf{X}^{n+1}} \left[\nu(dx_0) \prod_{i=0}^{n-1} P(x_i, dx_{i+1}) \right] \left(\prod_{i=0}^n h_i(x_i) \right) Ph_{n+1}(x_n) \\ &= \int_{\mathbf{X}^{n+1}} \left[\nu(dx_0) \prod_{i=0}^n P(x_i, dx_{i+1}) \right] \left(\prod_{i=0}^{n+1} h_i(x_i) \right), \end{aligned}$$

which shows that the law of $X_{0:n+1}$ is $\nu(dx_0) \prod_{i=0}^n P(x_i, dx_{i+1})$ and (H_{n+1}) is thus proved. \square

As a consequence, the marginal law of X_n is given by integrating $\nu(dx_0) \prod_{i=0}^{n-1} P(x_i, dx_{i+1})$ over $x_{0:n-1}$ and thus, for all $A \in \mathcal{X}$,

$$\mathbb{P}(X_n \in A) = \int_{\mathbf{X}^{n+1}} \mathbf{1}_A(X_n) \nu(dx_0) \prod_{i=0}^{n-1} P(x_i, dx_{i+1}) = \nu P^n(A),$$

that is, νP^n is the distribution of X_n or in a compact notation, $X_n \sim \nu P^n$.

4.3 Canonical space

In the previous sections, the $\{X_k : k \in \mathbb{N}\}$ are already given and we check the two items in the definition of a Markov chain (Definition 4.2) with Markov kernel P and initial distribution ν . We now turn to the reverse situation where a couple (ν, P) of initial distribution and Markov kernel are given beforehand and we intend to construct the random variables $\{X_k : k \in \mathbb{N}\}$ on some convenient (common) probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\{X_k : k \in \mathbb{N}\}$ is a Markov chain with Markov kernel P and initial distribution ν .

4.3.1 A simpler problem

Consider an easier problem where we only want to construct $\{X_k : k \in [0 : n-1]\}$ where n is some given positive integer. That is, we only consider a finite range of integers such that the first item in Definition 4.2 is satisfied. For a given $\nu \in \mathbf{M}_1(\mathbf{X})$, define the triplet $(\Omega_n, \mathcal{G}_n, \mathbb{P}_{\nu,n})$ as follows:

- $\Omega_n = \mathbf{X}^{n+1}$, $\mathcal{G}_n = \mathcal{X}^{\otimes(n+1)}$ and $\mathbb{P}_{\nu,n}$ is the probability measure defined on $(\Omega_n, \mathcal{G}_n)$ by

$$\mathcal{G}_n \ni A \mapsto \mathbb{P}_{\nu,n}(A) = \int_{\mathbf{X}^{n+1}} \mathbf{1}_A(\omega_{0:n}) \nu(d\omega_0) \prod_{i=1}^n P(\omega_{i-1}, d\omega_i),$$

and for $\omega \in \Omega_n$, set $X_k(\omega) = \omega_k$, i.e. $X_k(\omega)$ is the projection of the k -th component of ω .

We aim to show that for all $A \in \mathcal{X}$ and all $k \in [0 : n-1]$, we have $\mathbb{P}_{\nu,n}(X_{k+1} \in A | X_{0:k-1}) = P(X_k, A)$, $\mathbb{P}_{\nu,n}$ -a.s. To do so, write for any $k \in [0 : n-1]$, any non-negative measurable function h on \mathbf{X}^{k+1} and any $A \in \mathcal{X}$

$$\begin{aligned} \mathbb{E}_{\nu,n}[h(X_{0:k}) \mathbf{1}_A(X_{k+1})] &= \int_{\mathbf{X}^{k+2}} h(\omega_{0:k}) \mathbf{1}_A(\omega_{k+1}) \nu(d\omega_0) \prod_{i=1}^{k+1} P(\omega_{i-1}, d\omega_i) \\ &= \int_{\mathbf{X}^{k+2}} h(\omega_{0:k}) P(\omega_k, A) \nu(d\omega_0) \prod_{i=1}^k P(\omega_{i-1}, d\omega_i) \\ &= \mathbb{E}_{\nu,n}[h(X_{0:k}) P(X_k, A)]. \end{aligned}$$

Since h is arbitrary, this is equivalent to saying that $\mathbb{P}_{\nu,n}(X_{k+1} \in A | X_{0:k}) = P(X_k, A)$, \mathbb{P} -a.s.

4.3.2 The general case

We now consider the general case where $k \in \mathbb{N}$ instead of $k \in [0 : n-1]$. Define the coordinate process $(X_n)_{n \geq 0}$ by $X_n(\omega) = \omega_n$ for all $\omega \in \mathbf{X}^{\mathbb{N}}$. We will sometimes use $X_{k:\ell} : \omega \mapsto (\omega_k, \dots, \omega_\ell)$ for $k \leq \ell$ and by extension $X_{k:\infty} : \omega \mapsto (\omega_k, \dots, \omega_\ell, \omega_{\ell+1}, \dots)$. In particular, $X_{0:\infty}(\omega) = \text{Id}(\omega) = \omega$ where Id is the identity function.

Theorem 4.5. (The canonical space) *Let $(\mathbf{X}, \mathcal{X})$ be a measurable space and let P be a Markov kernel on $\mathbf{X} \times \mathcal{X}$. For every probability measure $\nu \in \mathbf{M}_1(\mathbf{X})$, there exists a unique probability measure*

\mathbb{P}_ν on the canonical space $(\mathbf{X}^\mathbb{N}, \mathcal{X}^{\otimes \mathbb{N}})$ such that, under \mathbb{P}_ν , the coordinate process $\{X_n : n \in \mathbb{N}\}$ is a Markov chain with Markov kernel P and initial distribution ν .

This result is often referred to as the *Ionescu-Tulcea theorem*. Its proof goes far beyond the scope of this course and we will admit it here. Some other, much simpler proofs exist and are based on the Kolmogorov extension theorem, but they hold at the price of additive assumptions on the space (in contrast, we only assume here that $(\mathbf{X}, \mathcal{X})$ is a measurable space, which is quite minimal). In the canonical representation, we therefore set $\Omega = \mathbf{X}^\mathbb{N}$, $\mathcal{G} = \mathcal{X}^{\otimes \mathbb{N}}$ and $\mathbb{P} = \mathbb{P}_\nu$. In the particular case where the initial distribution ν is a Dirac mass, we use the compact notation $\mathbb{P}_x = \mathbb{P}_{\delta_x}$. Thus, Theorem 4.5 allows to define not only one probability measure but a family of probability measures $(\mathbb{P}_\nu)_{\nu \in \mathbf{M}_1(\mathbf{X})}$ on the space of trajectories.

What are the relations between the probability measures $(\mathbb{P}_\nu)_{\nu \in \mathbf{M}_1(\mathbf{X})}$? A consequence of this theorem is that for all $A \in \mathcal{X}^{\otimes(n+1)}$, $\mathbb{P}_\nu(X_{0:n} \in A) = \int_A \nu(d\omega_0) \prod_{i=1}^n P(\omega_{i-1}, d\omega_i)$. Replacing ν by δ_{x_0} and comparing the two obtained expressions, we get

$$\mathbb{P}_\nu(X_{0:n} \in A) = \int_{\mathbf{X}} \nu(dx_0) \mathbb{P}_{x_0}(X_{0:n} \in A).$$

We can actually extend this result to any $A \in \mathcal{X}^{\otimes \mathbb{N}}$ by replacing the $n+1$ -tuple $X_{0:n}$ by the (infinite) trajectory $X_{0:\infty}$. First note the following equalities: $A = \{\omega \in A\} = \{\omega \in \Omega : X_{0:\infty}(\omega) \in A\} = \{X_{0:\infty} \in A\}$. We then obtain the following identity: for all $A \in \mathcal{X}^{\otimes \mathbb{N}}$,

$$\mathbb{P}_\nu(A) = \mathbb{P}_\nu(X_{0:\infty} \in A) = \int_{\mathbf{X}} \nu(dx_0) \mathbb{P}_{x_0}(X_{0:\infty} \in A) = \int_{\mathbf{X}} \nu(dx_0) \mathbb{P}_{x_0}(A). \quad (4.2)$$

An illustration is given in Exercise 9.2.

4.3.3 The Markov property.

Define the shift operator S by $S : \mathbf{X}^\mathbb{N} \ni \omega \mapsto \omega' \in \mathbf{X}^\mathbb{N}$ where $\omega = (\omega_i)_{i \in \mathbb{N}}$ and $\omega' = (\omega_{i+1})_{i \in \mathbb{N}}$.

Theorem 4.6. (The Markov property) For any $\nu \in \mathbf{M}_1(\mathbf{X})$, any non-negative or bounded function h on $\mathbf{X}^\mathbb{N}$ and any $n \in \mathbb{N}$,

$$\mathbb{E}_\nu [h \circ S^k | \mathcal{F}_k] = \mathbb{E}_{X_k} [h], \quad \mathbb{P}_\nu - a.s. \quad (4.3)$$

where $\mathcal{F}_k = \sigma(X_{0:k})$.

By definition of $X_{k:\infty}$, we have $\mathbb{P}_\nu - a.s.$,

$$\begin{aligned} \mathbb{E}_\nu [h(X_{k:\infty}) | \mathcal{F}_k] &= \mathbb{E}_\nu [h \circ S^k \circ X_{0:\infty} | \mathcal{F}_k] = \mathbb{E}_\nu [h \circ S^k \circ \text{Id} | \mathcal{F}_k] \\ &= \mathbb{E}_\nu [h \circ S^k | \mathcal{F}_k] = \mathbb{E}_{X_k} [h] = \mathbb{E}_{X_k} [h \circ \text{Id}] \\ &= \mathbb{E}_{X_k} [h(X_{0:\infty})]. \end{aligned}$$

Therefore, for any $\nu \in \mathbf{M}_1(\mathbf{X})$,

$$\boxed{\mathbb{E}_\nu [h(X_{k:\infty}) | \mathcal{F}_k] = \mathbb{E}_{X_k} [h(X_{0:\infty})], \quad \mathbb{P}_\nu - a.s.}$$

is another equivalent expression of the Markov property. This expression may seem easier to deal with but the reader has to be at ease with both formulations. A stronger version of the Markov property exists and is called (as expected) the strong Markov property: its statement is (4.3) with the exception that k is replaced by a stopping time τ and that the identity only holds on the event $\{\tau < \infty\}$. The strong Markov property is extremely important in Markov Chain theory but,

quite surprisingly, it is not needed in this course. Exercise 9.3 states and proves the Strong Markov property.

Chapter 5

Metropolis-Hastings algorithms

Contents

5.1 Invariant probability measures: existence	35
5.1.1 Metropolis-Hastings (MH) algorithms	36
5.2 Invariant probability measure: uniqueness	38
5.2.1 Application to Metropolis-Hastings algorithms.	40

5.1 Invariant probability measures: existence

Definition 5.1. We say that $\pi \in M_1(X)$ is an invariant probability measure for the Markov kernel P on $X \times \mathcal{X}$ 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 $k \in \mathbb{N}$, $\pi P^k = \pi$. This result on the (marginal) distribution of X_k may be extended to n -tuples.

More precisely, 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* under \mathbb{P}_π (in the sense that for all $n, p \in \mathbb{N}^*$, and all n -tuple $k_{1:n}$, the random vector $(X_{k_1}, \dots, X_{k_n})$ follows the same distribution as $(X_{k_1+p}, \dots, X_{k_n+p})$).

Exercises 9.4 and 9.5 illustrate the existence of stationary distributions for Markov chains. We now introduce the notion of reversibility for a Markov kernel. This will be of crucial importance for designing Markov kernels with a given invariant probability measure.

Definition 5.2. Let $\pi \in M_1(X)$ and P be a Markov kernel on $X \times \mathcal{X}$. We say that P is π -reversible if and only if (with infinitesimal notation)

$$\pi(dx)P(x, dy) = \pi(dy)P(y, dx), \quad (5.1)$$

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

$$\iint_{X^2} h(x, y) \pi(dx) P(x, dy) = \iint_{X^2} h(x, y) \pi(dy) P(y, dx). \quad (5.2)$$

A Markov kernel P is π -reversible if and only if the probability measure $\pi(dx)P(x, dy)$ is symmetric with respect to (x, y) .

Proposition 5.3. *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 $A \in \mathcal{X}$, we have by the reversibility relation

$$\pi P(A) = \iint_{\mathsf{X}^2} \mathbf{1}_A(y) \pi(dx) P(x, dy) = \iint_{\mathsf{X}^2} \mathbf{1}_A(y) \pi(dy) P(y, dx) = \int_A \pi(dy) \underbrace{P(y, \mathsf{X})}_1 = \pi(A),$$

which finishes the proof. \square

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

Exercise 9.6 gives an example of π -reversible kernel.

5.1.1 Metropolis-Hastings (MH) algorithms

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 $\pi(h) = \int_{\mathsf{X}} h(x) \pi(dx)$ of the form $n^{-1} \sum_{k=0}^{n-1} h(X_k)$ for any measurable function h .

5.1.1.1 Construction of the kernel

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 $\pi(dx) = \pi(x)\lambda(dx)$ and we assume that this density π is positive.

Moreover, let Q be Markov kernel on $\mathsf{X} \times \mathcal{X}$ such that $Q(x, dy) = q(x, y)\lambda(dy)$, 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 $\alpha : \mathsf{X}^2 \rightarrow [0, 1]$, Algorithm 5 describes the Metropolis algorithm.

Algorithm 5 The Metropolis Algorithm.

At $t = 0$, draw X_0 according to some arbitrary distribution.
for $t = 0$ to $t = n - 1$ **do**
 Draw independently $Y_{t+1} \sim Q(X_t, \cdot)$ and $U_{t+1} \sim \text{Unif}(0, 1)$.
 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 for

In words, 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 refused 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 . To do so, let us assume that for all $x, y \in \mathsf{X}$,

$$\pi(x)\alpha(x, y)q(x, y) = \pi(y)\alpha(y, x)q(y, x), \quad (5.3)$$

and let us show that it implies that the Markov kernel P associated with $(X_k)_{k \in \mathbb{N}}$ is π -reversible.

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) (which is equivalent to $\sigma(X_{0:t}) \subset \mathcal{F}_t$). Then, setting $\bar{\alpha}(x) = 1 - \int_{\mathbf{X}} Q(x, dy) \alpha(x, y)$, we have for any bounded or non-negative measurable function h on \mathbf{X} and any $t \in \mathbb{N}$,

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

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

$$P_{\langle \pi, Q \rangle}^{MH}(x, dy) = Q(x, dy) \alpha(x, y) + \bar{\alpha}(x) \delta_x(dy). \quad (5.4)$$

Lemma 5.4. *The Markov kernel $P_{\langle \pi, Q \rangle}^{MH}$ is π -reversible if and only if*

$$\pi(dx) Q(x, dy) \alpha(x, y) = \pi(dy) Q(y, dx) \alpha(y, x). \quad (5.5)$$

Equation (5.5) is often called the detailed balance condition.

Proof. First, note that

$$\pi(dx) \bar{\alpha}(x) \delta_x(dy) = \pi(dy) \bar{\alpha}(y) \delta_y(dx). \quad (5.6)$$

Indeed, for any measurable function h on \mathbf{X}^2 , we have

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

Combining (5.4) with (5.6), we obtain that $P_{\langle \pi, Q \rangle}^{MH}$ is π -reversible if and only if the detailed balance condition (5.5) is satisfied. This completes the proof. \square

5.1.1.2 Acceptance probability

We now make use of Lemma 5.4 in order to find an explicit expression of the acceptance probability α .

Lemma 5.5. *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, any $\alpha \in \{\alpha^{MH}, \alpha^b\}$ satisfies the detailed balance condition (5.5). Moreover, any other $\alpha \in [0, 1]$ that satisfies the detailed balance condition is dominated by α^{MH} in the sense that: for $\lambda^{\otimes 2}$ -almost all $x, y \in \mathbf{X}$,

$$\alpha(x, y) \leq \alpha^{MH}(x, y). \quad (5.7)$$

Proof. The fact that any $\alpha \in \{\alpha^{MH}, \alpha^b\}$ satisfies $\pi(x)q(x, y)\alpha(x, y) = \pi(y)q(y, x)\alpha(y, x)$ for $\lambda^{\otimes 2}$ -almost all $x, y \in \mathbf{X}$, is immediate, by plugging the value of α in the equation. It remains to check (5.7). Assume now that for $\lambda^{\otimes 2}$ -almost all $x, y \in \mathbf{X}$,

$$\pi(x)q(x, y)\alpha(x, y) = \pi(y)q(y, x)\alpha(y, x).$$

then, using that $\alpha(y, x) \leq 1$ shows that $\alpha(x, y) \leq \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}$. Moreover, $\alpha(x, y) \leq 1$ and this finally implies

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

which completes the proof. \square

According to (5.7), α^{MH} is actually the highest acceptance probability among the acceptance probabilities such that $P_{\langle \pi, Q \rangle}^{MH}$ is π -reversible and therefore, this acceptance is widely used in practice (in the sense that we expect that a Markov kernel that accepts often, explores the space more rapidly and therefore is preferable to another one with less acceptance probability). In what follows, unless otherwise stated, we implicitly assume that the Markov kernel $P_{\langle \pi, Q \rangle}^{MH}$ is associated with the acceptance probability α^{MH} .

Exercise 9.7 shows other examples of possible acceptance probabilities.

Example 5.6. (The independence sampler) If the proposition kernel is $Q(x, dy) = q(y)\lambda(dy)$ where q is a density with respect to λ on \mathbf{X} , then at each time step, the proposed candidate is drawn irrespective of the current value of the Markov chain (this is because, $Q(x, dy)$ does not depend on x), that is, in the step 2(a) of Algorithm 5, we draw $Y_{t+1} \sim q(\cdot)$. In such a case, the acceptance probability is

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

and the Metropolis-Hastings algorithm is called the *Independence Sampler*.

Example 5.7. (The random walk MH sampler) If $\mathbf{X} = \mathbb{R}^p$ and if the proposition kernel is $Q(x, dy) = q(y-x)\lambda(dy)$ where q is a symmetric density with respect to λ on \mathbf{X} , (by symmetric, we mean that $q(u) = q(-u)$ for all $u \in \mathbf{X}$) then at each time step in Algorithm 5, we draw a candidate $Y_{t+1} \sim q(y - X_t)\lambda(dy)$. In such a case, the acceptance probability is $\alpha(x, y) = \min(\pi(y)/\pi(x), 1)$ and the associated algorithm is called the *(symmetric) Random Walk Metropolis-Hasting*. Another way of writing the proposition update is $Y_{k+1} = X_k + \eta_k$ where $\eta_k \sim q(\cdot)$.

5.2 Invariant probability measure: uniqueness

We start with a very simple lemma that will be useful for finding sufficient conditions for uniqueness.

Lemma 5.8. *If P admits two distinct invariant probability measures, it also admits distinct invariant probability measures π_0 and π_1 that are mutually singular, i.e., such that there exists $A \in \mathcal{X}$ such that $\pi_0(A) = \pi_1(A^c) = 0$.*

Proof. Let ζ_0, ζ_1 be two distinct invariant probability measures for P . Both have densities with respect to some common dominating measure (for example, choosing $\zeta = \zeta_1 + \zeta_2$, we have that ζ dominates both ζ_0 and ζ_1 , since $\zeta(A) = 0 \Rightarrow (\zeta_1(A) = 0 \text{ and } \zeta_2(A) = 0)$ for any $A \in \mathcal{X}$ and according to the Radon Nikodym theorem, if a measure dominates another one, the latter has a density with respect to the former). Write then $\zeta_0(dx) = f_0(x)\zeta(dx)$ and $\zeta_1(dx) = f_1(x)\zeta(dx)$ where f_0, f_1 are non-negative measurable functions on \mathbf{X} . Define the positive part $(\zeta_1 - \zeta_0)^+$ and the negative part $(\zeta_1 - \zeta_0)^-$ of the signed measure $\zeta_1 - \zeta_0$ by $(\zeta_1 - \zeta_0)^+(dx) = [f_1(x) - f_0(x)]^+\zeta(dx)$ and $(\zeta_1 - \zeta_0)^-(dx) = [f_1(x) - f_0(x)]^-\zeta(dx)$. Then,

$$\begin{aligned}
(\zeta_1 - \zeta_0)^+ P \mathbf{1}_A &= \int_{\mathbf{X}} \zeta(dx) [f_1(x) - f_0(x)]^+ P(x, A) \\
&\geq \int_{\mathbf{X}} \zeta(dx) [f_1(x) - f_0(x)] P(x, A) \\
&\geq \zeta_1 P(A) - \zeta_0 P(A) = \zeta_1(A) - \zeta_0(A).
\end{aligned}$$

Therefore, $(\zeta_1 - \zeta_0)^+ P$ is a (non-negative) measure that is greater than the signed measure $\zeta_1 - \zeta_0$. Since the positive part $(\zeta_1 - \zeta_0)^+$ is also the smallest (non-negative) measure that is greater than $\zeta_1 - \zeta_0$, we conclude that $(\zeta_1 - \zeta_0)^+ \leq (\zeta_1 - \zeta_0)^+ P$. The measure $(\zeta_1 - \zeta_0)^+ P - (\zeta_1 - \zeta_0)^+$ is therefore non-negative and we have

$$[(\zeta_1 - \zeta_0)^+ P - (\zeta_1 - \zeta_0)^+](\mathbf{X}) = \int_{\mathbf{X}} (\zeta_1 - \zeta_0)^+(dx) \underbrace{P(x, \mathbf{X})}_1 - (\zeta_1 - \zeta_0)^+(\mathbf{X}) = 0.$$

Finally, $(\zeta_1 - \zeta_0)^+ = (\zeta_1 - \zeta_0)^+ P$. The probability measure $\pi_0 = (\zeta_1 - \zeta_0)^+ / (\zeta_1 - \zeta_0)^+(\mathbf{X})$ is thus an invariant probability measure for P . Replacing $(\zeta_1 - \zeta_0)^+$ by $(\zeta_1 - \zeta_0)^-$, we obtain in the same way that $\pi_1 = (\zeta_1 - \zeta_0)^- / (\zeta_1 - \zeta_0)^-(\mathbf{X})$ is an invariant probability measure. We can easily check that taking $A = \{f_0 \geq f_1\}$, we have $\pi_0(A) = \pi_1(A^c) = 0$, showing that these probability measures are mutually singular. \square

In the course of the proof, we actually need some results on the positive and negative part of a signed-measure. The following exercise is useful to fully understand the previous proof.

Exercise 5.9. Define $\mathbf{M}_s(\mathbf{X})$ the set of signed-measures. Let $\mu \in \mathbf{M}_s(\mathbf{X})$ and assume that $\mu \preceq \zeta$ where $\zeta \in \mathbf{M}_+(\mathbf{X})$ (in the sense that we have the implication: if for some $A \in \mathcal{X}$, $\zeta(A) = 0$, then $\mu(A) = 0$). According to the Radon-Nikodym theorem, there exists a measurable function h such that $\mu(dx) = h(x)\zeta(dx)$. Define $\mu^+(dx) = |h(x)|\zeta(dx)$.

1. Show that the measure μ^+ is well-defined (in the sense that the measure $|h(x)|\zeta(dx)$ does not depend on the measure ζ , provided that the ζ dominates μ .)
2. Show that for ζ -almost all $x \in \mathbf{X}$, $|h(x)| \leq 1$.
3. Assume that there exists $\nu \in \mathbf{M}_+(\mathbf{X})$ such that for all $A \in \mathcal{X}$, we have $\mu(A) \leq \nu(A)$. Show that $\mu^+(A) \leq \nu(A)$ for all $A \in \mathcal{X}$.

We now make use of Lemma 5.8 in order to give a sufficient condition for uniqueness.

Proposition 5.10. *Assume that there exists a non-null measure $\mu \in \mathbf{M}_+(\mathbf{X})$ satisfying the following property:*

- *For all $A \in \mathcal{X}$ such that $\mu(A) > 0$ and for all $x \in \mathbf{X}$, there exists $n \in \mathbb{N}$ such that $P^n(x, A) > 0$.*

Then, P admits at most one invariant probability measure.

If the assumption of Proposition 5.10 holds, we say that P is μ -irreducible and in such case, μ is called an irreducibility measure for P .

Proof. The proof is by contradiction. Assume that there exist two distinct invariant probability measures. According to Lemma 5.8, we can consider two invariant probability measures π_1 and π_2 that are mutually singular. Under the assumptions of Proposition 5.10, let $A \in \mathcal{X}$ such that $\mu(A) > 0$. Then, for any $i \in \{1, 2\}$, we have

$$0 < \int_{\mathbf{X}} \pi_i(dx) \underbrace{\sum_{n=0}^{\infty} P^n(x, A)}_{>0} = \sum_{n=0}^{\infty} \pi_i P^n(A) = \sum_{n=0}^{\infty} \pi_i(A),$$

which in turn implies that $\pi_i(A) > 0$. The contraposed implication gives that if for some $i \in \{1, 2\}$, $\pi_i(A) = 0$, then $\mu(A) = 0$. Now, since π_1 and π_2 are mutually singular, there exists $A \in \mathcal{X}$ such that $\pi_1(A) = \pi_2(A^c) = 0$ and this shows that $\mu(A) = \mu(A^c) = 0$ which is impossible. \square

Exercise 9.9 gives an example where Proposition 5.10 applies.

5.2.1 Application to Metropolis-Hastings algorithms.

We have already seen that $P_{\langle\pi, Q\rangle}^{MH}$ is π -invariant and we have assumed that $Q(x, dy) = q(x, y)\lambda(dy)$ and $\pi(dy) = \pi(y)\lambda(dy)$ and for simplicity, that for all $x, y \in \mathbf{X}$, $q(x, y) > 0$ and $\pi(y) > 0$. This in turn implies that $\alpha(x, y) = \alpha^{MH}(x, y) > 0$ and therefore if $\lambda(A) > 0$, then for all $x \in \mathbf{X}$,

$$P(x, A) \geq \int_A q(x, y)\alpha(x, y)\lambda(dy) > 0.$$

This shows that P is λ -irreducible and therefore π is the unique invariant probability measure for P .

Chapter 6

Ergodicity and Law of Large numbers

Contents

6.1	Dynamical systems.	41
6.2	Markov chains and ergodicity	42

We now focus on properties of Markov chains with unique invariant probability measures. We show in this chapter that such Markov chains turn out to be ergodic in some sense to be defined and this, in turn, allows to apply the Birkhoff ergodic theorem and use the law of large numbers.

6.1 Dynamical systems.

Definition 6.1. (Dynamical system) A dynamical system \mathcal{D} is a quadruplet $\mathcal{D} = (\Omega, \mathcal{F}, \mathbb{P}, T)$ where $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $T : \Omega \rightarrow \Omega$ is a measurable mapping such that $\mathbb{P} = \mathbb{P} \circ T^{-1}$.

Lemma 6.2. The collection of sets $\mathcal{I} = \{A \in \mathcal{F} : \mathbf{1}_A = \mathbf{1}_A \circ T\}$ is a σ -field and any set in \mathcal{I} is called an invariant set.

Proof. Indeed, obviously $\Omega \in \mathcal{I}$. Moreover, if $A \in \mathcal{I}$, then for all $\omega \in \Omega$,

$$\mathbf{1}_{A^c}(\omega) = 1 - \mathbf{1}_A(\omega) = 1 - \mathbf{1}_A \circ T(\omega) = \mathbf{1}_{A^c} \circ T(\omega),$$

showing that $A^c \in \mathcal{I}$. Now, consider a countable family of $A_i \in \mathcal{I}$ where $i \in \mathbb{N}$. Then $\omega \in \bigcap_{i \in \mathbb{N}} A_i$ if and only if for all $i \in \mathbb{N}$, $\omega \in A_i$ which is in turn equivalent to $T(\omega) \in A_i$. All in all we have shown that $\omega \in \bigcap_{i \in \mathbb{N}} A_i$ if and only if $T(\omega) \in \bigcap_{i \in \mathbb{N}} A_i$. This shows that \mathcal{I} is a σ -field. \square

Definition 6.3. (Ergodicity) A dynamical system $(\Omega, \mathcal{F}, \mathbb{P}, T)$ is said to be *ergodic* if \mathcal{I} is trivial, that is, $A \in \mathcal{I}$ implies that either $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$.

What interests us in the first place is that ergodic dynamical system satisfies the Birkhoff ergodic theorem, as detailed below. We first define the iterates of T by $T^0 = I$, and for all $k \geq 1$, $T^k = T^{k-1} \circ T$.

Theorem 6.4. (The Birkhoff theorem) Let $\mathcal{D} = (\Omega, \mathcal{F}, \mathbb{P}, T)$ be an ergodic dynamical system and let $h \in \mathcal{L}_1(\Omega)$. Then,

$$\lim_{n \rightarrow \infty} n^{-1} \sum_{k=0}^{n-1} h \circ T^k = \mathbb{E}[h], \quad \mathbb{P} - a.s.$$

Exercise 9.11 proves this theorem.

6.2 Markov chains and ergodicity

Let us now relate Markov chains to dynamical systems. Recall the shift operator $S : \mathbf{X}^{\mathbb{N}} \ni \omega \mapsto \omega' \in \mathbf{X}^{\mathbb{N}}$ where $\omega = (\omega_i)_{i \in \mathbb{N}}$ and $\omega' = (\omega_{i+1})_{i \in \mathbb{N}}$. It is important to note that in general $(\mathbf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes \mathbb{N}}, \mathbb{P}_\nu, S)$ is not a dynamical system except if the initial distribution ν is actually invariant wrt P .

Lemma 6.5. *Let P be a Markov kernel admitting an invariant probability measure π . Then, the quadruplet $(\mathbf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes \mathbb{N}}, \mathbb{P}_\pi, S)$ is a dynamical system.*

Proof. Indeed, the relation $\mathbb{P}_\pi = \mathbb{P}_\pi \circ S^{-1}$ is equivalent to the fact that for any $A \in \mathcal{X}^{\otimes \mathbb{N}}$, $\mathbb{P}_\pi(A) = \mathbb{P}_\pi(X_{0:\infty} \in A) = \mathbb{P}_\pi(X_{0:\infty} \in S^{-1}(A)) = \mathbb{P}_\pi(X_{1:\infty} \in A)$, which is a consequence of the fact that the sequence of random variables $\{X_k : k \in \mathbb{N}\}$ is strongly stationary under \mathbb{P}_π . \square

The relation $\mathbb{P}_\pi = \mathbb{P}_\pi \circ S^{-1}$ tells us that for any $A \in \mathcal{X}^{\otimes \mathbb{N}}$, we have

$$\mathbb{E}_\pi[\mathbf{1}_A] = \mathbb{P}_\pi(A) = \mathbb{P}_\pi \circ S^{-1}(A) = \mathbb{P}_\pi(S^{-1}(A)) = \mathbb{E}_\pi[\mathbf{1}_{S^{-1}(A)}] = \mathbb{E}_\pi[\mathbf{1}_A \circ S],$$

This, in turn, implies that for any $h \in \mathbf{F}_+(\mathbf{X})$,

$$\mathbb{E}_\pi[h] = \mathbb{E}_\pi[h \circ S].$$

We now provide conditions under which a Markov kernel induces an ergodic dynamical system $(\mathbf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes \mathbb{N}}, \mathbb{P}_\pi, S)$. In what follows, $\mathcal{F}_k = \sigma(X_{0:k})$. Since invariant sets A belong to $\mathcal{X}^{\otimes \mathbb{N}}$, it can be (and it will be) useful to get approximations of A by sets in \mathcal{F}_k where k is conveniently chosen.

Lemma 6.6. (The approximation lemma) *Any set $A \in \mathcal{X}^{\otimes \mathbb{N}}$ satisfies the following approximation property:*

- for all $\delta > 0$, there exist $k \in \mathbb{N}$ and $B \in \mathcal{F}_k$ such that

$$\mathbb{E}_\pi[|\mathbf{1}_A - \mathbf{1}_B|] \leq \delta. \quad (6.1)$$

Proof. This is a typical use of the monotone class theorem. Consider the class \mathcal{M} of sets $A \in \mathcal{X}^{\otimes \mathbb{N}}$, for which the approximation (6.1) holds.

- If $A_0, A_1 \in \mathcal{M}$ and $A_0 \subset A_1$, then $A_1 \setminus A_0 \in \mathcal{M}$. This is actually immediate from the following identities, valid for all sets A_0, A_1, B_0, B_1 ,

$$\mathbf{1}_{A_1 \setminus A_0} - \mathbf{1}_{B_1 \setminus B_0} = \mathbf{1}_{A_1} \mathbf{1}_{A_0^c} - \mathbf{1}_{B_1} \mathbf{1}_{B_0^c} = \mathbf{1}_{A_1} (\mathbf{1}_{A_0^c} - \mathbf{1}_{B_0^c}) + (\mathbf{1}_{A_1} - \mathbf{1}_{B_1}) \mathbf{1}_{B_0^c},$$

which implies $\mathbb{E}[|\mathbf{1}_{A_1 \setminus A_0} - \mathbf{1}_{B_1 \setminus B_0}|] \leq \mathbb{E}[|\mathbf{1}_{A_0} - \mathbf{1}_{B_0}|] + \mathbb{E}[|\mathbf{1}_{A_1} - \mathbf{1}_{B_1}|]$

- If $A_n \uparrow A$ where $A_n \in \mathcal{M}$ and $A_n \subset A_{n+1}$ for all $n \geq 0$. Then, setting $A = \cup_n A_n \in \mathcal{M}$, we have

$$\lim_{n \rightarrow \infty} \mathbf{1}_{A_n} = \mathbf{1}_A,$$

and this immediately implies that $A \in \mathcal{M}$ (check it carefully).

Then, \mathcal{M} is a monotone class that contains all the $(\mathcal{F}_k)_{k \geq 0}$ and therefore, it contains $\sigma(\cup_{k=0}^\infty \mathcal{F}_k) = \mathcal{X}^{\otimes \mathbb{N}}$, which completes the proof. \square

Theorem 6.7. *Let P be a Markov kernel on $\mathbf{X} \times \mathcal{X}$. Assume that P admits a unique invariant probability measure π . Then, the dynamical system $(\mathbf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes \mathbb{N}}, \mathbb{P}_\pi, S)$ is ergodic.*

Proof. Let P be a Markov kernel that admits a unique invariant probability measure π and let $A \in \mathcal{I} = \{A \in \mathcal{X}^{\otimes \mathbb{N}} : \mathbf{1}_A = \mathbf{1}_A \circ S\}$.

Assume that $\mathbb{P}_\pi(A) > 0$. We will show that $\mathbb{P}_\pi(A) = 1$ and this will prove that the dynamical system $(\mathbf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes \mathbb{N}}, \mathbb{P}_\pi, S)$ is ergodic.

Before diving into the proof, let us take a few minutes to analyse the situation... The quantity of interest is $\mathbb{P}_\pi(A)$ while the assumption is on π (it is the unique invariant probability measure for P). A first step is to relate $\mathbb{P}_\pi(A)$ with π ... That reminds us (4.2), which allows to write

$$\mathbb{P}_\pi(A) = \int_{\mathbf{X}} \pi(dx) \mathbb{E}_x[\mathbf{1}_A] = \pi(h_A), \quad \text{where we have set } h_A(x) = \mathbb{E}_x[\mathbf{1}_A].$$

The proof proceeds in two steps. We first establish some results on h_A and then, deduce properties on $\mathbb{P}_\pi(A)$ by constructing, from π and A , another invariant probability measure π_A .

(i) ($h_A(X_n)$ **does not depend on n** , $\mathbb{P}_\pi - a.s.$). To see this, first write for any $n \in \mathbb{N}$,

$$\begin{aligned} \mathbb{E}_\pi[|h_A(X_0) - \mathbf{1}_A|] &= \mathbb{E}_\pi[|h_A(X_0) - \mathbf{1}_A| \circ S^n] \\ &= \mathbb{E}_\pi[|h_A(\underbrace{X_0 \circ S^n}_{X_n}) - \underbrace{\mathbf{1}_A \circ S^n}_{\mathbf{1}_A}|] = \mathbb{E}_\pi[|h_A(X_n) - \mathbf{1}_A|]. \end{aligned} \quad (6.2)$$

We now show that the rhs tends to 0. First, we find another expression for $h_A(X_n)$. Define $\mathcal{F}_n = \sigma(X_0, \dots, X_n)$. Then, $\mathbb{P}_\pi - a.s.$,

$$h_A(X_n) = \mathbb{E}_{X_n}[\mathbf{1}_A] \stackrel{(1)}{=} \mathbb{E}_\pi[\mathbf{1}_A \circ S^n | \mathcal{F}_n] \stackrel{(2)}{=} \mathbb{E}_\pi[\mathbf{1}_A | \mathcal{F}_n],$$

where $\stackrel{(1)}{=}$ comes from the Markov property and $\stackrel{(2)}{=}$ from the fact that $A \in \mathcal{I}$. Fix some $k \in \mathbb{N}$ and let $B \in \mathcal{F}_k$. Then, for all $n \geq k$, we have $\mathbb{E}[\mathbf{1}_B | \mathcal{F}_n] = \mathbf{1}_B$ and thus,

$$\mathbb{E}_\pi[|h_A(X_n) - \mathbf{1}_B|] = \mathbb{E}_\pi[|\mathbb{E}_\pi[\mathbf{1}_A - \mathbf{1}_B | \mathcal{F}_n]|] \leq \mathbb{E}_\pi[\mathbb{E}_\pi[|\mathbf{1}_A - \mathbf{1}_B| | \mathcal{F}_n]] = \mathbb{E}_\pi[|\mathbf{1}_B - \mathbf{1}_A|]$$

This implies, by using the triangular inequality and then taking the limsup,

$$\limsup_{n \rightarrow \infty} \mathbb{E}_\pi[|h_A(X_n) - \mathbf{1}_A|] \leq 2\mathbb{E}_\pi[|\mathbf{1}_B - \mathbf{1}_A|].$$

Now, according to the approximation lemma (Lemma 6.6), the rhs can be made arbitrarily small for a convenient choice of k and $B \in \mathcal{F}_k$. Therefore, $\lim_{n \rightarrow \infty} \mathbb{E}_\pi[|h_A(X_n) - \mathbf{1}_A|] = 0$. Combining this limiting result with (6.2), we deduce that $\mathbb{E}_\pi[|h_A(X_n) - \mathbf{1}_A|]$ is a constant that tends to 0 as n tends to infinity. It is thus equal to 0 for all $n \in \mathbb{N}$, and we have

$$h_A(X_0) = h_A(X_n) = \mathbf{1}_A, \quad \mathbb{P}_\pi - a.s. \quad (6.3)$$

(ii) Now, define the probability measure π_A on $(\mathbf{X}, \mathcal{X})$ by $\pi_A(f) = \frac{\mathbb{E}_\pi[h_A(X_0)f(X_0)]}{\mathbb{P}_\pi(A)}$ for any non-negative measurable function f on \mathbf{X} . Then, using the Markov property, (6.3) with $n = 1$ and $X_1 \stackrel{\mathcal{L}}{=} X_0$ under \mathbb{P}_π (that is, they share the same distribution under \mathbb{P}_π), we get

$$\begin{aligned} \mathbb{E}_\pi[h_A(X_0) \times Pf(X_0)] &= \mathbb{E}_\pi[h_A(X_0) \times f(X_1)] = \mathbb{E}_\pi[h_A(X_1) \times f(X_1)] \\ &= \mathbb{E}_\pi[h_A(X_0) \times f(X_0)], \end{aligned}$$

Therefore, $\pi_A P(f) = \pi_A(Pf) = \pi_A(f)$, showing that π_A is an invariant probability measure for P and thus, $\pi = \pi_A$. Then,

$$\mathbb{P}_\pi(A) = \int_{\mathbf{X}} \pi(dx) \mathbb{P}_x(A) = \pi(h_A) = \pi_A(h_A) = \frac{\mathbb{E}_\pi[h_A(X_0) \times h_A(X_0)]}{\mathbb{P}_\pi(A)}.$$

Applying again (6.3) yields $\mathbb{P}_\pi(A) = \mathbb{E}_\pi[\mathbf{1}_A \times \mathbf{1}_A] / \mathbb{P}_\pi(A) = 1$ which completes the proof of the theorem. \square

As a consequence, the Birkhoff theorem for dynamical systems, Theorem 6.4, yields the following result.

Theorem 6.8. *Let P be a Markov kernel admitting a unique invariant probability measure π . Then, for all $h \in F(X^{\mathbb{N}})$ such that $\mathbb{E}_{\pi}[|h|] < \infty$, we have*

$$\lim_{n \rightarrow \infty} n^{-1} \sum_{k=0}^{n-1} h(X_{k:\infty}) = \mathbb{E}_{\pi}[h], \quad \mathbb{P}_{\pi} - a.s.$$

or equivalently,

$$\lim_{n \rightarrow \infty} n^{-1} \sum_{k=0}^{n-1} h \circ S^k = \mathbb{E}_{\pi}[h], \quad \mathbb{P}_{\pi} - a.s.$$

A particular case of Theorem 6.8 is when $h(X_{0:\infty}) = f(X_0)$. In such case, we have the following corollary.

Corollary 6.9. *Let P be a Markov kernel admitting a unique invariant probability measure π . Then, for all $f \in F(X)$ such that $\pi(|f|) = \int_X \pi(dx) |f(x)| < \infty$, we have*

$$\lim_{n \rightarrow \infty} n^{-1} \sum_{k=0}^{n-1} f(X_k) = \pi(f), \quad \mathbb{P}_{\pi} - a.s. \quad (6.4)$$

The limiting result (6.4) is nice but it holds $\mathbb{P}_{\pi} - a.s.$ We now try to overcome this issue. Under the assumptions of Corollary 6.9, set $A = \left\{ \lim_{n \rightarrow \infty} n^{-1} \sum_{k=0}^{n-1} f(X_k) = \pi(f) \right\}$. Combining Theorem 6.8 with (4.2), $0 = \mathbb{P}_{\pi}(A^c) = \int_X \pi(dx) \mathbb{P}_x(A^c)$ and this implies that $\mathbb{P}_x(A^c) = 0$ (i.e. $\mathbb{P}_x(A) = 1$) for π -almost all $x \in X$.

Corollary 6.10. *Let P be a Markov kernel admitting a unique invariant probability measure π . Then, for all $f \in F(X)$ such that $\pi(|f|) = \int_X \pi(dx) |f(x)| < \infty$, we have for π -almost all $x \in X$,*

$$\lim_{n \rightarrow \infty} n^{-1} \sum_{k=0}^{n-1} f(X_k) = \pi(f), \quad \mathbb{P}_x - a.s. \quad (6.5)$$

What about Metropolis-Hastings algorithms? In Section 5.2.1, we have seen that π is the unique invariant probability measure for $P_{\langle \pi, Q \rangle}^{MH}$ provided that $Q(x, dy) = q(x, y) \lambda(dy)$ and $\pi(dy) = \pi(y) \lambda(dy)$ with $q > 0$ and $\pi > 0$. Therefore, (6.4) and (6.5) hold.

Theorem 6.11. *Let Q be a Markov kernel on $X \times X$ and $\pi \in M_1(X)$. Assume that $Q(x, dy) = q(x, y) \lambda(dy)$ and $\pi(dy) = \pi(y) \lambda(dy)$ where $q > 0$, $\pi > 0$ and λ is a σ -finite measure on (X, \mathcal{X}) . Then, the Markov chain $\{X_n : n \in \mathbb{N}\}$ with Markov kernel $P_{\langle \pi, Q \rangle}^{MH}$, i.e. the Markov chain generated by the Metropolis-Hastings algorithm is such that: for all initial distributions $\nu \in M_1(X)$ and all $f \in F(X)$ such that $\pi(|f|) = \int_X \pi(dx) |f(x)| < \infty$,*

$$\lim_{n \rightarrow \infty} n^{-1} \sum_{k=0}^{n-1} f(X_k) = \pi(f), \quad \mathbb{P}_{\nu} - a.s. \quad (6.6)$$

Proof. Let $\nu \in M_1(X)$. Set $A = \left\{ \lim_{n \rightarrow \infty} n^{-1} \sum_{k=0}^{n-1} f(X_k) = \pi(f) \right\}$. From Section 5.2.1, we know that π is the unique invariant probability measure for $P_{\langle \pi, Q \rangle}^{MH}$ and therefore by Corollary 6.9, $0 = \mathbb{P}_{\pi}(A^c) = \int_X \pi(dx) h_{A^c}(x) = \pi(h_{A^c})$ where we have set $h_{A^c}(x) = \mathbb{E}_x[\mathbf{1}_{A^c}]$. Fix an arbitrary $x \in X$. Since

$$\left(\lim_{n \rightarrow \infty} n^{-1} \sum_{k=0}^{n-1} f(X_k) = \pi(f) \right) \Leftrightarrow \left(\lim_{n \rightarrow \infty} n^{-1} \sum_{k=1}^n f(X_k) = \pi(f) \right),$$

the set $A \in \mathcal{I}$ and hence $A^c \in \mathcal{I}$ since \mathcal{I} is a σ -field. Then, using the Markov property and $\mathbf{1}_{A^c} \circ S = \mathbf{1}_{A^c}$, we obtain

$$P_{\langle \pi, Q \rangle}^{MH} h_{A^c}(x) = \mathbb{E}_x [\mathbb{E}_{X_1} [\mathbf{1}_{A^c}]] = \mathbb{E}_x [\mathbb{E}_x [\mathbf{1}_{A^c} \circ S | \mathcal{F}_1]] = \mathbb{E}_x [\mathbf{1}_{A^c} \circ S] = \mathbb{E}_x [\mathbf{1}_{A^c}] = h_{A^c}(x).$$

This implies, by combining with (5.4)

$$h_{A^c}(x) = P_{\langle \pi, Q \rangle}^{MH} h_{A^c}(x) = \int \frac{q(x, y) \alpha(x, y)}{\pi(y)} h_{A^c}(y) \pi(y) \lambda(dy) + \bar{\alpha}(x) h_{A^c}(x).$$

Since $q > 0$ and $\pi > 0$, we can easily check that $\alpha > 0$ and $\bar{\alpha}(x) < 1$ (check it carefully). The first term in the rhs is null since $\pi(h_{A^c}) = 0$. Therefore, $(1 - \bar{\alpha}(x)) h_{A^c}(x) = 0$ and since $\bar{\alpha}(x) < 1$, we can conclude that $h_{A^c}(x) = \mathbb{E}_x [\mathbf{1}_{A^c}] = 0$. Finally, x being arbitrary, we obtain

$$\mathbb{P}_\nu(A^c) = \int_X \nu(dx) \mathbb{E}_x [\mathbf{1}_{A^c}] = 0,$$

and this is equivalent to $\mathbb{P}_\nu \left(\lim_{n \rightarrow \infty} n^{-1} \sum_{k=0}^{n-1} f(X_k) = \pi(f) \right) = 1$. The proof is concluded. \square

Theorem 6.11 is nice since the LLN holds \mathbb{P}_ν -a.s. for all starting distributions ν but the problem is that the result of Theorem 6.11 only concerns MH kernels. For a general kernel P , we have the nice following result.

Theorem 6.12. *If P is a Markov kernel on $X \times \mathcal{X}$ that admits a unique invariant probability measure π . Assume in addition that for all bounded functions h and all measures $\nu \in \mathbf{M}_1(X)$,*

$$\lim_{n \rightarrow \infty} \nu P^n h = \pi(h) \quad (6.7)$$

Then, for all initial distributions $\nu \in \mathbf{M}_1(X)$ and all $f \in \mathbf{F}(X)$ such that $\pi(|f|) = \int_X \pi(dx) |f(x)| < \infty$,

$$\lim_{n \rightarrow \infty} n^{-1} \sum_{k=0}^{n-1} f(X_k) = \pi(f), \quad \mathbb{P}_\nu - a.s. \quad (6.8)$$

Proof. We start as in the proof of Theorem 6.11. Let $\nu \in \mathbf{M}_1(X)$. Set

$$A = \left\{ \lim_{n \rightarrow \infty} n^{-1} \sum_{k=0}^{n-1} f(X_k) = \pi(f) \right\}.$$

Since π is the unique invariant probability measure for P , we have, by Corollary 6.9, $0 = \mathbb{P}_\pi(A^c) = \int_X \pi(dx) h_{A^c}(x) = \pi(h_{A^c})$ where we have set $h_{A^c}(x) = \mathbb{E}_x [\mathbf{1}_{A^c}]$. As in the proof of Theorem 6.11, $A^c \in \mathcal{I}$. Moreover,

$$\nu P^n h_{A^c} = \mathbb{E}_\nu [h_{A^c}(X_n)] = \mathbb{E}_\nu [\mathbb{E}_{X_n} [\mathbf{1}_{A^c}]] \stackrel{(1)}{=} \mathbb{E}_\nu [\mathbb{E}_\nu [\mathbf{1}_{A^c} \circ S^n | \mathcal{F}_n]] \stackrel{(2)}{=} \mathbb{E}_\nu [\mathbf{1}_{A^c} \circ S^n] \stackrel{(3)}{=} \mathbb{P}_\nu(A^c)$$

where $\stackrel{(1)}{=}$ follows from the Markov property, $\stackrel{(2)}{=}$ is the tower property, and $\stackrel{(3)}{=}$ follows from $\mathbf{1}_{A^c} \circ S^n = \mathbf{1}_{A^c}$, since $A^c \in \mathcal{I}$. Combined with (6.7), we get

$$\mathbb{P}_\nu(A^c) = \nu P^h h_{A^c} = \pi(h_{A^c}) = \int \pi(dx) \mathbb{E}_x [\mathbf{1}_{A^c}] = \mathbb{P}_\pi(A^c) = 0$$

Finally $\mathbb{P}_\nu(A^c) = 0$, that is $\mathbb{P}_\nu(A) = 1$, which is equivalent to

$$\lim_{n \rightarrow \infty} n^{-1} \sum_{k=0}^{n-1} f(X_k) = \pi(f), \quad \mathbb{P}_\nu - a.s.$$

□

Chapter 7

Geometric ergodicity and Central Limit theorems

Contents

7.1	Total variation norm and coupling	47
7.2	Geometric ergodicity	50
7.3	The Poisson Equation	53
7.3.1	Definition	53
7.3.2	Poisson equation and martingales	54
7.3.3	Central Limit theorems	55

Let (X_n) be a Markov chain with Markov kernel P and assume that P admits an invariant probability measure π . In this chapter, we are interested in finding conditions under which we can bound the error between the marginal distribution of X_n and the distribution π . To do so, we first need to define some notion of distance or pseudo-distance between probability measures. In this chapter we focus on the total variation distance. Then, we aim at comparing π and $P^n(x, \cdot)$, the distribution of the n -th iterate of the Markov kernel starting from an arbitrary point $x \in \mathbf{X}$, by bounding the total variation distance between them.

7.1 Total variation norm and coupling

We start with the notion of coupling between probability measures. If μ, ν are two probability measures on $(\mathbf{X}, \mathcal{X})$, then a coupling γ of (μ, ν) is a probability measure on the product space $(\mathbf{X}^2, \mathcal{X}^{\otimes 2})$ such that if $(X, Y) \sim \gamma$, then $X \sim \mu$ and $Y \sim \nu$.

Definition 7.1. Let $(\mathbf{X}, \mathcal{X})$ be a measurable space and let ν, μ be two probability measures $\mu, \nu \in \mathbf{M}_1(\mathbf{X})$. We define $\mathcal{C}(\mu, \nu)$, the coupling set associated to (μ, ν) as follows

$$\mathcal{C}(\mu, \nu) = \{\gamma \in \mathbf{M}_1(\mathbf{X}^2) : \forall A \in \mathcal{X}, \gamma(A \times \mathbf{X}) = \mu(A), \gamma(\mathbf{X} \times A) = \nu(A)\}$$

Any $\gamma \in \mathcal{C}(\mu, \nu)$ is called a coupling of (μ, ν) .

We can for example construct a coupling of (μ, μ) by sampling $X \sim \mu$ and by setting $Y = X$. The distribution of (X, Y) is then a coupling of (μ, μ) . This is a very "dependent" coupling (since we chose $X = Y$ by construction). We can also construct an "independent" coupling as follows: draw independently X and Y according to the same distribution μ , then the distribution of (X, Y) is a coupling of (μ, μ) (since $X \sim \mu$ and $Y \sim \mu$).

Before going further into coupling techniques, define the total variation norm and let us link it with coupling.

Definition 7.2. Let (X, \mathcal{X}) be a measurable space and let μ, ν be two probability measures $\mu, \nu \in \mathbf{M}_1(X)$. Then the total variation norm between μ and ν noted $\|\mu - \nu\|_{TV}$, is defined by

$$\|\mu - \nu\|_{TV} = 2 \sup \{ |\mu(f) - \nu(f)| : f \in F(X), 0 \leq f \leq 1 \} \quad (7.1)$$

$$= \int |\varphi_0 - \varphi_1|(x) \zeta(dx) \quad (7.2)$$

$$= 2 \inf \{ \mathbb{P}(X \neq Y) : (X, Y) \sim \gamma \text{ where } \gamma \in \mathcal{C}(\mu, \nu) \} \quad (7.3)$$

where $\mu(dx) = \varphi_0(x)\zeta(dx)$ and $\nu(dx) = \varphi_1(x)\zeta(dx)$.

Before showing that these different expressions of the total variation are indeed equivalent, let us make a few comments.

- (a) The reader might wonder why we can always write $\mu(dx) = \varphi_0(x)\zeta(dx)$ and $\nu(dx) = \varphi_1(x)\zeta(dx)$ for some well-chosen measure ζ , and measurable functions φ_0 and φ_1 . Actually, if we take $\mu, \nu \in \mathbf{M}_1(X)$, then setting $\zeta = \mu + \nu$ yields the two implications:

$$(\zeta(A) = 0) \implies (\mu(A) = 0) \quad \text{and} \quad (\zeta(A) = 0) \implies (\nu(A) = 0)$$

Therefore, the measure ζ dominates the measure μ and it also dominates the measure ν . By the Radon Nikodym theorem, the measures μ and ν have densities wrt ζ , densities that we call φ_0 and φ_1 in Definition 7.2.

- (b) The first definition, (7.1), is expressed as a supremum over functions, while the last one is an infimum over coupling measures... These two equalities can thus be considered as a *duality formula*.
- (c) The expression in the middle allows to write the total variation as a L_1 -norm between the two densities of the distributions wrt to a common dominating measure.
- (d) An immediate consequence of the equivalent definitions of the total variation norm is that if f is a measurable function taking values in $[0, 1]$ and if X, Y are random variables such that $(X, Y) \sim \gamma$ with $\gamma \in \mathcal{C}(\mu, \nu)$, then we have the coupling inequality

$$|\mu(f) - \nu(f)| \leq \mathbb{P}(X \neq Y)$$

This inequality will often be used in practice. It is due to such inequalities that coupling techniques are so successful.

Proof of the equivalences in Definition 7.2. Call A , B and C the quantities that appear respectively in (7.1), (7.2) and (7.3). Any function f satisfies $0 \leq f \leq 1$ if and only if the function $g = 2f - 1$ satisfies $|g| \leq 1$. This implies immediately

$$A = 2 \sup \{ |\mu(f) - \nu(f)| : f \in F(X), 0 \leq f \leq 1 \} = \sup \{ |\mu(g) - \nu(g)| : g \in F(X), |g| \leq 1 \}$$

Moreover, for any $g \in F(X)$ such that $|g| \leq 1$,

$$|\mu(g) - \nu(g)| = \left| \int (\varphi_1 - \varphi_0)(x)g(x)\zeta(dx) \right| \leq \int |\varphi_1 - \varphi_0|(x)|g(x)|\zeta(dx) = B.$$

Therefore, $A \leq B$. Moreover, setting $g^*(x) = \text{sign}(\varphi_0(x) - \varphi_1(x))$, we have $|g^*| = 1$ and therefore,

$$\begin{aligned} B &= \int |\varphi_0 - \varphi_1|(x)\zeta(dx) = \int (\varphi_0(x) - \varphi_1(x))g^*(x)\zeta(dx) \\ &= \mu(g^*) - \nu(g^*) \leq \sup \{ |\mu(g) - \nu(g)| : g \in F(X), |g| \leq 1 \} = A. \end{aligned}$$

Thus $A = B$. Now let $f \in F(X)$ be such that $0 \leq f \leq 1$ and let X, Y be random variables such that $(X, Y) \sim \gamma$ with $\gamma \in \mathcal{C}(\mu, \nu)$, then

$$|\mu(f) - \nu(f)| = |\mathbb{E}[f(X) - f(Y)]| = |\mathbb{E}[\{f(X) - f(Y)\} \mathbf{1}_{X \neq Y}]| \leq \mathbb{E}[|f(X) - f(Y)| \mathbf{1}_{X \neq Y}] \leq \mathbb{P}(X \neq Y).$$

This shows that $A \leq C$. To finish the proof, we will show that $C \leq B$ and to do so, we will exhibit an optimal coupling of (μ, ν) . Define

$$\epsilon = \int_{\mathbf{X}} \varphi_0 \wedge \varphi_1(x) \zeta(dx) \quad \text{and} \quad \zeta'(dx) = \frac{\varphi_0 \wedge \varphi_1(x)}{\epsilon} \zeta(dx)$$

Then, it can be readily checked that $\zeta' \in \mathbf{M}_1(\mathbf{X})$, and $\mu(dx) \geq \epsilon \zeta'(dx)$ and $\nu(dx) \geq \epsilon \zeta'(dx)$. This implies that there exist $\mu_1, \nu_1 \in \mathbf{M}_1(\mathbf{X})$ such that

$$\begin{aligned} \mu(dx) &= \epsilon \zeta'(dx) + (1 - \epsilon) \mu_1(dx) \\ \nu(dx) &= \epsilon \zeta'(dx) + (1 - \epsilon) \nu_1(dx) \end{aligned}$$

Define $\gamma(dxdy) = \epsilon \zeta'(dx) \delta_x(dy) + (1 - \epsilon) \mu_1(dx) \nu_1(dy)$. Obviously, $\gamma \in \mathcal{C}(\mu, \nu)$. Let X, Y be two random variables such that $(X, Y) \sim \gamma$. We can draw (X, Y) in the following way: draw a Bernoulli variable $U \sim \text{Ber}(\epsilon)$. If $U = 1$, draw $X \sim \zeta'$ and set $Y = X$. If $U = 0$, then draw independently $X \sim \mu_1$ and $Y \sim \nu_1$. Then, clearly

$$\begin{aligned} \mathbb{P}(X \neq Y) &= 1 - \mathbb{P}(X = Y) \leq 1 - \epsilon = 1 - \int_{\mathbf{X}} \varphi_0 \wedge \varphi_1(x) \zeta(dx) \\ &= \frac{1}{2} \int_{\mathbf{X}} \varphi_0 + \varphi_1(x) - 2\varphi_0 \wedge \varphi_1(x) \zeta(dx) \end{aligned}$$

This shows that $C \leq B$ and the proof is completed. \square

Exercise 7.3. Show that the supremum in (7.1) is attained with a convenient choice of f .

Another equivalent expression of the total variation distance is

$$\|\mu - \nu\|_{\text{TV}} = 2 \inf \{ \gamma(\Delta) : \gamma \in \mathcal{C}(\mu, \nu) \} \quad (7.4)$$

where we have used the notation $\Delta(x, y) = \mathbf{1}_{x \neq y}$ (we also say that $\Delta(x, y)$ is the Hamming distance between x and y).

Example 7.4. See Figure 7.1. Let $\mu = \mathcal{N}(-1, 1)$ and $\nu = \mathcal{N}(1, 1)$. Let $X \sim \mathcal{N}(-1, 1)$ and set $Y = X + 2$. Then, (X, Y) is a coupling of (μ, ν) but it is not the optimal coupling for the Hamming distance since $\mathbb{P}(X \neq Y) = 1$, whereas using Definition 7.2,

$$\|\mu - \nu\|_{\text{TV}} = 2 \left(1 - \int_{-\infty}^{\infty} \phi(x+1) \wedge \phi(x-1) dx \right) = 2 \left(1 - 2 \int_1^{\infty} \frac{e^{-u^2/2}}{\sqrt{2\pi}} du \right),$$

where ϕ the density of the standard Gaussian distribution.

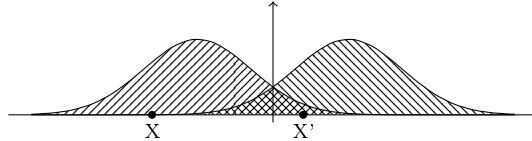


Fig. 7.1 An example of coupling of two probability measures.

7.2 Geometric ergodicity

In what follows, we assume that for some measurable function $V : \mathbf{X} \rightarrow [1, \infty)$, we have

(A1) [**Minorizing condition**] for all $d > 0$, there exists $\epsilon_d > 0$ and a probability measure ν_d such that

$$\forall x \in C_d := \{V \leq d\}, \quad P(x, \cdot) \geq \epsilon_d \nu_d(\cdot) \quad (7.5)$$

(A2) [**Drift condition**] there exists a constants $(\lambda, b) \in (0, 1) \times \mathbb{R}^+$ such that for all $x \in \mathbf{X}$,

$$PV(x) \leq \lambda V(x) + b$$

Typically, the function V is unbounded (but in particular situations, it can also be bounded) and the level set $\{V \leq d\}$ is typically compact (when the chain takes value a topological space)... Roughly speaking, (A1) tells you that wherever x moves in a set C_d , the measure $P(x, \cdot)$ is lower bounded by the non-trivial measure $\epsilon_d \nu_d(\cdot)$. In many cases, $\mathbf{X} = \mathbb{R}^n$, and P is dominated by the Lebesgue measure: $P(x, dy) = p(x, y)dy$. In that case, we usually take $P(x, A) \geq \epsilon_d \nu_d(A)$ where

$$\epsilon_d = \int_{\mathbf{X}} \left[\inf_{x \in C_d} p(x, y) \right] dy, \quad \nu_d(A) = \frac{\int_A \inf_{x \in C_d} p(x, y) dy}{\epsilon_d}$$

i.e. we only need to bound from below the kernel density $p(x, y)$ when $x \in C_d$. If C_d is compact, then it is quite easy to check such lower-bound. In the Markov chain terminology, if (7.5) holds, we say that C_d is a small set.

The drift condition (A2) tells you that in the mean sense, the drift function V is shrinked by a factor λ up to the additive constant b ... Intuitively speaking, the Markov kernel P does not bring to regions where V is too large so that the chain does not go to infinity too quickly (since limited values of V corresponds typically to bounded sets). And we can easily imagine that such chains will have nice ergodic properties.

Before stating the result, we must say that, in practise, for a given Markov kernel P , there is no general rule for guessing the expression of a drift function V that satisfies (A2), and we have to try different functions V for checking the assumptions... For example, if $X_{k+1} = \alpha X_k + \epsilon_k$ where (ϵ_k) are iid and $\alpha \in (0, 1)$. If we know that $\mathbb{E}[|\epsilon_1|^r] < \infty$, then we can try a drift function $V(x) = |x|^r$ and if $\mathbb{E}[e^{\beta \epsilon_1}] < \infty$, then we can try $V(x) = e^{\beta x}$. For MH algorithms, we also sometimes use a negative power of the target density. But once again, the choice of V is very model specific. We now show that assumptions (A1) and (A2) imply that the Markov kernel P is "geometrically ergodic" in the following sense.

Theorem 7.5. [Geometric ergodicity] Assume (A1) and (A2) for some measurable function $V \geq 1$. Then, there exists a constant $\varrho \in (0, 1)$ such that for all $x, x' \in \mathbf{X}$ and all $n \in \mathbb{N}$,

$$\|P^n(x, \cdot) - P^n(x', \cdot)\|_{TV} \leq \varrho^n [V(x) + V(x')].$$

Remark 7.6. Assume that there exist a constant $\epsilon > 0$ and a probability measure ν such that for all $x \in \mathbf{X}$, $P(x, \cdot) \geq \epsilon \nu(\cdot)$. In that case, (A1) and (A2) are satisfied with the constant function $V(x) = 1$ and Theorem 7.5 then shows that

$$\|P^n(x, \cdot) - P^n(x', \cdot)\|_{TV} \leq 2\varrho^n.$$

for some constant $\varrho \in (0, 1)$. Such a Markov chain is usually said to be uniformly ergodic.

The proof needs several steps. To bound $\|P^n(x, \cdot) - P^n(x', \cdot)\|_{TV}$, we will construct a bivariate Markov chain (X_k, X'_k) such that first component process (X_k) behaves marginally as a Markov chain starting from x with Markov kernel P , while the second component process (X'_k) behaves marginally as a Markov chain starting from x' with Markov kernel P . Let us be more specific... In what follows, we choose d sufficiently large so that

$$\bar{\lambda} := \lambda + \frac{2b}{1+d} < 1 \quad (7.6)$$

Definition of the joint kernel \bar{P}

Define $Q(x_k, dx_{k+1}) = \frac{P(x_k, dx_{k+1}) - \epsilon_d \nu_d(dx_{k+1})}{1 - \epsilon_d}$ and set

$$\begin{aligned} \bar{P}((x_k, x'_k), dx_{k+1} dx'_{k+1}) &= \mathbf{1}_{x_k = x'_k} P(x_k, dx_{k+1}) \delta_{x_{k+1}}(x'_{k+1}) \\ &+ \mathbf{1}_{x_k \neq x'_k} \mathbf{1}_{(x_k, x'_k) \notin \bar{C}_d} [P(x_k, dx_{k+1}) P(x'_k, dx'_{k+1})] \\ &+ \mathbf{1}_{x_k \neq x'_k} \mathbf{1}_{(x_k, x'_k) \in \bar{C}_d} [\epsilon_d \nu_d(dx_{k+1}) \delta_{x_{k+1}}(x'_{k+1}) + (1 - \epsilon_d) Q(x_k, dx_{k+1}) Q(x'_k, dx'_{k+1})] \end{aligned}$$

Actually, \bar{P} is a Markov kernel on $\mathbf{X}^2 \times \mathcal{X}^{\otimes 2}$ and it can be easily checked that

$$\bar{P}((x, x'), \cdot) \in \mathcal{C}(P(x, \cdot), P(x', \cdot)) \quad (7.7)$$

This will indeed imply by induction that for any $n \in \mathbb{N}$,

$$\bar{P}^n((x, x'), \cdot) \in \mathcal{C}(P^n(x, \cdot), P^n(x', \cdot)) \quad (7.8)$$

Interpretation of the joint kernel \bar{P}

Set $\bar{X}_k = (X_k, X'_k)$ and $\bar{C}_d = C_c \times C_d$. If $(\bar{X}_k)_{k \in \mathbb{N}}$ is a Markov chain with the Markov kernel \bar{P} , the transition from $\bar{X}_k = (x_k, x'_k)$ to $\bar{X}_{k+1} = (X_{k+1}, X'_{k+1})$ can be seen as follows

- If $x_k = x'_k$, draw $X_{k+1} \sim P(x_k, \cdot)$ and set $X'_{k+1} = X_{k+1}$.
- Otherwise,
 - If $(x_k, x'_k) \notin \bar{C}_d$, then
 - Draw independently $X_{k+1} \sim P(x_k, \cdot)$ and $X'_{k+1} \sim P(x'_k, \cdot)$
 - If $(x_k, x'_k) \in \bar{C}_d$, then
 - Draw $U \sim \text{Ber}(\epsilon_d)$.
 - If $U = 1$, draw $X_{k+1} \sim \nu_d$ and set $X'_{k+1} = X_{k+1}$.
 - If $U = 0$, draw independently $X_{k+1} \sim Q(x_k, \cdot)$ and $X'_{k+1} \sim Q(x'_k, \cdot)$.
- Set $\bar{X}_{k+1} = (X_{k+1}, X'_{k+1})$.

Therefore, the bivariate Markov chain $(\bar{X}_k)_{k \in \mathbb{N}} = (X_k, X'_k)_{k \in \mathbb{N}}$ is such that it tries to couple its two components with probability ϵ_d each time it falls into \bar{C}_d and once it couples (ie $X_k = X'_k$) then, it stays together for ever (ie for all $n \geq k$, $X_n = X'_n$).

Some nice properties of \bar{P}

The following inequalities are immediate:

1. Set $\Delta(x, x') = \mathbf{1}_{x \neq x'}$, then
 - (a) if $(x, x') \in \bar{C}_d$, $\bar{P}\Delta(x, x') \leq (1 - \epsilon_d)\Delta(x, x')$
 - (b) if $(x, x') \notin \bar{C}_d$, $\bar{P}\Delta(x, x') \leq \Delta(x, x')$
2. Setting $\bar{V}(x, x') = (V(x) + V(x'))/2$, we have $\bar{P}\bar{V}(x, x') = 2^{-1}(PV(x) + PV(x')) \leq \lambda \bar{V}(x, x') + b$. This implies
 - (a) if $(x, x') \in \bar{C}_d$, $\bar{P}\bar{V}(x, x') \leq (\lambda + b)\bar{V}(x, x')$

$$(b) \text{ if } (x, x') \notin \bar{C}_d, \bar{P}\bar{V}(x, x') \leq \underbrace{\left(\lambda + \frac{2b}{1+d}\right)}_{\bar{\lambda}} \bar{V}(x, x')$$

We now have all the tools for proving Theorem 7.5.

of Theorem 7.5. For any $\beta \in (0, 1)$, define

$$\varrho_\beta = \max((1 - \epsilon_d)^{1-\beta}(\lambda + b)^\beta, \bar{\lambda}^\beta) \quad (7.9)$$

The expression of ϱ_β may seem a bit complicated (we will understand why we choose ϱ_β like this in (7.10) below) but, since $\bar{\lambda}$ and $1 - \epsilon_d$ are both in $(0, 1)$, we can always pick β sufficiently small (but positive) so that $\varrho_\beta \in (0, 1)$. This ϱ_β being chosen, set $W = \Delta^{1-\beta} \bar{V}^\beta$. Then, using Holder's inequality and the inequalities in the section **Some nice properties of \bar{P}** , we have for all $(x, x') \in \mathbf{X}^2$,

$$\begin{aligned} \bar{P}W(x, x') &= \bar{P}(\Delta^{1-\beta} \bar{V}^\beta)(x, x') \leq (\bar{P}\Delta(x, x'))^{1-\beta} (\bar{P}\bar{V}(x, x'))^\beta \\ &\leq (\Delta^{1-\beta} \bar{V}^\beta)(x, x') \times \begin{cases} (1 - \epsilon_d)^{1-\beta}(\lambda + b)^\beta & \text{if } (x, x') \in C_d^2 \\ \bar{\lambda}^\beta & \text{if } (x, x') \notin C_d^2 \end{cases} \\ &\leq \varrho_\beta W(x, x') \end{aligned}$$

This implies by induction that for all $n \in \mathbb{N}$ and all $(x, x') \in \mathbf{X}^2$,

$$\bar{P}^n W(x, x') \leq \varrho_\beta^n W(x, x') \quad (7.10)$$

Then

$$\|P^n(x, \cdot) - P^n(x', \cdot)\|_{\text{TV}} \stackrel{(1)}{\leq} 2\bar{P}^n \Delta(x, x') \stackrel{(2)}{\leq} 2\bar{P}^n W(x, x') \stackrel{(3)}{\leq} 2\varrho_\beta^n W(x, x') \stackrel{(4)}{\leq} \varrho_\beta^n (V(x) + V(x'))$$

where (1) comes from (7.8) and (7.4), (2) from $\Delta(x, x') = \Delta^{1-\beta}(x, x') \leq W(x, x')$ because $V \geq 1$, (3) from (7.10) and (4) from

$$W(x, x') \leq \left(\frac{V(x) + V(x')}{2} \right)^\beta \leq \frac{V(x) + V(x')}{2}$$

since $V \geq 1$ and $\beta \in (0, 1)$. □

Corollary 7.7. *Assume that (A1) and (A2) hold for some measurable function $V \geq 1$. Then, the Markov kernel P admits a unique invariant probability measure π . Moreover, $\pi(V) < \infty$ and there exists constants $(\varrho, \alpha) \in (0, 1) \times \mathbb{R}^+$ such that for all $\mu \in \mathbf{M}_1(\mathbf{X})$ and all $n \in \mathbb{N}$,*

$$\|\mu P^n - \pi\|_{\text{TV}} \leq \alpha \varrho^n \mu(V).$$

Proof. For any $\mu, \nu \in \mathbf{M}_1(\mathbf{X})$ and any $h \in \mathbf{F}(\mathbf{X})$ such that $|h| \leq 1$, we have, using Theorem 7.5,

$$|\mu P^n h - \nu P^n h| = \left| \int_{\mathbf{X}^2} \mu(dx) \nu(dy) [P^n h(x) - P^n h(y)] \right| \leq \int_{\mathbf{X}^2} \mu(dx) \nu(dy) |[P^n h(x) - P^n h(y)]| \leq \varrho^n [\mu(V) + \nu(V)]$$

Thus,

$$\|\mu P^n - \nu P^n\|_{\text{TV}} \leq \varrho^n [\mu(V) + \nu(V)] \quad (7.11)$$

Replacing μ by δ_x and ν by $P(x, \cdot)$, we get for all $x \in \mathbf{X}$,

$$\|P^n(x, \cdot) - P^{n+1}(x, \cdot)\|_{\text{TV}} \leq \varrho^n [V(x) + PV(x)] \leq \varrho^n [(1 + \lambda)V(x) + b]$$

This implies that $\{P^n(x, \cdot)\}$ is a Cauchy sequence and since $(M_1(X), \|\cdot\|_{TV})$ is complete, it converges to a limit $\pi \in M_1(X)$. Then, for all $x \in X$ and all $h \in F(X)$ such that $|h| \leq 1$, we also have $|Ph| \leq 1$ and therefore

$$\pi(Ph) = \lim_{n \rightarrow \infty} P^n(Ph)(x) = \lim_{n \rightarrow \infty} P^{n+1}h(x) = \pi(h)$$

showing that π is **P -invariant**. We now show **uniqueness** of an invariant probability measure. To see this, note that π actually does not depend on the choice of x . Indeed, replacing μ by δ_x and ν by $\delta_{x'}$ in (7.11), we get that $\lim_{n \rightarrow \infty} \|P^n(x, \cdot) - P^n(x', \cdot)\|_{TV} = 0$. Therefore, for all $x \in X$, $\lim_{n \rightarrow \infty} P^n h(x) = \pi(h)$. Let π' be an invariant probability measure for P , then

$$\pi'(h) = \pi' P^n(h) = \int \pi'(dx) \underbrace{P^n h(x)}_{\rightarrow \pi(h)} \rightarrow_{n \rightarrow \infty} \pi(h)$$

where the last equality comes from Lebesgue's dominated convergence theorem. Since $PV \leq \lambda V + b$, we have by induction for all $n \in \mathbb{N}$,

$$P^n V(x) \leq \lambda^n V(x) + b \left(\sum_{k=0}^{n-1} \lambda^k \right) \leq \lambda^n V(x) + \frac{b}{1-\lambda}$$

Therefore, for any $M > 0$, by Jensen's inequality applied to the convex function $u \mapsto u \wedge M$, we have $P^n(V \wedge M)(x) \leq (P^n V(x)) \wedge M \leq \left(\lambda^n V(x) + \frac{b}{1-\lambda} \right) \wedge M$. We then integrate wrt π and use $\pi = \pi P^n$:

$$\pi(V \wedge M) = \pi P^n(V \wedge M) \leq \int \pi(dx) \left(\lambda^n V(x) + \frac{b}{1-\lambda} \right) \wedge M$$

The Lebesgue dominated convergence theorem then shows by letting n to infinity, $\pi(V \wedge M) \leq \frac{b}{1-\lambda} \wedge M$. Then, letting M to infinity, we get $\pi(V) \leq b/(1-\lambda) < \infty$. To complete the proof, apply (7.11) with $\nu = \pi$, we get

$$\|\mu P^n - \underbrace{\pi P^n}_{\pi}\|_{TV} \leq \varrho^n [\mu(V) + \pi(V)] \leq \alpha \varrho^n [\mu(V)]$$

with $\alpha = 1 + \pi(V) < \infty$. □

Exercise 9.12 shows that under (A1) and (A2) a Law of Large number is valid, starting from any initial distribution.

Theorem 7.7 is nice but it is expressed only as a bound on the total variation norm, this implies that we can bound $|\mu P^n h - \pi(h)|$ where $|h| \leq 1$. Under the same assumptions, we can actually obtain more general bounds for possibly unbounded functions h . The proof is slightly more complicated so we decide to just state the result.

Theorem 7.8. *Assume that (A1) and (A2) hold for some function $V \geq 1$. Then, there exist constants $(\varrho, \alpha) \in (0, 1) \times \mathbb{R}^+$ such that for all $\mu \in M_1(X)$ satisfying $\mu(V) < \infty$, all $n \in \mathbb{N}$ and all measurable functions $|h| \leq V$,*

$$|\mu P^n h - \pi(h)| \leq \alpha \varrho^n \mu(V).$$

7.3 The Poisson Equation

7.3.1 Definition

We start with a general definition of a Poisson equation.

Definition 7.9. For a given measurable function h such that $\pi|h| < \infty$, the Poisson equation is defined by

$$\hat{h} - P\hat{h} = h - \pi(h) \quad (7.12)$$

A solution to Poisson equation (7.12) is a function \hat{h} such that $P|\hat{h}|(x) < \infty$ for all $x \in X$ and for all $x \in X$, $\hat{h}(x) - P\hat{h}(x) = h(x) - \pi(h)$.

The following result holds under the set of assumptions (A1) and (A2).

Theorem 7.10. Assume (A1) and (A2) hold for some measurable function $V \geq 1$. Then, for any function h such that $|h| \leq V$, the function

$$\hat{h} = \sum_{n=0}^{\infty} \{P^n h - \pi(h)\} \quad (7.13)$$

is well-defined. Moreover, \hat{h} is a solution of the Poisson equation associated to h and there exists a constant γ such that for all $x \in X$,

$$|\hat{h}(x)| \leq \gamma V(x)$$

Proof. To see the existence of a solution to the Poisson equation under (A1) and (A2), note that by Theorem 7.8, $\sum_{n=0}^{\infty} \{P^n h(x) - \pi(h)\}$ converges for any $|h| \leq V$ and we can thus define

$$\hat{h}(x) = \sum_{n=0}^{\infty} \{P^n h(x) - \pi(h)\}$$

Then,

$$P\hat{h}(x) = \sum_{n=1}^{\infty} \{P^n h(x) - \pi(h)\}$$

which immediately shows (7.12). Moreover, setting \hat{h} as in (7.13), Theorem 7.8 shows that for all $x \in X$,

$$|\hat{h}(x)| \leq \frac{\alpha}{1-\varrho} V(x)$$

□

7.3.2 Poisson equation and martingales

The interest of Poisson equation is that it allows to link quantities of interest of our Markov chain with a well-chosen martingale. Then, we apply limiting results on martingales and the impact of those results to our Markov chain.

We start with a refresher on martingales.

7.3.2.1 A refresh on martingales

Let $(M_n)_{n \in \mathbb{N}}$ be a sequence of random variables on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let $(\mathcal{F}_n)_{n \in \mathbb{N}}$ be a filtration (ie for all $n \in \mathbb{N}$, $\mathcal{F}_n \subset \mathcal{F}_{n+1} \subset \mathcal{F}$). We say that $(M_n)_{n \in \mathbb{N}}$ is a (\mathcal{F}_n) -martingale if for all $n \in \mathbb{N}$, M_n is integrable and for all $n \geq 1$,

$$\mathbb{E}[M_n | \mathcal{F}_{n-1}] = M_{n-1}$$

The *increment process* of the martingale is by definition $(M_{n+1} - M_n)_{n \in \mathbb{N}}$.

The following CLT result holds for martingales with stationary increments. It is stated without proof.

Theorem 7.11. *If a sequence $(M_n)_{n \in \mathbb{N}}$ is a (\mathcal{F}_n) -martingale with stationary and square integrable increments, then*

$$n^{-1/2}M_n \xrightarrow{\mathcal{L}_{\mathbb{P}}} \mathcal{N}(0, \mathbb{E}[(M_1 - M_0)^2])$$

7.3.2.2 Link with martingales

Define

$$S_n(h) = \sum_{k=0}^{n-1} \{h(X_k) - \pi(h)\}$$

The solution of the Poisson equation allows us to relate $S_n(h)$ to a martingale by writing:

$$S_n(h) = M_n(\hat{h}) + \hat{h}(X_0) - \hat{h}(X_n) \quad (7.14)$$

where

$$M_n(\hat{h}) = \sum_{k=1}^n \left\{ \hat{h}(X_k) - P\hat{h}(X_{k-1}) \right\} \quad (7.15)$$

Note that $\{M_n(\hat{h})\}_{n \in \mathbb{N}}$ is indeed a (\mathcal{F}_k) -martingale where $\mathcal{F}_k = \sigma(X_0, \dots, X_k)$ since:

$$\mathbb{E}[M_n(\hat{h}) | \mathcal{F}_{n-1}] - M_{n-1}(h) = \mathbb{E}[\hat{h}(X_n) - P\hat{h}(X_{n-1}) | \mathcal{F}_{n-1}] = P\hat{h}(X_{n-1}) - P\hat{h}(X_{n-1}) = 0$$

This link with martingales allows to obtain LLN and Central Limit theorems for our Markov chain from limiting results on martingales. Since LLN has been already studied in a different approach in the previous chapter, we only focus here on CLT.

7.3.3 Central Limit theorems

Theorem 7.12. *Let P be a Markov kernel with a unique invariant probability measure π . Let $h \in L_2(\pi)$. Assume that there exists a solution $\hat{h} \in L_2(\pi)$ to the Poisson equation $\hat{h} - P\hat{h} = h$. Then*

$$n^{-1/2} \sum_{k=0}^{n-1} \{h(X_k) - \pi(h)\} \xrightarrow{\mathcal{L}_{\mathbb{P}_\pi}} \mathcal{N}(0, \sigma_\pi^2(h)),$$

where

$$\sigma_\pi^2(h) = \mathbb{E}_\pi[\{\hat{h}(X_1) - P\hat{h}(X_0)\}^2] \quad (7.16)$$

Proof. Without loss of generality, we assume $\pi(h) = 0$. The sequence $(M_n(\hat{h}))_{n \in \mathbb{N}}$ defined in (7.15) is such that

$$\begin{aligned} \mathbb{E}_\pi[(M_n(\hat{h}) - M_{n-1}(\hat{h}))^2] &= \mathbb{E}_\pi[\{\hat{h}(X_1) - P\hat{h}(X_0)\}^2] \leq 2\mathbb{E}_\pi[\hat{h}^2(X_1) + (P\hat{h}(X_0))^2] \\ &= 2 \left[\pi(\hat{h}^2) + \pi((P\hat{h})^2) \right] \stackrel{(1)}{\leq} 2 \left[\pi(\hat{h}^2) + \underbrace{\pi P(\hat{h})^2}_\pi \right] = 2\pi(\hat{h}^2) < \infty \end{aligned}$$

where $\stackrel{(1)}{\leq}$ follows from Cauchy-Schwarz inequality. Therefore, the sequence $(M_n(\hat{h}))_{n \in \mathbb{N}}$ is a martingale with stationary and square integrable increments under \mathbb{P}_π . By Theorem 7.11, we have

$$n^{-1/2}M_n(\hat{h}) \xrightarrow{\mathcal{L}_{\mathbb{P}_\pi}} \mathcal{N}\left(0, \mathbb{E}_\pi[\{\hat{h}(X_1) - P\hat{h}(X_0)\}^2]\right). \quad (7.17)$$

Since the Markov chain $(X_k)_{k \in \mathbb{N}}$ is stationary under \mathbb{P}_π , we get $\mathbb{E}_\pi[|\hat{h}(X_0) + \hat{h}(X_n)|] \leq 2\pi(|\hat{h}|)$ which implies that

$$n^{-1/2}\{\hat{h}(X_0) + \hat{h}(X_n)\} \xrightarrow{\mathbb{P}_\pi\text{-prob}} 0.$$

Combining it with (7.17) and (7.14) and using Slutsky's lemma gives:

$$n^{-1/2} \sum_{k=0}^{n-1} h(X_k) \xrightarrow{\mathcal{L}_{\mathbb{P}_\pi}} \mathcal{N}(0, \sigma_\pi^2(h))$$

□

Theorem 7.13. *Assume that (A1) and (A2) hold for some function V . Then, for all measurable functions h such that $|h|^2 \leq V$,*

$$n^{-1/2} \sum_{k=0}^{n-1} \{h(X_k) - \pi(h)\} \xrightarrow{\mathcal{L}_{\mathbb{P}_\pi}} \mathcal{N}(0, \sigma_\pi^2(h)),$$

where

$$\sigma_\pi^2(h) = \mathbb{E}_\pi[\{\hat{h}(X_1) - P\hat{h}(X_0)\}^2] \quad (7.18)$$

and \hat{h} is defined as in (7.13).

Proof. Assume that (A1) and (A2) hold for some function V . Then, (A1) also holds with V replaced by $V^{1/2}$. Moreover, since $PV \leq \lambda V + b$, we have by Cauchy-Schwarz,

$$P(V^{1/2}) \leq (PV)^{1/2} \leq (\lambda V + b)^{1/2} \leq \lambda^{1/2}V^{1/2} + b^{1/2}$$

Finally, (A1) and (A2) hold for the function $V^{1/2}$. We can therefore apply Theorem 7.10 with V replaced by $V^{1/2}$. Then, for all $h \leq V^{1/2}$, the function \hat{h} defined by (7.13) is solution to the Poisson equation and there exists a constant $\gamma > 0$ such that $\hat{h} \leq \gamma V^{1/2}$. This implies that $\pi(\hat{h}^2) \leq \gamma\pi(V) < \infty$ by Theorem 7.7. Therefore $\hat{h} \in L_2(\pi)$ and Theorem 7.12 applies. The proof is completed. □

Under the assumptions of Theorem 7.13, the CLT holds under \mathbb{P}_π . We can actually extend this result to all \mathbb{P}_ν where ν is any probability measure in $\mathbf{M}_1(\mathbf{X})$.

Variants of MH algorithms

Contents

8.1	Generalisation of MH Algorithms	57
8.2	Pseudo marginal Monte Carlo methods	58
8.3	Hamiltonian Monte Carlo	59
8.3.1	MH with deterministic moves	60
8.3.2	Hamiltonian dynamics	61
8.3.3	The leapfrog integrator	63
8.4	Data augmentation	64
8.4.1	Two-stage Gibbs sampler	67

In this chapter, we describe diverse variants of MH algorithms. Recall that we are given a target distribution π . In classical MH, we construct a Markov chain (X_k) that admits π as invariant probability measure. In most of the variants presented in this chapter, we extend (X_k) by adding a component, say (U_k) and such that (X_k, U_k) is a Markov chain that admits an invariant probability measure Π with the property that Π has π as its marginal distribution wrt the first component. Finally $(X_k)_{k \in \mathbb{N}}$ alone is not a Markov chain, on the contrary to $(X_k, U_k)_{k \in \mathbb{N}}$.

We start with a general version of MH algorithms that will be useful in many contexts.

8.1 Generalisation of MH Algorithms

Let $\pi \in \mathcal{M}_1(\mathcal{X})$ and let Q be a Markov kernel on $\mathcal{X} \times \mathcal{X}$. In Section 5.1.1, we have presented the Metropolis-Hastings algorithm when π and $Q(x, \cdot)$ have both densities wrt a common dominating measure λ . Here we do not make such an assumption so that the expression of α^{MH} given in Lemma 5.5 is not available anymore and should be adapted. Instead, we will need the following assumption. Define

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

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

$$\int h(x, y) \mu_1(dx dy) = \int h(x, y) r(x, y) \mu_0(dx dy) \tag{8.1}$$

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

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

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 5.5 is replaced by $\alpha(x, y) = r(x, y) \wedge 1$ is π -reversible.

Lemma 8.1. *Assume (B1). Then, setting $\alpha(x, y) = r(x, y) \wedge 1$, the MH kernel:*

$$P_{\langle \pi, Q \rangle}^{MH}(x, dy) = Q(x, dy)\alpha(x, y) + \bar{\alpha}(x)\delta_x(dy) \quad \text{where} \quad \bar{\alpha}(x) = 1 - \int_{\mathbf{X}} Q(x, dy)\alpha(x, y)$$

is π -reversible.

Proof. Similarly to Lemma 5.4, we only need to check the detailed balance condition. Let $h \in F_+(\mathbf{X})$, then,

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

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

8.2 Pseudo marginal Monte Carlo methods

Assume that π and Q are dominated by a common dominating measure λ and write by abuse of notation, $\pi(dx) = \pi(x)\lambda(dx)$ and $Q(x, dy) = q(x, y)\lambda(dy)$. When considering a Metropolis-Hastings algorithm, we need an explicit expression of $\pi(x)$ for any $x \in \mathbf{X}$, 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, say that you draw $W \sim R(x, dw)$ where R is a Markov kernel from \mathbf{X} to \mathbb{R}_*^+ , that is, a Markov kernel on $\mathbf{X} \times \mathcal{B}(\mathbb{R}_*^+)$ such that $\int_{\mathbb{R}_*^+} wR(x, dw) = \pi(x)$ (the *unbiasedness* condition).

The pseudo marginal algorithm works as described in Algorithm 8.2 below. Finally, this algo-

Algorithm 6 The Pseudo-Marginal MH Algorithm.

At $t = 0$, draw X_0 according to some arbitrary distribution and draw $W_0 \sim R(X_0, \cdot)$.

for $t = 0 \rightarrow n - 1$ **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 $(X_{t+1}, W_{t+1}) = \begin{cases} (\tilde{X}_{t+1}, \tilde{W}_{t+1}) & \text{with prob. } \frac{\tilde{W}_{t+1}q(\tilde{X}_{t+1}, X_t)}{W_t q(X_t, \tilde{X}_{t+1})} \wedge 1 \\ (X_t, W_t) & \text{with prob. } 1 - \frac{\tilde{W}_{t+1}q(\tilde{X}_{t+1}, X_t)}{W_t q(X_t, \tilde{X}_{t+1})} \wedge 1 \end{cases}$

end for

gorithm is very close to the classical MH except that we replace $\pi(x)$ by its unbiased estimator.

We now justify Pseudo-marginal Monte Carlo methods by showing that it is actually a ("disguised") generalized MH algorithm (as described in Lemma 8.1) by considering "extended Markov chain", $(\tilde{X}_k)_{k \in \mathbb{N}} = (X_k, W_k)_{k \in \mathbb{N}}$ on an extended space and with an extended target. Define the

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

$$\iint_{X \times \mathbb{R}_*^+} \Pi(dx dw) = \int_X \left(\int_{\mathbb{R}_*^+} wR(x, dw) \right) \lambda(dx) = \int_X \pi(x) \lambda(dx) = 1$$

Moreover, in Algorithm 8.2, the candidate $(\tilde{X}_{t+1}, W_{t+1})$ is proposed according to \bar{Q} where the proposal kernel \bar{Q} is defined by $\bar{Q}(\bar{x}, d\bar{x}') = Q(x, dx')R(x', dw')$.

In order to check (B1), we first set

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

Then, writing $Q(x, dy) = q(x, y)\lambda(dy)$, we obtain for all $h \in F_+(\bar{X}^2)$,

$$\begin{aligned} \int_{\bar{X}^2} h(\bar{x}, \bar{x}') \mu_1(d\bar{x}d\bar{x}') &= \int_{\bar{X}^2} h(\bar{x}, \bar{x}') w' q(x', x) [R(x, dw)R(x', dw')\lambda(dx)\lambda(dx')] \\ &= \int_{\bar{X}^2} h(\bar{x}, \bar{x}') \underbrace{\frac{w' q(x', x)}{w q(x, x')}}_{r(\bar{x}, \bar{x}')} \mu_0(d\bar{x}d\bar{x}'). \end{aligned}$$

Since $r > 0$, we can apply Lemma 8.1 with $\alpha(\bar{x}, \bar{x}') = r(\bar{x}, \bar{x}') \wedge 1$ and we finally get that $P_{\langle \Pi, \bar{Q} \rangle}^{MH}(\bar{x}, d\bar{x}')$ is Π -reversible. Since Algorithm 8.2 corresponds to applying the Markov kernel $P_{\langle \Pi, \bar{Q} \rangle}^{MH}$, this completes the proof. Note that the extended target distribution Π has the marginal π wrt the first component:

$$\Pi(A \times \mathbb{R}_*^+) = \int_A \int_{\mathbb{R}_*^+} wR(x, dw)\lambda(dx) = \int_A \pi(dx) = \pi(A).$$

To sum up, $(\bar{X}_k)_{k \in \mathbb{N}} = (X_k, W_k)_{k \in \mathbb{N}}$ produced by Algorithm 8.2 is a generalized Metropolis-Hastings algorithm where the target distribution Π admits π as the marginal distribution on the first component. Note that $(X_k)_{k \in \mathbb{N}}$ is not a Markov chain anymore (but $(\bar{X}_k)_{k \in \mathbb{N}}$ is).

8.3 Hamiltonian Monte Carlo

In Hamiltonian Monte Carlo (HMC), we again extend the target density and construct a Markov chain on an extended space. We assume here that we have a target distribution on \mathbb{R}^d , say π , and we write $\pi(q) \propto e^{-U(q)}$ (in this litterature, the "mute" variable x is replaced by q). Nothing very restrictive so far... We may consider π as the marginal of the extended target

$$\Pi(q, p) \propto \exp \{ -U(q) - p^T p / 2 \}, \quad p, q \in \mathbb{R}^d \quad (8.2)$$

We can see that this extended target density can be written as the product of two densities: π for the first component and the normal density of $\mathcal{N}(0, I_d)$ for the second component. At this stage, adding a second component in the target distribution that is completely independent of the first component and with such a classical distribution as $\mathcal{N}(0, I_d)$ does not seem to bring too much excitement in the problem but let's be patient... In what follows, we make use of the following terminology (that comes from physicists:)

- $q \in \mathbb{R}^d$ is the position and $U(q)$ is called the *potential energy*.
- $p \in \mathbb{R}^d$ is the momentum and $K(p) = p^T p / 2$ 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 $(X_t)_{t \in \mathbb{N}} = (q_t, p_t)_{t \in \mathbb{N}}$ as described in Algorithm 8.3.

Algorithm 7 The Leapfrog HMC.

At $t = 0$, draw X_0 according to some arbitrary distribution.
for $t = 0 \rightarrow n - 1$ **do**
 Set $q_{t+1}^0 = q_t$ and draw $p_{t+1}^0 \sim \mathcal{N}(0, I_d)$.
 for $k = 0$ to $k = L - 1$ **do**
 $p_{t+1}^{k+1/2} = p_{t+1}^k - (h/2) \nabla U(q_{t+1}^k)$.
 $q_{t+1}^{k+1} = q_{t+1}^k - h p_{t+1}^{k+1/2}$.
 $p_{t+1}^{k+1} = p_{t+1}^{k+1/2} - (h/2) \nabla U(q_{t+1}^{k+1})$.
 end for
 With probability $\frac{\Pi(q_{t+1}^L, p_{t+1}^L)}{\Pi(q_{t+1}^0, p_{t+1}^0)} \wedge 1$, set $(q_{t+1}, p_{t+1}) = (q_{t+1}^L, p_{t+1}^L)$.
 Otherwise set $(q_{t+1}, p_{t+1}) = (q_{t+1}^0, p_{t+1}^0)$.
end for

A transition of this algorithm can be decomposed into two different (sub-)transitions:

- The first transition is $\begin{pmatrix} q_t \\ p_t \end{pmatrix} \rightarrow \begin{pmatrix} q_{t+1}^0 \\ p_{t+1}^0 \end{pmatrix}$ where one component is freezed $q_{t+1}^0 = q_t$, while the second one has been refreshed with $\mathcal{N}(0, I_d)$ which turns out to be also the conditional law $\Pi(\mathrm{d}p|q)|_{q=q_t}$ (see (8.2)). A move where one component is fixed whereas the second one is according to the conditional distribution of the target is actually a Gibbs move and as a Gibbs move, the transition $\begin{pmatrix} q_t \\ p_t \end{pmatrix} \rightarrow \begin{pmatrix} q_{t+1}^0 \\ p_{t+1}^0 \end{pmatrix}$ is Π -reversible (please, check it carefully).
- The second transition concerns $\begin{pmatrix} q_{t+1}^0 \\ p_{t+1}^0 \end{pmatrix} \rightarrow \begin{pmatrix} q_{t+1} \\ p_{t+1} \end{pmatrix}$. It consists in (a) constructing a candidate $\begin{pmatrix} q_{t+1}^L \\ p_{t+1}^L \end{pmatrix}$ deterministically from $\begin{pmatrix} q_{t+1}^0 \\ p_{t+1}^0 \end{pmatrix}$ using L steps, and then (b) accept or refuse the proposed candidate according to some well-chosen probability. It looks like a **MH transition with deterministic moves** and we then have to check carefully that it is Π -reversible.

8.3.1 MH with deterministic moves

We start with a very simple question: can we construct a MH algorithm with target Π and where the proposal candidate is deterministic: $Q(x, \mathrm{d}y) = \delta_{\varphi(x)}(\mathrm{d}y)$? Of course due to the Dirac mass, we are not in a dominated framework but the question is: can we use a generalized MH algorithm as described in Section 8.1?

If yes, then we have to see if (B1) is satisfied. Set

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

and write for any non-negative function h ,

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

Let us focus on the last term $\Pi(\mathrm{d}v)h(v, \varphi^{-1}(v))$. If we want to let appear the integral of $h(x, y)$ wrt to $\mu_0(\mathrm{d}x\mathrm{d}y) = \Pi(\mathrm{d}x)\delta_{\varphi(x)}(\mathrm{d}y)$, we need to assume that $\varphi^{-1}(v) = \varphi(v)$ that is φ is an **involution** and in such a case:

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

and the acceptance probability is then

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

Remark 8.2. (i) A first point is that if we only use the involution, then after two steps we land up to the initial state... Not very interesting... Therefore, this deterministic transition is often combined with another move that is not deterministic. (In the Leapfrog, the first move is not deterministic: while we freeze the position, we refresh the momentum according to a Normal distribution).

(ii) For any involution, you can get a Metropolis Hastings with a theoretical expression of the acceptance probability as

$$\frac{\mathrm{d}\Pi \circ \varphi^{-1}}{\mathrm{d}\Pi}(x) \wedge 1$$

but the ideal HMC goes one step further since we can show that this is equal to 1. To get this, if we work on \mathbb{R}^d and if Π has density wrt the Lebesgue measure that we still denote Π , we get

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

where the second term $\left| \frac{\partial \varphi^{-1}(x)}{\partial x} \right|$ 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 $\frac{\Pi(\varphi^{-1}(x))}{\Pi(x)}$ is one if the involution stays on **the same level set** (ie the moves according to φ does not change the value of Π) and 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

$$\frac{\Pi(\varphi^{-1}(x))}{\Pi(x)} = \frac{\Pi(\varphi(x))}{\Pi(x)} \quad \text{since} \quad \varphi \circ \varphi = \text{I}$$

MH with deterministic moves is possible but the mapping φ should be an involution. If the mapping should stay on the same level set and is volume preserving, the acceptance probability is even exactly equal to one.

8.3.2 Hamiltonian dynamics

8.3.2.1 Level sets

How can we find deterministic moves that stays on the same level set, ie, such that Π and consequently, $H = U + K$ (the Hamiltonian) is constant.

If we now let (p, q) depend 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 gives the idea of using the following dynamics: for all $i \in [1 : d]$,

$$\begin{aligned} \frac{\partial H}{\partial q^{t,i}}(q^t, p^t) &= \frac{\partial U(q^t)}{\partial q^{t,i}} = -\frac{dp^{t,i}}{dt} \\ \frac{\partial H}{\partial p^{t,i}}(q^t, p^t) &= \frac{\partial K(p^t)}{\partial p^{t,i}} = p^{t,i} = \frac{dq^{t,i}}{dt} \quad \blacktriangleright \textbf{(Hamiltonian dynamics)} \end{aligned} \quad (8.3)$$

The very last equation leads to the interpretation of $p^{t,i}$ as a speed since it is the derivative of the position wrt time. Note $\phi^t(q, p) = (q^t, p^t)$ the (deterministic) position and momentum at time t when (q^s, p^s) follows the Hamiltonian dynamics (8.3). So, Hamiltonian dynamics moves along the **same level sets**. It can also be shown that it is **volume-preserving**. But unfortunately, it is not an involution. That's the bad news. But there is also a good news: we can add a flip mapping on the second component after a move so that the resulting mapping is an involution. We will see it in the next paragraph.

8.3.2.2 The flip operator trick and the involution

Denote by $s(q, p) = (q, -p)$ the flip operator on the second component. The flip is also volume-preserving and moves along the level set so that finally, setting $f^T = s \circ \phi^T$, we obtain that f^T is volume-preserving and level-set invariant. We now show that f^T is an involution. Indeed, write $f^T(q, p) = (q^T, -p^T)$. To see what we obtain by applying again f^T , set $\tilde{q}^t = q^{T-t}$ and $\tilde{p}^t = -p^{T-t}$ so that $(\tilde{q}^0, \tilde{p}^0) = f^T(q, p)$. Then,

$$\begin{aligned} \frac{d\tilde{q}^{t,i}}{dt} &= \frac{dq^{T-t,i}}{dt} = -\left. \frac{dq^{s,i}}{ds} \right|_{s=T-t} = -p^{T-t,i} = \tilde{p}^{t,i} \\ \frac{d\tilde{p}^{t,i}}{dt} &= -\frac{dp^{T-t,i}}{dt} = -\left. \frac{dp^{s,i}}{ds} \right|_{s=T-t} = -\frac{\partial U(q^{T-t})}{\partial q^{T-t,i}} = -\frac{\partial U(\tilde{q}^{T-t})}{\partial \tilde{q}^{T-t,i}} \end{aligned}$$

Finally, the process $(\tilde{q}^t, \tilde{p}^t)$ follows the Hamiltonian dynamics so that $f^T(\tilde{q}^0, \tilde{p}^0) = (\tilde{q}^T, -\tilde{p}^T)$ and by definition this quantity is equal to (q^0, p^0) . We finally obtain that f^T is an involution.

The ideal HMC can be described as follows (Algorithm 8.3.2.2). We see in this ideal algorithm that the candidate is always accepted and that we don't even need to apply the flip operator (since it does not change at all the algorithm).

Algorithm 8 The ideal HMC.

At $t = 0$, draw X_0 according to some arbitrary distribution.

for $t = 0 \rightarrow n - 1$ **do**

Set $q_{t+1}^0 = q_t$ and draw $p_{t+1}^0 \sim \mathcal{N}(0, I_d)$.

Set $(q_{t+1}^L, p_{t+1}^L) = \phi^T(q_{t+1}^0, p_{t+1}^0)$.

Set $(q_{t+1}, p_{t+1}) = (q_{t+1}^L, p_{t+1}^L)$.

end for

This is just an ideal algorithm since we don't know how to solve exactly Hamiltonian dynamics. The leapfrog is based on an approximation of these dynamics.

8.3.3 The leapfrog integrator

8.3.3.1 Discretization and volume-preserving property

Recall that the Hamiltonian dynamics are: for all $i \in [1 : d]$,

$$\frac{\partial U(q^t)}{\partial q^{t,i}} = -\frac{dp^{t,i}}{dt}, \quad \text{and} \quad p^{t,i} = \frac{dq^{t,i}}{dt}$$

A first idea of discretization would be: choose a small stepsize h and a number of steps L and move according to: (for $k \in [1 : L]$),

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

Unfortunately, this discretization is associated to a mapping $(q_{k+1}, p_{k+1}) = \varphi(q_k, p_k)$ that is not volume-preserving... That is the absolute value of the Jacobian determinant of φ is not equal to one.

Instead, for any differentiable function ψ , the Jacobian matrix of a mapping where one component is freezed and the other one is updated by an additive term: $(x, y) \mapsto (x, y + \psi(x))$ is given by $\begin{pmatrix} 1 & 0 \\ \star & 1 \end{pmatrix}$ so that the determinant is one and this mapping is volume preserving...

Now, we will focus on the leapfrog discretization. Interested readers in other discretizations may try to solve Exercise 9.13 which focus on another discretization scheme.

The leapfrog discretization is defined by the following scheme: for $k \in [1 : L]$

$$\begin{aligned} p^{k+1/2} &= p^k - (h/2)\nabla U(q^k) \\ q^{k+1} &= q^k + hp^{k+1/2} \\ p^{k+1} &= p^{k+1/2} - (h/2)\nabla U(q^{k+1}) \end{aligned} \tag{8.4}$$

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

$$(q^k, p^k) \xrightarrow{\varphi_1} (q^k, p^{k+1/2}) \xrightarrow{\varphi_2} (q^{k+1}, p^{k+1/2}) \xrightarrow{\varphi_1} (q^{k+1}, p^{k+1}). \tag{8.5}$$

where

$$\varphi_1(x, y) = (x, y - (h/2)\nabla U(x)) \quad \text{and} \quad \varphi_2(x, y) = (x + hy, y) \tag{8.6}$$

Each of these mappings keep one component freezed while the other component is updated with an additive term, so 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 but not level-set invariant, so the acceptance probability will not be equal to 1, we still hope this is still high since it is an approximation of the Hamiltonian...

A very last property on the Leapfrog HMC should be checked (in order to say that the second step in Algorithm 8.3 is indeed a MH with deterministic moves): the involution property.

8.3.3.2 The flip operator trick and the involution property

To be specific, if we add the flip mapping s on the second component, then do we obtain an involution as for the ideal HMC?

Lemma 8.3. *For any $L \geq 1$, write $\Phi^{h,L}$ the Leapfrog mapping and define $f^{h,L} = s \circ \Phi^{h,L}$. Then $f^{h,L}$ is an involution.*

Proof. We will show that $f^{h,L} \circ f^{h,L} = Id$ by induction on L .

(i) We first show that $f^{h,1}$ is an involution.

Using (9.2), (8.5) and (8.6), we can check that

$$(q^k, p^k) \xrightarrow{\varphi_1} (q^k, p^{k+1/2}) \xrightarrow{\varphi_2} (q^{k+1}, p^{k+1/2}) \xrightarrow{\varphi_1} (q^{k+1}, p^{k+1}) \xrightarrow{s} (q^{k+1}, -p^{k+1}).$$

and

$$(q^{k+1}, -p^{k+1}) \xrightarrow{\varphi_1} (q^{k+1}, -p^{k+1/2}) \xrightarrow{\varphi_2} (q^k, -p^{k+1/2}) \xrightarrow{\varphi_1} (q^k, -p^k) \xrightarrow{s} (q^k, p^k).$$

Therefore: $f^{h,1} = s \circ \Phi^{h,1} = s \circ \varphi_1 \circ \varphi_2 \circ \varphi_1$ is an involution, i.e.

$$s \circ \Phi^{h,1} \circ s \circ \Phi^{h,1} = Id$$

Moreover, applying s on both sides of the equation and noting that s is an involution, we get $\Phi^{h,1} \circ s \circ \Phi^{h,1} = s$ (this will be useful two lines below).

(ii) Assume that for some $L \geq 1$, $f^{h,L}$ is an involution, that is $f^{h,L} \circ f^{h,L} = Id$. Now, write

$$f^{h,L+1} \circ f^{h,L+1} = s \circ \Phi^{h,L} \circ \underbrace{\Phi^{h,1} \circ s \circ \Phi^{h,1}}_s \circ \Phi^{h,L} = f^{h,L} \circ f^{h,L} = Id$$

by the induction assumption.

This completes the proof. \square

8.4 Data augmentation

Throughout this section, (X, \mathcal{X}) and (Y, \mathcal{Y}) are Polish spaces equipped with their Borel σ -fields. Again, we wish to simulate from a probability measure π defined on (X, \mathcal{X}) using a sequence $\{X_k, k \in \mathbb{N}\}$ of X -valued random variables. Data augmentation algorithms consist in writing the target distribution π as the marginal of the distribution π^* on the product space $(X \times Y, \mathcal{X} \otimes \mathcal{Y})$ defined by $\pi^* = \pi \otimes R$ where R is a kernel on $X \times Y$. There exists also a kernel S on $Y \times X$ and a probability measure $\tilde{\pi}$ on (Y, \mathcal{Y}) such that $\pi^*(C) = \iint \mathbf{1}_C(x, y) \tilde{\pi}(dy) S(y, dx)$ for $C \in \mathcal{X} \otimes \mathcal{Y}$. In other words, if (X, Y) is a pair of random variables with distribution π^* , then $R(x, \cdot)$ is the distribution of Y conditionally on $X = x$ and $S(y, \cdot)$ is the distribution of X conditionally on $Y = y$. The bivariate distribution π^* can then be expressed as follows

$$\pi^*(dxdy) = \pi(dx)R(x, dy) = S(y, dx)\tilde{\pi}(dy). \quad (8.7)$$

A data augmentation algorithm consists in running a Markov Chain $\{(X_k, Y_k), k \in \mathbb{N}\}$ with invariant probability π^* and to use $n^{-1} \sum_{k=0}^{n-1} f(X_k)$ as an approximation of $\pi(f)$. A significant difference between this general approach and a Metropolis-Hastings algorithm associated to the target distribution π is that $\{X_k, k \in \mathbb{N}\}$ is no longer constrained to be a Markov chain. The transition from (X_k, Y_k) to (X_{k+1}, Y_{k+1}) is decomposed into two successive steps: Y_{k+1} is first drawn given (X_k, Y_k) and then X_{k+1} is drawn given (X_k, Y_{k+1}) . Intuitively, Y_{k+1} can be used as an auxiliary variable, which directs the moves of X_k toward interesting regions with respect to the target distribution.

When sampling from R and S is feasible, a classical choice consists in following the two successive steps: given (X_k, Y_k) ,

- (i) sample Y_{k+1} from $R(X_k, \cdot)$,
- (ii) sample X_{k+1} from $S(Y_{k+1}, \cdot)$.

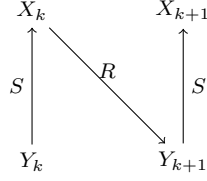


Fig. 8.1 In this example, sampling from R and S is feasible.

It turns out that $\{X_k, k \in \mathbb{N}\}$ is a Markov chain with Markov kernel RS and π is reversible wrt RS .

Lemma 8.4. *The distribution π is reversible with respect to the kernel RS .*

Proof. We must prove that the measure $\pi \otimes RS$ on \mathbf{X}^2 is symmetric. For $A, B \in \mathcal{X}$, applying (8.7), we have

$$\begin{aligned} \pi \otimes RS(A \times B) &= \int_{\mathbf{X} \times \mathbf{Y}} \pi(\mathrm{d}x) R(x, \mathrm{d}y) \mathbf{1}_A(x) S(y, B) = \int_{\mathbf{X} \times \mathbf{Y}} \mathbf{1}_A(x) S(y, B) \pi^*(\mathrm{d}x \mathrm{d}y) \\ &= \int_{\mathbf{X} \times \mathbf{Y}} \mathbf{1}_A(x) S(y, B) S(y, \mathrm{d}x) \tilde{\pi}(\mathrm{d}y) = \int_{\mathbf{Y}} S(y, A) S(y, B) \tilde{\pi}(\mathrm{d}y). \end{aligned}$$

This proves that $\pi \otimes RS$ is symmetric. \square

Assume now that sampling from R or S is infeasible. In this case, we consider two instrumental kernels Q on $(\mathbf{X} \times \mathbf{Y}) \times \mathcal{Y}$ and T on $(\mathbf{X} \times \mathbf{Y}) \times \mathcal{X}$ which will be used to propose successive candidates for Y_{k+1} and X_{k+1} . For simplicity, assume that $R(x, \mathrm{d}y')$ and $Q(x, y; \mathrm{d}y')$ (resp. $S(y', \mathrm{d}x')$ and $T(x, y'; \mathrm{d}x')$) are dominated by the same measure and call r and q (resp. s and t) the associated transition densities. We assume that r and s are known up to a normalizing constant. Define the Markov chain $\{(X_k, Y_k), k \in \mathbb{N}\}$ as follows. Given $(X_k, Y_k) = (x, y)$,

- (DA1) draw a candidate \tilde{Y}_{k+1} according to the distribution $Q(x, y; \cdot)$ and accept $Y_{k+1} = \tilde{Y}_{k+1}$ with probability $\alpha(x, y, \tilde{Y}_{k+1})$ defined by

$$\alpha(x, y, y') = \frac{r(x, y') q(x, y; y')}{r(x, y) q(x, y; y')} \wedge 1;$$

otherwise, set $Y_{k+1} = Y_k$; the Markov kernel on $\mathbf{X} \times \mathbf{Y} \times \mathcal{Y}$ associated to this transition is denoted by K_1 ;

- (DA2) draw then a candidate \tilde{X}_{k+1} according to the distribution $T(x, Y_{k+1}; \cdot)$ and accept $X_{k+1} = \tilde{X}_{k+1}$ with probability $\beta(x, Y_{k+1}, \tilde{X}_{k+1})$ defined by

$$\beta(x, y, x') = \frac{s(y, x') t(x', y; x)}{s(y, x) t(x, y; x')} \wedge 1;$$

otherwise, set $X_{k+1} = X_k$; the Markov kernel on $\mathbf{X} \times \mathbf{Y} \times \mathcal{X}$ associated to this transition is denoted by K_2 .

For $i = 1, 2$, let K_i^* be the kernels associated to K_1 and K_2 as follows: for $x \in \mathbf{X}$, $y \in \mathbf{Y}$, $A \in \mathcal{X}$ and $B \in \mathcal{Y}$,

$$K_1^*(x, y; A \times B) = \mathbf{1}_A(x) K_1(x, y; B). \quad (8.8)$$

$$K_2^*(x, y; A \times B) = \mathbf{1}_B(y) K_2(x, y; A). \quad (8.9)$$

Then, the kernel of the chain $\{(X_n, Y_n), n \in \mathbb{N}\}$ is $K = K_1^* K_2^*$. The process $\{X_n, n \in \mathbb{N}\}$ is in general not a Markov chain since the distribution of X_{k+1} conditionally on (X_k, Y_k) depends

on (X_k, Y_k) and on X_k only, except in some special cases. Obviously, this construction includes the previous one where sampling from R and S was feasible. Indeed, if $Q(x, y; \cdot) = R(x, \cdot)$ and $T(x, y; \cdot) = S(x, \cdot)$, then the acceptance probabilities α and β defined above simplify to one, the candidates are always accepted and we are back to the previous algorithm.

Proposition 8.5. *The extended target distribution π^* is reversible wrt the kernels K_1^* and K_2^* and invariant with respect to K .*

Proof. The reversibility of π^* with respect to K_1^* and K_2^* implies its invariance and consequently its invariance with respect to the product $K = K_1^* K_2^*$. Let us prove the reversibility of π^* with respect to K_1^* . For each $x \in \mathbf{X}$, the kernel $K_1(x, \cdot; \cdot)$ on $\mathbf{Y} \times \mathbf{Y}$ is the kernel of a Metropolis-Hastings algorithm with target density $r(x, \cdot)$, proposal kernel density $q(x, \cdot; \cdot)$ and acceptance probability $\alpha(x, \cdot, \cdot)$. It implies that the distribution $R(x, \cdot)$ is reversible with respect to the kernel $K_1(x, \cdot; \cdot)$. Applying the definition (8.8) of K_1^* and $\pi^* = \pi \otimes R$ yields, for $A, C \in \mathcal{X}$ and $B, D \in \mathcal{Y}$,

$$\begin{aligned} \pi^* \otimes K_1^*(A \times B \times C \times D) &= \iint_{A \times B} \pi(\mathrm{d}x \mathrm{d}y) K_1^*(x, y; C \times D) \\ &= \iint_{A \times B} \pi(\mathrm{d}x) R(x, \mathrm{d}y) \mathbf{1}_C(x) K_1(x, y, D) \\ &= \int_{A \cap C} \pi(\mathrm{d}x) [R(x, \cdot) \otimes K_1(x, \cdot; \cdot)](B \times D). \end{aligned}$$

We have seen that for each $x \in \mathbf{X}$, the measure $R(x, \cdot) \otimes K_1(x, \cdot; \cdot)$ is symmetric, thus $\pi^* \otimes K^*$ is also symmetric. The reversibility of π^* with respect to K_2^* is proved similarly. \square

Example 8.6 (The slice sampler). Set $\mathbf{X} = \mathbb{R}^d$ and $\mathcal{X} = \mathcal{B}(\mathbf{X})$. Let μ be a σ -finite measure on $(\mathbf{X}, \mathcal{X})$ and let h be the density with respect to μ of the target distribution. We assume that for all $x \in \mathbf{X}$,

$$h(x) = C \prod_{i=0}^k f_i(x),$$

where C is a constant (which is not necessarily known) and $f_i : \mathbb{R}^d \rightarrow \mathbb{R}^+$ are nonnegative measurable functions. For $y = (y_1, \dots, y_k) \in \mathbb{R}^{+k}$, define

$$L(y) = \{x \in \mathbb{R}^d : f_i(x) \geq y_i, i = 1, \dots, k\}.$$

The f_0 -slice-sampler algorithm proceeds as follows:

- given X_n , draw independently a k -tuple $Y_{n+1} = (Y_{n+1,1}, \dots, Y_{n+1,k})$ of independent random variables such that $Y_{n+1,i} \sim \text{Unif}(0, f_i(X_n))$, $i = 1, \dots, k$.
- sample X_{n+1} from the distribution with density proportional to $f_0 \mathbf{1}_{L(Y_{n+1})}$.

Set $\mathbf{Y} = (\mathbb{R}^+)^k$ and for $(x, y) \in \mathbf{X} \times \mathbf{Y}$,

$$h^*(x, y) = C f_0(x) \mathbf{1}_{L(y)}(x) = h(x) \prod_{i=1}^k \frac{\mathbf{1}_{[0, f_i(x)]}(y_i)}{f_i(x)}.$$

Let π^* be the probability measure with density h^* with respect to Lebesgue's measure on $\mathbf{X} \times \mathbf{Y}$. Then $\int_{\mathbf{Y}} h^*(x, y) \mathrm{d}y = h(x)$ i.e. π is the first marginal of π^* . Let R be the kernel on $\mathbf{X} \times \mathbf{Y}$ with kernel density r defined by

$$r(x, y) = \frac{h^*(x, y)}{h(x)} \mathbf{1}_{h(x) > 0}.$$

Then $\pi^* = \pi \otimes R$. Define the distribution $\tilde{\pi} = \pi R$, its density $\tilde{h}(y) = \int_{\mathbf{X}} h^*(u, y) du$ and the kernel S on $\mathbf{Y} \times \mathcal{X}$ with density s by

$$s(y, x) = \frac{h^*(x, y)}{\tilde{h}(y)} \mathbf{1}_{\tilde{h}(y) > 0}.$$

If (X, Y) is a vector with distribution π^* , then $S(y, \cdot)$ is the conditional distribution of X given $Y = y$ and the Markov kernel of the chain $\{X_n, n \in \mathbb{N}\}$ is RS and Lemma 8.4 can be applied to prove that π is reversible, hence invariant, with respect to RS .

8.4.1 Two-stage Gibbs sampler

The Gibbs sampler is a simple method which decomposes a complex multidimensional distribution into a collection of smaller dimensional ones. Let $(\mathbf{X}, \mathcal{X})$ and $(\mathbf{Y}, \mathcal{Y})$ be complete separable metric spaces endowed with their Borel σ -fields. To construct the Markov chain $\{(X_n, Y_n), n \in \mathbb{N}\}$ with π^* as an invariant probability, we proceed exactly as in data-augmentation algorithms. Assume that π^* may be written as

$$\pi^*(dx dy) = \pi(dx)R(x, dy) = \tilde{\pi}(dy)S(y, dx) \quad (8.10)$$

where π and $\tilde{\pi}$ are probability measures on \mathbf{X} and \mathbf{Y} respectively and R and S are kernels on $\mathbf{X} \times \mathbf{Y}$ and $\mathbf{Y} \times \mathcal{X}$ respectively.

The deterministic updating (two-stage) Gibbs (DUGS) sampler

When sampling from R and S is feasible, the DUGS sampler proceeds as follows: given (X_k, Y_k) ,

- (DUGS1) sample Y_{k+1} from $R(X_k, \cdot)$,
- (DUGS2) sample X_{k+1} from $S(Y_{k+1}, \cdot)$.

For both the Data Augmentation algorithms and the two-stage Gibbs sampler we consider a distribution π^* on the product space $\mathbf{X} \times \mathbf{Y}$. In the former case, the distribution of interest is a marginal distribution of π^* and in the latter case the target distribution is π^* itself.

We may associate to each update (DUGS1)-(DUGS2) of the algorithm a transition kernel on $(\mathbf{X} \times \mathbf{Y}) \times (\mathcal{X} \otimes \mathcal{Y})$ defined for $(x, y) \in \mathbf{X} \times \mathbf{Y}$ and $A \times B \in \mathcal{X} \otimes \mathcal{Y}$ by

$$R^*(x, y; A \times B) = \mathbf{1}_A(x)R(x, B), \quad (8.11)$$

$$S^*(x, y; A \times B) = \mathbf{1}_B(y)S(y, A). \quad (8.12)$$

The transition kernel of the DUGS is then given by

$$P_{\text{DUGS}} = R^* S^*. \quad (8.13)$$

Note that for $A \times B \in \mathcal{X} \otimes \mathcal{Y}$,

$$\begin{aligned} P_{\text{DUGS}}(x, y; A \times B) &= \iint_{\mathbf{X} \times \mathbf{Y}} R^*(x, y; dx' dy') S^*(x', y'; A \times B) \\ &= \iint_{\mathbf{X} \times \mathbf{Y}} R(x, dy') \mathbf{1}_B(y') S(y', A) \\ &= \int_B R(x, dy') S(y', A) = R \otimes S(x, B \times A). \end{aligned} \quad (8.14)$$

As a consequence of Proposition 8.5, we obtain the invariance of π^* .

Lemma 8.7. *The distribution π^* is reversible with respect to the kernels R^* and S^* and invariant with respect to P_{DUGS} .*

The Random Scan Gibbs sampler (RSGS)

At each iteration, the RSGS algorithm consists in updating one component chosen at random. It proceeds as follows: given (X_k, Y_k) ,

- (RSGS1) sample a Bernoulli random variable B_{k+1} with probability of success $1/2$.
- (RSGS2) If $B_{k+1} = 0$, then sample Y_{k+1} from $R(X_k, \cdot)$ else sample X_{k+1} from $S(Y_{k+1}, \cdot)$.

The transition kernel of the RSGS algorithm can be written

$$P_{\text{RSGS}} = \frac{1}{2}R^* + \frac{1}{2}S^*. \quad (8.15)$$

Lemma 8.7 implies that P_{RSGS} is reversible wrt π^* and therefore π^* is invariant for P_{RSGS} .

If sampling from R or S is infeasible, the Gibbs transitions can be replaced by a Metropolis-Hastings algorithm on each component as in the case of the DUGS algorithm. The algorithm is then called the Two-Stage Metropolis-within-Gibbs algorithm.

Illustrations, exercises, extensions

9.1 Illustrations

A “collaboratory” **Jupyter Notebook** that illustrates the results of some exercises or some algorithms seen in the course is at the disposal of the reader by following this link.

9.1.1 Illustrations of HMC

Figure 9.1 displays several trajectories of the leapfrog integrator when the target density is a Gaussian distribution or a mixture of Gaussian distributions. Trajectories are initialized randomly and then the leapfrog integrator is run with step size $h = 0.01$.

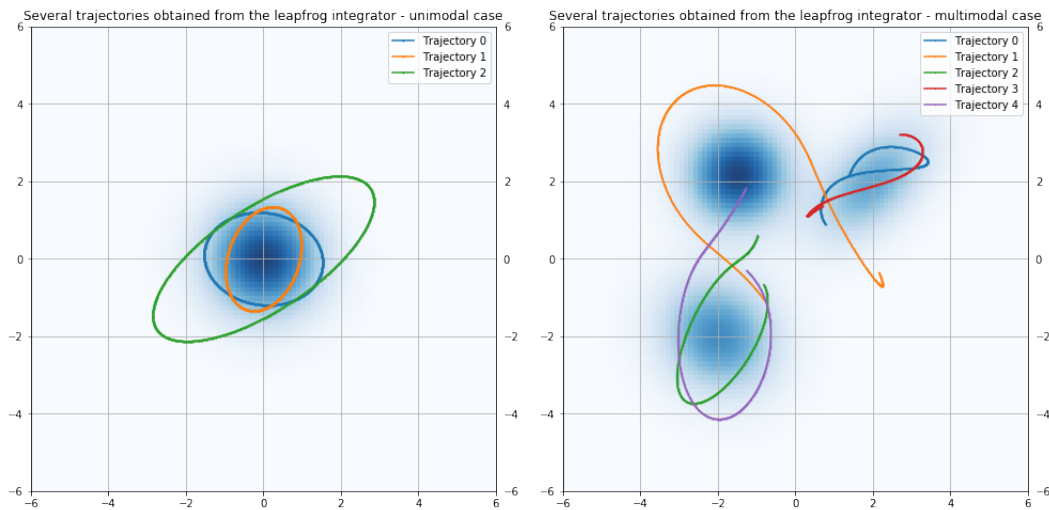
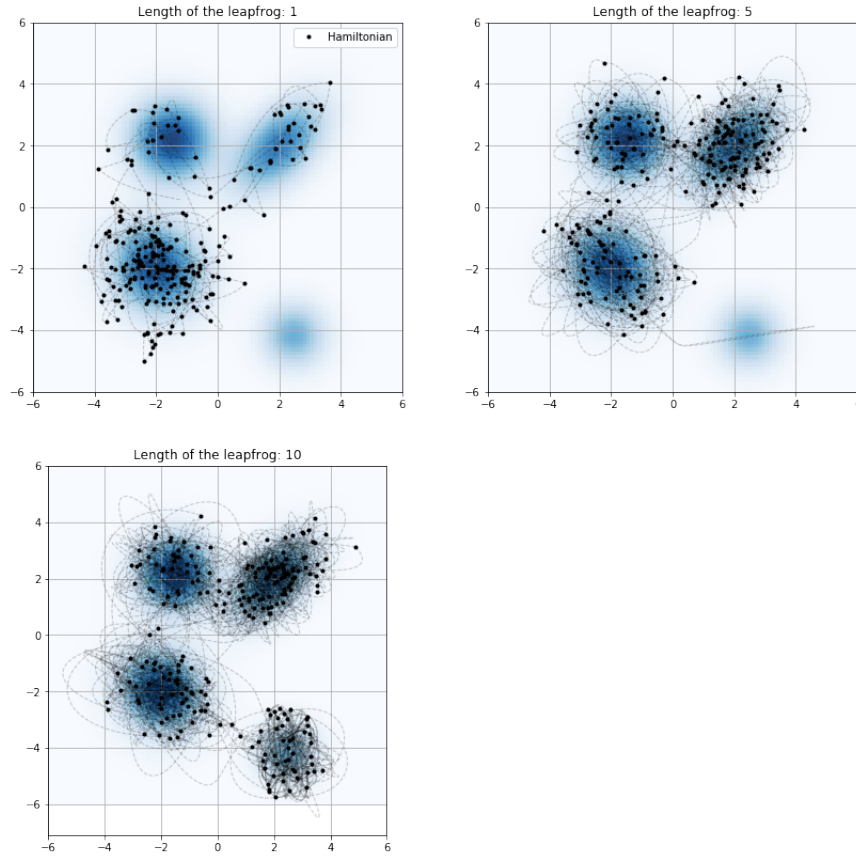


Figure 9.1.1 displays several trajectories of a HMC sampler with leapfrog integrators with various lengths when the target density is a mixture of Gaussian distributions. This figure highlights the fact that the proposed moves provided by the leapfrog integrator allow to explore widely the space and to jump from a mode to another.



9.2 Exercises

Exercise 9.1. Let

$$X_t = \sigma_t Z_t, \quad \sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2, \quad t \geq 1$$

where the coefficients α_0, α_1 are positive and $\{Z_t : t \in \mathbb{N}\}$ is an iid sequence of random variables such that $\mathbb{E}[Z_0] = 0$, $\mathbb{E}[Z_0^2] = 1$, and $\{Z_t : t \in \mathbb{N}\}$ is independent of X_0

1. Assuming that Z_0 has the density q wrt the Lebesgue measure, show that $\{X_t : t \in \mathbb{N}\}$ is a Markov chain with transition density

$$p(x, x') = \frac{1}{\sqrt{\alpha_0 + \alpha_1 x^2}} q\left(\frac{x'}{\sqrt{\alpha_0 + \alpha_1 x^2}}\right).$$

Exercise 9.2. Let $\mu \in \mathbf{M}_1(\mathbf{X})$ and assume that for some function $h \in \mathbf{F}(\mathbf{X})$, and some constant C ,

$$\lim_{n \rightarrow \infty} n^{-1} \sum_{k=0}^{n-1} h(X_k) = C, \quad \mathbb{P}_\mu - a.s.$$

Then show that for μ -almost all $x \in \mathbf{X}$,

$$\lim_{n \rightarrow \infty} n^{-1} \sum_{k=0}^{n-1} h(X_k) = C, \quad \mathbb{P}_x - a.s.$$

Exercise 9.3. In this exercise, we will show the strong Markov property: for any $\nu \in \mathbf{M}_1(\mathbf{X})$, any non-negative or bounded function h on $\mathbf{X}^{\mathbb{N}}$, any $n \in \mathbb{N}$ and any stopping time σ ,

$$\mathbb{E}_{\nu} [h \circ S^{\sigma} \mathbf{1}_{\{\sigma < \infty\}} | \mathcal{F}_{\sigma}] = \mathbb{E}_{X_{\sigma}} [h], \quad \mathbb{P}_{\nu} - a.s.$$

where $\mathcal{F}_{\sigma} = \{A \in \mathcal{F} : A \cap \{\sigma = k\} \in \mathcal{F}_k, \forall k \geq 1\}$ and $\mathcal{F}_k = \sigma(X_{0:k})$.

For any filtration (\mathcal{F}_k) on \mathcal{F} , recall that σ is a (\mathcal{F}_k) -stopping time if for all $n \in \mathbb{N}$, $\{\sigma \leq n\} \in \mathcal{F}_n$.

1. Define $\mathcal{F}_{\sigma} = \{A \in \mathcal{F} : A \cap \{\sigma = k\} \in \mathcal{F}_k, \forall k \geq 1\}$. Show that \mathcal{F}_{σ} is a σ -field.
2. Using the decomposition, $\mathbf{1}_{\{\sigma < \infty\}} = \sum_{k=0}^{\infty} \mathbf{1}_{\{\sigma = k\}}$, show the strong Markov property.

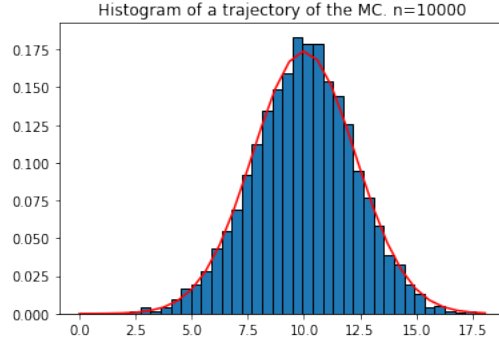
Exercise 9.4. Consider a Gaussian AR(1) process, $X_t = \mu + \phi X_{t-1} + \sigma Z_t$, where $\{Z_t : t \in \mathbb{N}\}$ is an iid sequence of standard Gaussian random variables, independent of X_0 . Assume that $|\phi| < 1$ and that X_0 is Gaussian with mean μ_0 and variance γ_0^2 .

1. Show that if X_1 has the same distribution as X_0 then

$$\begin{cases} \mu + \phi \mu_0 = \mu_0 \\ \phi^2 \gamma_0^2 + \sigma^2 = \gamma_0^2 \end{cases}$$

2. Deduce an invariant distribution for $(X_t)_{t \in \mathbb{N}}$.

Details of the numerics are given in the Jupyter Notebook.



Exercise 9.5. Consider a Markov chain whose state space $\mathbf{X} = (0, 1)$ is the open unit interval. If the chain is at x , it picks one of the two intervals $(0, x)$ or $(x, 1)$ with equal probability $1/2$, and then moves to a point y which is uniformly distributed in the chosen interval.

1. Show that this Markov chain has a transition density wrt the Lebesgue measure on the interval $(0, 1)$, which is given by

$$k(x, y) = \frac{1}{2} \frac{1}{x} \mathbf{1}_{(0, x)}(y) + \frac{1}{2} \frac{1}{1-x} \mathbf{1}_{(x, 1)}(y)$$

2. Show that this Markov chain can be equivalently represented as an iterated random sequence.

$$X_t = \epsilon_t [X_{t-1} U_t] + (1 - \epsilon_t) [X_{t-1} + U_t(1 - X_{t-1})]$$

where $\{U_n : n \in \mathbb{N}\}$ and $\{\epsilon_n : n \in \mathbb{N}\}$ are iid random variables whose distribution should be given.

3. Assuming that the stationary distribution has a density p wrt the Lebesgue measure show that

$$p(y) = \frac{1}{2} \int_y^1 \frac{p(x)}{x} dx + \frac{1}{2} \int_0^y \frac{p(x)}{1-x} dx$$

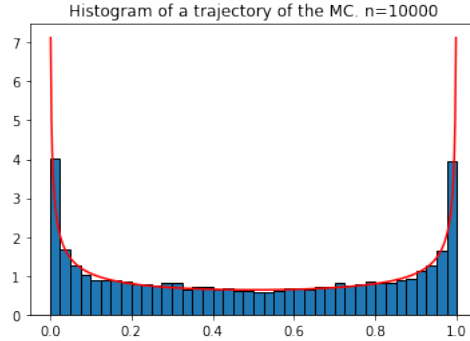
4. Deduce that

$$\int_0^z p(y)dy = 2C \arcsin(\sqrt{z})$$

for some constant C .

5. Conclude that $C = 1/\pi$.

Details of the numerics are given in the Jupyter Notebook.



Exercise 9.6. Let $\pi(dx dy) = \pi(x, y) \lambda_X(dx) \lambda_Y(dy)$ be probability measure on $(X \times Y, \mathcal{X} \otimes \mathcal{Y})$ where λ_X (resp. λ_Y) is σ -finite measure on (X, \mathcal{X}) (resp. (Y, \mathcal{Y})). Define $\pi(y|x) = \frac{\pi(x, y)}{\pi(x)}$ whenever $\pi(x) \neq 0$.

Define: $P((x, y), dx' dy') = \delta_x(dx') \pi(y'|x) \lambda_Y(dy')$. Show that P is a π -reversible Markov kernel on $(X \times Y, \mathcal{X} \otimes \mathcal{Y})$.

Exercise 9.7. In a MH algorithm, we want to find an acceptance probability $\alpha(x, y) = f\left(\frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right)$

1. Show that the detailed balance condition is satisfied if and only if for all $u \geq 0$, $f(u) = uf(1/u)$.
2. Show that it is sufficient to check that for all $u \in (0, 1)$, $f(u) = uf(1/u)$.
3. Find all the functions f such that the detailed balance condition is satisfied with $\alpha(x, y) = f\left(\frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right)$.

Exercise 9.8. Let P be a Markov kernel with invariant probability measure π . Let $A \in \mathcal{X}$ such that $\pi(A) = 1$. Show that there exists $B \subset A$ such that $P(x, B) = 1$ for all $x \in B$ (i.e. such set B is said to be *absorbing* in the sense that starting from any point in B , the Markov chain stays in B forever with probability one).

Exercise 9.9. Let (X_t) be the AR(p) process defined by: $X_t = \sum_{i=1}^p a_i X_{t-i} + \sigma \epsilon_t$ for all $t \geq i$, where $\epsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ and $\sigma > 0$.

Define $Y_t = \begin{pmatrix} X_t \\ \vdots \\ X_{t-p+1} \end{pmatrix}$ and show that (Y_t) is a Markov chain. Give the expression of the associated Markov kernel. Show that it admits at most one invariant probability measure.

Exercise 9.10. Let P be a Markov kernel on $X \times X$, admitting an invariant probability measure π . We assume that there exist two measurable positive functions V and f and a constant K such that

$$PV + f \leq V + K.$$

Show that $\pi(f) < \infty$.

Exercise 9.11. In this exercise, we prove Theorem 6.4. Set $S_n = \sum_{k=0}^{n-1} h \circ T^k$, $L_n = \inf \{S_k : k \in [1 : n]\}$ and $A = \{\inf_{n \in \mathbb{N}^*} L_n = -\infty\}$. We first assume that $\mathbb{E}[h] > 0$.

1. Show that $L_n \geq h + \inf(0, L_n \circ T)$.

2. Deduce that

$$\mathbb{E}[\mathbf{1}_A h] \leq \mathbb{E}[\mathbf{1}_A (L_n)^+]$$

3. Deduce $\mathbb{P}(A) = 0$.

4. Prove The Birkhoff theorem (Theorem 6.4).

Exercise 9.12. Assume that (A1) and (A2) hold for some measurable function $V \geq 1$. We want to prove that: for all initial distributions $\nu \in \mathbf{M}_1(X)$ and all $f \in \mathbf{F}(X)$ such that $\pi(|f|) = \int_X \pi(dx) |f(x)| < \infty$,

$$\lim_{n \rightarrow \infty} n^{-1} \sum_{k=0}^{n-1} f(X_k) = \pi(f), \quad \mathbb{P}_\nu - a.s \quad (9.1)$$

1. Prove the result by combining Corollary 7.7 and Theorem 7.5 with some results in Chapter 3 (give the exact references of the results that you pick from Chapter 3)

Exercise 9.13. Consider the following discretization of the Hamiltonian dynamics.

$$\begin{aligned} p^{k+1} &= p^k - (h) \nabla U(q^k) \\ q^{k+1} &= q^k + h p^{k+1} \end{aligned} \quad (9.2)$$

1. Is it volume-preserving?

2. Can we to the flip operator trick to obtain an involution?

References

- Blundell et al., 2015. Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR.
- 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.
- Burda et al., 2015. Burda, Y., Grosse, R., and Salakhutdinov, R. (2015). Importance weighted autoencoders. *International Conference on Learning Representations*.
- Chen et al., 2023. Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. R. (2023). Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions.
- Conforti et al., 2023. Conforti, G., Durmus, A., and Silveri, M. G. (2023). Score diffusion models without early stopping: finite fisher information is all you need.
- De Bortoli et al., 2021. De Bortoli, V., Thornton, J., Heng, J., and Doucet, A. (2021). Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709.
- Durmus and Moulines, 2017. Durmus, A. and Moulines, E. (2017). Nonasymptotic convergence analysis for the unadjusted langevin algorithm.
- 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.
- 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.
- Ho et al., 2020. Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*.
- Hyvärinen and Dayan, 2005. Hyvärinen, A. and Dayan, P. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4).
- Karatzas and Shreve, 2012. Karatzas, I. and Shreve, S. (2012). *Brownian motion and stochastic calculus*, volume 113. Springer Science & Business Media.
- Kingma and Welling, 2013. Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Kingma et al., 2019. Kingma, D. P., Welling, M., et al. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392.

- 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.
- Ng and Jordan, 2001. Ng, A. and Jordan, M. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14.
- 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.
- 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.
- Sohl-Dickstein et al., 2015. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France. PMLR.
- Song and Ermon, 2019. Song, Y. and Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.
- Song and Kingma, 2021. Song, Y. and Kingma, D. P. (2021). How to train your energy-based models. *arXiv:2101.03288*.
- Song et al., 2021. Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2021). Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations (ICLR)*.
- Tang and Yang, 2021. Tang, R. and Yang, Y. (2021). On empirical bayes variational autoencoder: An excess risk bound. In Belkin, M. and Kpotufe, S., editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4068–4125. PMLR.
- Tomczak and Welling, 2018. Tomczak, J. and Welling, M. (2018). Vae with a vampprior. In *International conference on artificial intelligence and statistics*, pages 1214–1223. PMLR.
- Vincent, 2011. Vincent, P. (2011). A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674.
- Williams, 1992. Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256.

Index

Acceptance probability, 37
approximation lemma, 42

bayesian inference, 6
Birkhoff's theorem, 41

canonical space, 32
central limit theorem, 55
coupling, 47
 inequality, 48
 maximal, 49

data augmentation, 64
drift condition, 50
DUGS, 67
dynamical system, 41

ebm, 7
ELBO, 9
ergodic dynamical system, 41
exploration, 11

filtration, 30
 natural, 30
flip operator, 62

generalized Metropolis Hastings, 57
geometric ergodicity, 50
Gibbs sampler, 67

Hamiltonian, 59
 Monte Carlo, 59
 dynamics, 62
hidden Markov models, 8

independence sampler, 38
invariant
 probability measure, 35
 set, 41

involution, 61
irreducibility, 39

kinetic energy, 59

latent variables, 8
leapfrog
 integrator, 63
Leapfrog algorithm, 60
level set, 61

Markov chain, 30
Markov kernel, 29
Markov property, 33
martingales, 54
 central limit theorem, 55
Metropolis Hastings, 36
minorizing condition, 50

Poisson equation, 53
positive part of a measure, 38
potential energy, 59
pseudo marginal, 58

Radon Nikodym, 38
random scan, 68
random walk MH sampler, 38
resampling, 11
reversibility, 35
reweighting, 12
rsgs, 68

sequential Monte Carlo, 11
slice sampler, 66
small set, 50

total variation distance, 48

variational inference, 9
volume-preserving, 61

