Monte Carlo Statistical Methods

Christian P. Robert
CREST, Insee, Paris

George Casella
Cornell University, Ithaca, NY

Draft Version 1.1

February 27, 1998

CHAPTER 1

Introduction

Version 1.1 February 27, 1998

Until the advent of powerful and accessible computing methods, the experimenter was confronted with a difficult choice. Either describe an accurate model of a phenomenon, which would usually prevent the computation of explicit answers, or choose a standard model which would allow this computation, but would often not be a close representation of a realistic model. This dilemma is present in many branches of statistical applications, for example in electrical engineering, aeronautics, biology, networks, and astronomy. To use realistic models, the researchers in these disciplines have often developed original approaches for model fitting that are customized for their own problems. (This is particularly true of physicists, the originators of Markov chain Monte Carlo methods.) Traditional methods of analysis, such as the usual numerical analysis techniques, turn out to be not well adapted for such settings. The first section of this chapter presents a number of examples of statistical models, some of which were instrumental in developing the field of simulation-based inference. The remaining sections describe the difficulties specific to most common statistical methods, while the final section contains a comparison with numerical analysis techniques.

1.1 Statistical Models

In a purely statistical setup, computational difficulties occur at both the level of *probabilistic modeling* of the inferred phenomenon and at the level of *statistical inference* on this model (estimation, prediction, tests, variable selection, etc.). In the first case, a detailed representation of the causes of the phenomenon, such as accounting for potential explanatory variables linked to the phenomenon, can lead to a probabilistic structure which is too complex to allow for a parametric representation of the model. Moreover, there may be no provision for getting closed-form estimates of quantities of interest. A frequent setup with this type of complexity is an *expert systems* (in medicine, physics, finance, etc.) or more generally a *graph structure*. Figure 1.1.1 gives an example of such a structure analyzed in Spiegelhalter et al. (1993). It is related to the detection of a left ventricle hypertrophia (LVH), where the links between causes represent probabilistic dependen-

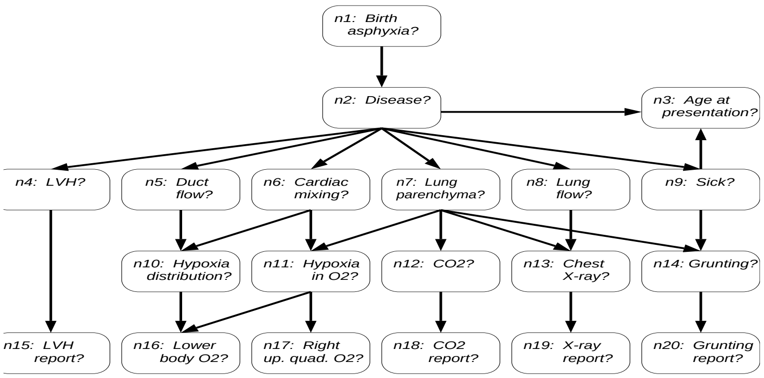


Figure 1.1.1. *Probabilistic representation of links between causes of left ventricle hypertrophia (Source: Spiegelhalter et al., 1993)*

cies. (The motivation behind the analysis is to improve the prediction of this disease.) In this case, the conditional distributions of nodes with respect to their parents lead to the joint distribution. See Robert[1] (1991) or Lauritzen (1993) for other examples of complex expert systems where this reconstitution is impossible.

A second setup where model complexity prohibits an explicit representation appears in econometrics (and in many other areas) for structures of *latent* (or *missing*) variable models. Given a "simple" model, aggregation or removal of some components of this model may sometimes induce such involved structures that simulation is truly the only way to draw an inference. (Chapter 9 provides a series of examples of such models where simulation methods are necessary.)

Example 1.1.1 –**Censored data models**– *Censored data models* can be considered to be missing data models where densities are not sampled directly. To obtain estimates, and make inferences, usually requires programming or computing time and precludes analytical answers.

Barring cases where the censoring phenomenon can be ignored (see Chapter 9), several types of censoring can be categorized by their relation with an underlying (unobserved) model, $Y_i^* \sim f^*(y_i^*|\theta)$:

(i) Given random variables Y_i^*, which may be times of observation or concentrations, the actual observations are $Y_i = \min\{Y_i^*, \overline{u}\}$ where $\overline{u}$ is the maximal observation duration or the smallest measurable concentration

[1] Claudine, not Christian!

rate.

(ii) The original variables Y_i^* are kept in the sample with probability $\rho(y_i^*)$ and the number of discarded variables is either known or unknown.

(iii) The variables Y_i^* are associated with auxiliary variables $X_i \sim g$ such that $y_i = h(y_i^*, x_i)$ is the observation. Typically, $h(y_i^*, x_i) = \min(y_i^*, x_i)$. The fact that truncation occurred, namely the variable $\mathbb{I}_{y_i^* > x_i}$, may be either known or unknown.

As an example, if $X \sim \mathcal{N}(\theta, \sigma^2)$ and $Y \sim \mathcal{N}(\mu, \rho^2)$, the variable $Z = X \wedge Y = \min(X, Y)$ is distributed as

$$\left[1 - \Phi\left(\frac{z-\theta}{\sigma}\right)\right] \quad \rho^{-1}\varphi\left(\tfrac{z-\mu}{\rho}\right) \tag{1.1.1}$$
$$+ \qquad \left[1 - \Phi\left(\frac{z-\mu}{\rho}\right)\right] \; \sigma^{-1}\varphi\left(\frac{z-\theta}{\sigma}\right)$$

where φ is the density of the normal $\mathcal{N}(0,1)$ distribution and Φ is the corresponding cdf, which is not easy to compute. Similarly, if X has a Weibull distribution with two parameters, $\mathcal{W}e(\alpha, \beta)$ and density

$$f(x) = \alpha\beta x^{\alpha-1} e^{-\beta x^\alpha}$$

on $\mathbb{R}^+$, the observation of the censored variable $Z = X \wedge \omega$, where ω is constant, has the density

$$f(z) = \alpha\beta z^\alpha e^{-\beta z^\alpha} \; \mathbb{I}_{z \le \omega} + \left(\int_\omega^\infty \alpha\beta x^\alpha e^{-\beta x^\alpha} dx \right) \delta_\omega(z) \ , \tag{1.1.2}$$

where $\delta_a(\cdot)$ is the Dirac mass at a. In this case, the weight of the Dirac mass, $P(X \ge \omega)$, cannot be explicitly computed.

The distributions (1.1.1) and (1.1.2) appear naturally in quality control applications. There, testing of a product may be of a duration ω, where the quantity of interest is time to failure. If the product is still functioning at the end of the experiment, the observation on failure time is censored. Similarly, in a longitudinal study of a disease, some patients may leave the study either due to other death causes or by simply dropping out. ||

In some cases, the additive form of a density, while formally explicit, prohibits the computation of the density of a sample $(X_1, \cdots, X_n)$ for n large. (Here, "explicit" has the restrictive meaning that "it can be computed in a reasonable time".)

Example 1.1.2 –**Mixture models**– Models of *mixtures of distributions* are based on the assumption that the observations are generated from one of k elementary distributions f_i with probability p_i, the overall density being

$$p_1 f_1(x) + \cdots + p_k f_k(x) \ .$$

An expansion of the distribution of $(X_1, \cdots, X_n)$,

$$\prod_{i=1}^{n} \{p_1 f_1(x_i) + \cdots + p_k f_k(x_i)\} \ ,$$

involves k^n elementary terms, which is prohibitive for large samples. While the computation of standard moments like the mean or the variance of these distributions is feasible in most setups (and thus the derivation of moment estimators, see Problem 1.3), the representation of the likelihood (and therefore the analytical computation of maximum likelihood or Bayes estimates) is generally impossible for mixtures. ||

Lastly, we look at a particularly important example in the processing of temporal (or time series) data where the likelihood cannot be written explicitly.

Example 1.1.3 –**Moving average model**– An MA(q) model describes variables (X_t) that can be modeled as $(t = 0, \ldots, n)$

$$X_t = \varepsilon_t + \sum_{j=1}^{q} \beta_j \varepsilon_{t-j} \ , \tag{1.1.3}$$

where for $i = -q, -(q-1), \cdots$, the ε_i's are i.i.d. random variables $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and for $j = 1, \cdots, q$, the β_js are unknown parameters. If the sample consists of the observation $(X_0, \cdots, X_n)$, where $n > q$, the sample density is (Problem 1.13)

$$\begin{aligned} \int_{\mathbb{R}^q} \sigma^{-(n+q)} \quad & \prod_{i=1}^{q} \varphi\left(\frac{\varepsilon_{-i}}{\sigma}\right) \varphi\left(\frac{x_0 - \sum_{i=1}^{q} \beta_i \varepsilon_{-i}}{\sigma}\right) \\ \times \quad & \varphi\left(\frac{x_1 - \beta_1 \hat{\varepsilon}_o - \sum_{i=2}^{q} \beta_i \varepsilon_{1-i}}{\sigma}\right) \cdots \\ \times \quad & \varphi\left(\frac{x_n - \sum_{i=1}^{q} \beta_i \hat{\varepsilon}_{n-i}}{\sigma}\right) \ d\varepsilon_{-1} \cdots d\varepsilon_{-q} \ , \end{aligned} \tag{1.1.4}$$

with

$$\begin{aligned} \hat{\varepsilon}_0 &= x_0 - \sum_{i=1}^{q} \beta_i \varepsilon_{-i} \ , \\ \hat{\varepsilon}_1 &= x_1 - \sum_{i=2}^{q} \beta_i \varepsilon_{1-i} - \beta_1 \hat{\varepsilon}_0 \ , \\ & \quad \cdots \\ \hat{\varepsilon}_n &= x_n - \sum_{i=1}^{q} \beta_i \hat{\varepsilon}_{n-i} \ . \end{aligned}$$

The iterative definition of the $\hat{\varepsilon}_i$'s is a real obstacle to an explicit integration in (1.1.4) which hinders statistical inference in these models. Note that for $i = -q, -(q-1), \cdots, -1$ the perturbations ε_{-i} can be interpreted as missing data (see Chapter 9). ||

Before the introduction of simulation-based inference, computational difficulties encountered in the modeling of a problem often forced the use of "standard models" and "standard" distributions. One course would be to use models based on *exponential families* (1.3.2) (see Lehmann, 1983, Brown, 1986 or Robert, 1994), which enjoy numerous regularity properties (see Note 1.8.1). Another course was to abandon parametric representations for non-parametric approaches which are by definition robust against modeling errors. In econometrics, the computing bottleneck created by the need for explicit solutions has led to the use of *linear* structures of dependence (see Gouriéroux and Monfort, 1989, 1995).

1.2 Statistical Inference

The statistical techniques that we will be most concerned with are *maximum likelihood* and *Bayesian* methods, and the inferences that can be drawn from their use. In their implementation, these approaches are customarily associated with specific mathematical computations, the former with maximization problems—and thus to an *implicit* definition of estimators as solutions of maximization problems—, the later with integration problems—and thus to a (formally) *explicit* representation of estimators as an integral. (See Lehmann 1983, Berger 1985, Casella and Berger 1990 or Robert 1994 for an introduction to these techniques.) As previously mentioned, reduction to simple, perhaps non-realistic, distributions was often necessitated by computational limitations, but it is also the case that the reduction to simple distributions does not necessarily eliminate the issue of non-explicit expressions, whatever the statistical technique. Our major focus is the application of simulation-based techniques to provide solutions and inference for a more realistic set of models, and hence circumvent the problems associated with the need for explicit or computationally simple answers.

Alternative approaches (see, for instance, Gouriéroux and Monfort 1996) involve solving implicit equations for *methods of moments* or minimization of generalized distances (for *M-estimators*). Approaches by minimal distance can in general be reformulated as maximizations of formal likelihoods as illustrated in Example 1.2.1 below, while the method of moments can sometimes be expressed as a derivation of a maximization problem, that is as a difference equation. Note however that such an interpretation is rare and also that the method of moments is generally sub-optimal when compared with Bayesian or maximum likelihood approaches, these latter two methods using more efficiently the information contained in the distribution of the observations, according to the *Likelihood Principle* (see Robert 1994). But the moment estimators are still of interest as starting values for iterative methods aiming at maximizing the likelihood, since they are convergent in most setups. For instance, in the case of normal mixtures, while the likelihood is not bounded (see Example 1.3.4 below) and therefore there is no maximum likelihood estimator, it can be shown that the solution of the likelihood equations which is closer to the moment estimator is a convergent estimator (see Lehmann 1983).

Example 1.2.1 –**Least Squares Estimators**– Estimation by *least squares* can be traced back to Gauss (1810) and Legendre (1805) (see Stigler 1985). In the particular case of linear regression we observe (x_i, Y_i), $i = 1, \cdots, n$, where

$$Y_i = a + bx_i + \varepsilon_i, \quad i = 1, \cdots, n, \tag{1.2.1}$$

and the variables ε_i represent errors. The parameter (a, b) is estimated by minimizing the distance

$$\sum_{i=1}^{n} (y_i - ax_i - b)^2 \tag{1.2.2}$$

in (a, b), yielding the least squares estimates. If we add more structure to the error term, in particular that $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, independent (equivalently, $Y_i|x_i \sim \mathcal{N}(ax_i + b, \sigma^2)$), the log-likelihood function for (a, b) is proportional to

$$\log(\sigma^{-n}) \; - \sum_{i=1}^{n} (y_i - ax_i - b)^2/2\sigma^2,$$

and it follows that the maximum likelihood estimates of a and b are identical to the least squares estimates. However, the likelihood structure also provides an estimator of σ^2.

Therefore, if, in (1.2.2) we assume $\mathbb{E}(\varepsilon_i) = 0$, or equivalently that the linear relationship $\mathbb{E}[Y|x] = ax+b$ holds, minimization of (1.2.2) is equivalent, from a computational point of view, to imposing a normality assumption on Y conditionally on x and applying maximum likelihood. In this latter case the additional estimator of σ^2 is consistent if the normal approximation is asymptotically valid. ||

Although somewhat obvious, this formal equivalence between the optimization of a function depending on the observations and the maximization of a likelihood associated with the observations has a nontrivial outcome, and applies in many other cases. For example, in the case where the parameters are constrained Robertson et al. (1988) consider a $p \times q$ table of random variables Y_{ij} with means θ_{ij}, where the means are increasing in i and j. Estimation of the θ_{ij}'s by minimizing the sum of the $(y_{ij} - \theta_{ij})^2$'s is possible through the (numerical) algorithm called *"pool-adjacent-violators"* and developed by Robertson *et al.* (1988) to solve this specific problem. (See Problems 1.17 and 1.18.) An alternative is to use an algorithm based on simulation and a representation by a normal likelihood of the problem (see §5.2.4).

1.3 Likelihood Methods

The method of maximum likelihood estimation is quite a popular technique for deriving estimators. Starting from an iid sample $X_1, \ldots, X_n$ from a population with density $f(x|\theta_1,\ldots,\theta_k)$, the *likelihood function* is

$$\begin{aligned} L(\theta_1,\ldots,\theta_k|x) &= L(\theta_1,\ldots,\theta_k|x_1,\ldots,x_n) \\ &= \prod_{i=1}^{n} f(x_i|\theta_1,\ldots,\theta_k). \end{aligned} \tag{1.3.1}$$

More generaly, when the x_i's are not iid, the likelihood is defined as the joint density $f(x_1,\ldots,x_n|\theta)$ taken as a function of θ. The value of θ, say $\hat{\theta}$, which is the parameter value at which $L(\theta|x)$ attains its maximum as a function of θ, with x held fixed, is known as a *maximum likelihood estimator (MLE)*. Notice that, by its construction, the range of the MLE coincides with the range of the parameter. The justification of the maximum likelihood method are primarily asymptotic, in the sense that the MLE is converging almost surely to the true value of the parameter, under fairly general conditions (see Lehmann and Casella 1997 and Problem 1.12), although it can also be interpreting as being at the fringe of the Bayesian paradigm (see, e.g., Berger and Wolpert 1989).

In the context of exponential families, that is, distributions with density

$$f(x) = h(x)\, e^{\theta \cdot x - \psi(\theta)}, \qquad \theta, x \in \mathbb{R}^k, \tag{1.3.2}$$

the approach by maximum likelihood is straightforward. The maximum likelihood estimator of θ is the solution of

$$x = \nabla\psi\{\hat{\theta}(x)\}\ , \tag{1.3.3}$$

which also is the equation yielding a method of moments estimator, since $\mathbb{E}_\theta[X] = \nabla\psi(\theta)$. In practice, there are settings where ψ, the log-Laplace transform, or *cumulant generating function* of h, cannot be computed explicitly. Even if that can be done, it may still be the case that the solution of (1.3.3) is not explicit, or there are constraints on θ such that the maximum of (1.3.2) is not a solution of (1.3.3). This last situation occurs in the estimation of the table of θ_{ij}'s in the discussion following Example 1.2.1.

Example 1.3.1 –**Beta MLE**– The beta $\mathcal{B}e(\alpha,\beta)$ distribution is a particular case of exponential family since its density,

$$f(y|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\, y^{\alpha-1}(1-y)^{\beta-1}\ , \qquad 0 \le y \le 1,$$

can be written as (1.3.2), with $\theta = (\alpha,\beta)$, $x = (\log y, \log(1-y))$. Equation (1.3.3) is then

$$\begin{aligned} \log y &= \Psi(\alpha) - \Psi(\alpha+\beta), \\ \log(1-y) &= \Psi(\beta) - \Psi(\alpha+\beta)\ , \end{aligned} \tag{1.3.4}$$

where $\Psi(z) = d\log\Gamma(z)/dz$ denotes the *digamma function* (see Abramowitz and Stegun 1964). There is no explicit solution to (1.3.4). While it may seem absurd to estimate both parameters of the $\mathcal{B}e(\alpha,\beta)$ distribution from a single observation, Y, the formal computing problem at the core of this example remains valid for a sample $Y_1,\ldots,Y_n$ since (1.3.4) is then replaced with

$$\begin{aligned}\frac{1}{n}\sum_i \log y_i &= \Psi(\alpha) - \Psi(\alpha+\beta), \\ \frac{1}{n}\sum_i \log(1-y_i) &= \Psi(\beta) - \Psi(\alpha+\beta)\ .\end{aligned}$$

||

When the parameter of interest λ is not a one-to-one function of θ, that is when there exists *nuisance* parameters, the maximum likelihood estimator of λ, is still well defined. If the parameter vector is of the form $\theta = (\lambda, \psi)$, where ψ is a nuisance parameter, a typical approach is to calculate the full MLE $\hat\theta = (\hat\lambda, \hat\psi)$, and use the resulting $\hat\lambda$ to estimate λ. In principle, this does not require more complex calculations although the distribution of the maximum likelihood estimator of λ, $\hat\lambda$, may be quite involved. Many other options exist, such as *conditional, marginal, or profile* likelihood. (See Barndorff-Nielsen and Cox 1994.)

Example 1.3.2 –Noncentrality Parameters– If $X \sim \mathcal{N}_p(\theta, I_p)$ and if $\lambda = ||\theta||^2$ is the parameter of interest, the nuisance parameters are the angles Ψ in the polar representation of θ (see Problem 1.2) and the maximum likelihood estimator of λ is $\hat\lambda(x) = ||x||^2$, which has a constant bias equal to p. Surprisingly, an observation $Y = ||X||^2$ which has a non-central chi-squared distribution, $\chi^2_p(\lambda)$ (see Appendix 1), leads to a maximum likelihood estimator of λ which differs[2] from Y, since it is the solution of the implicit equation

$$(1.3.5) \qquad \sqrt{\lambda}\, I_{(p-1)/2}\left(\sqrt{\lambda y}\right) = \sqrt{y}\, I_{p/2}\left(\sqrt{\lambda y}\right) \ , \qquad y > p\ ,$$

where I_ν is the *modified Bessel function* (see Problem 1.14)

$$\begin{aligned} I_\nu(t) &= \frac{(z/2)^\nu}{\sqrt{\pi}\Gamma(\nu+\frac{1}{2})}\int_0^\pi e^{t\cos(\theta)}\sin^{2\nu}(\theta)d\theta \\ &= \left(\frac{t}{2}\right)^\nu \sum_{k=0}^\infty \frac{(z/2)^{2k}}{k!\Gamma(\nu+k+1)} \end{aligned}$$

(see also Abramowitz and Stegun 1964). So even in the favorable context of exponential families we are not necessarily free from computational problems, since the resolution of (1.3.5) requires us first to evaluate the special functions $I_{p/2}$ and $I_{(p-1)/2}$ (see Saxena and Alam 1982). Note also that the maximum likelihood estimator is not a solution of (1.3.5) when $y < p$ (see Problem 1.24). ||

When we leave the exponential family setup, we face increasingly challenging difficulties in using maximum likelihood techniques. One reason for this is the lack of a *sufficient statistic* of fixed dimension outside exponential

[2] This phenomenon is not paradoxical as $Y = ||X||^2$ is not a sufficient statistic in the original problem.

1.48 Show that a Student's t-distribution $\mathcal{T}_p(\nu, \theta, \tau^2)$ does not allow for a conjugate family, apart from the trivial family $\mathcal{F}_0$.

1.49 Show that, for exponential families, a multiplication of the number of hierarchical levels does not modify the conjugate nature of the resulting prior if conjugate distributions with fixed scale parameters are used at every level of the hierarchy. (*Hint:* Consider, for instance, the normal case.)

1.50 Show that the Bayes estimator of $\eta = ||\theta||^2$ under quadratic loss for $\pi(\eta) = 1/\sqrt{\eta}$ and $x \sim \mathcal{N}(\theta, I_p)$ can be written as

$$\delta^\pi(x) = \frac{{}_1F_1(3/2; p/2; ||x||^2/2)}{{}_1F_1(1/2; p/2; ||x||^2/2)},$$

where ${}_1F_1$ is the confluent hypergeometric function. Deduce from the series development of ${}_1F_1$ the asymptotic development of δ^π (for $||x||^2 \to +\infty$) and compare with $\delta_0(x) = ||x||^2 - p$. Study the behavior of these estimators under the weighted quadratic loss

$$L(\delta, \theta) = \frac{(||\theta||^2 - \delta)^2}{2||\theta||^2 + p}$$

and conclude.

1.51 Assuming that $\pi(\theta) = 1$ is an acceptable prior for real parameters, show that this generalized prior leads to $\pi(\sigma) = 1/\sigma$ if $\sigma \in \mathbb{R}^+$ and to $\pi(\varrho) = 1/\varrho(1-\varrho)$ if $\varrho \in [0, 1]$ by considering the natural transformations $\theta = \log(\sigma)$ and $\theta = \log(\varrho/(1-\varrho))$.

1.52 (Dawid et al. 1973) Consider n random variables $x_1, \ldots, x_n$, such that the first ξ of these variables has an $\mathcal{E}xp(\eta)$ distribution and the $n-\xi$ other have a $\mathcal{E}xp(c\eta)$ distribution, where c is known and ξ takes its values in $\{1, 2, \ldots, n-1\}$.

(a) Give the shape of the posterior distribution of ξ when $\pi(\xi, \eta) = \pi(\xi)$ and show that it only depends on $z = (z_2, \ldots, z_n)$, with $z_i = x_i/x_1$.

(b) Show that the distribution of z, $f(z|\xi)$, only depends on ξ.

(c) Show that the posterior distribution $\pi(\xi|x)$ cannot be written as a posterior distribution for $z \sim f(z|\xi)$, whatever $\pi(\xi)$, although it only depends on z. How do you explain this phenomenon?

(d) Show that the paradox does not occur when $\pi(\xi, \eta) = \pi(\xi)\eta^{-1}$.

1.53 (Dawid et al. 1973) Consider u_1, u_2, s^2 such that

$$u_1 \sim \mathcal{N}(\mu_1, \sigma^2), \qquad u_2 \sim \mathcal{N}(\mu_2, \sigma^2), \qquad s^2 \sim \sigma^2\chi^2_\nu/\nu,$$

and $\zeta = (\mu_1 - \mu_2)/(\sigma\sqrt{2})$ is the parameter of interest. The prior distribution is

$$\pi(\mu_1, \mu_2, \sigma) = \frac{1}{\sigma}.$$

(a) Show that the posterior distribution $\pi(\zeta|x)$ only depends on

$$z = \frac{u_1 - u_2}{s\sqrt{2}}.$$

(b) Show that the distribution of z only depends on ζ, but that a paradox occurs; it is still impossible to derive $\pi(\zeta|x)$ from $f(z|\zeta)$, even though $\pi(\zeta|x)$ only depends on z.

(c) Show that the paradox does not occur when

$$\pi(\mu_1,\mu_2,\sigma)=\frac{1}{\sigma^2}.$$

1.54 (Dawid et al. 1973) Consider $2n$ independent random variables,

$$\begin{array}{lcl} x_{11},\ldots,x_{1n} & \sim & \mathcal{N}(\mu_1,\sigma^2), \\ x_{21},\ldots,x_{2n} & \sim & \mathcal{N}(\mu_2,\sigma^2). \end{array}$$

(a) The parameter of interest is $\xi=(\xi_1,\xi_2)=(\mu_1/\sigma,\mu_2/\sigma)$ and the prior distribution is

$$\pi(\mu_1,\mu_2,\sigma)=\sigma^{-p}.$$

Show that $\pi(\xi|x)$ only depends on $z=(z_1,z_2)=(\bar{x}_1/s,\bar{x}_2/s)$ and that the distribution of z only depends on ξ. Derive the value of p which avoids the paradox.

(b) The parameter of interest is now $\zeta=\xi_1$. Show that $\pi(\zeta|x)$ only depends on z_1 and that $f(z_1|\xi)$ only depends on ζ. Give the value of p which avoids the paradox.

(c) Consider the previous questions when $\sigma\sim\mathcal{P}a(\alpha,\sigma_0)$.

1.55 (Dawid et al. 1973) Consider (x_1,x_2) with the following distribution:

$$f(x_1,x_2|\theta)\propto\int_0^{+\infty}t^{2n-1}\exp\left[-\frac{1}{2}\left\{t^2+n(x_1t-\zeta)^2+n(x_2t-\xi)^2\right\}\right]dt,$$

with $\theta=(\zeta,\xi)$. Justify this distribution by considering the setting of Exercise 1.54. The prior distribution on θ is $\pi(\theta)=1$.

(a) Show that $\pi(\zeta|x)$ only depends on x_1 and that $f(x_1|\theta)$ only depends on ζ, but that $\pi(\zeta|x)$ cannot be obtained from $x_1\sim f(x_1|\zeta)$.

(b) Show that, for any distribution $\pi(\theta)$ such that $\pi(\zeta|x)$ only depends on x_1, $\pi(\zeta|x)$ cannot be proportional to $\pi(\zeta)f(x_1|\zeta)$.

1.56 (Jaynes 1980) Consider

$$f(y,z|\eta,\zeta)\propto\frac{\zeta^z\eta^y(1-\eta)^{z-y}}{y!(z-y)!}\ ,\qquad 0\le y\le z,$$

with $0<\eta<1$.

(a) Show that $f(z|\eta,\zeta)$ only depends on ζ and derive the distribution $f(y,z|\eta,\zeta)$ from $f(y|z,\eta,\zeta)$.

(b) Show that, for every $\pi(\eta)$, the paradox does not occur.

1.57 (Dawid et al. 1973) Consider $x=(y,z)$ with distribution $f(x|\theta)$ and $\theta=(\eta,\xi)$. Assume that $\pi(\xi|x)$ only depends on z and that $f(z|\theta)$ only depends on ξ.

(a) Show that the paradox does not occur if $\pi(\theta)$ is proper.

(b) Generalize to the case where $\int\pi(\eta,\xi)\,d\eta=\pi(\xi)$ and examine whether the paradox is evacuated.

1.58 Consider $x_1,\ldots,x_n\sim\mathcal{N}(\mu+\nu,\sigma^2)$, with $\pi(\mu,\nu)\propto 1/\sigma$.

(a) Show that the posterior distribution is not defined for every n.

(b) Extend this result to overparametrized models with improper priors.

1.8 Notes

1.8.1 Conjugate priors

When prior information about the model is quite limited, the prior distribution is often chosen in a parametered family so as to keep the subjective input as limited as possible. Families $\mathcal{F}$ that are *closed under sampling*, that is such that, for every $\pi \in \mathcal{F}$, the posterior distribution $\pi(\theta|x)$ also belongs to $\mathcal{F}$, are of particular interest, for both parsimony and invariance motivations. These families are also called *conjugate* and another justification found in Diaconis and Ylvisaker (1979) is that some Bayes estimators are then linear. But the main motivation for using conjugate priors is their tractability, while they can only be found in exponential families, for reasons related to the *Pitman–Koopman lemma* (see Robert, 1994).

In fact, if the sampling density is of the form

$$f(x|\theta) = C(\theta)h(x)\exp\{R(\theta)\cdot T(x)\}, \tag{1.8.1}$$

which include many common continuous and discrete distributions (see Brown, 1986), a conjugate family for $f(x|\theta)$ is given by

$$\pi(\theta|\mu,\lambda) = K(\mu,\lambda)\, e^{\theta.\mu-\lambda\psi(\theta)},$$

since the posterior distribution is $\pi(\theta|\mu+x,\lambda+1)$. In particular, normal $\mathcal{N}(\theta,\sigma^2)$ distributions are associated with normal $\mathcal{N}(\mu,\tau^2)$ conjugate priors, Poisson $\mathcal{P}(\theta)$ with Gamma $\mathcal{G}(\alpha,\beta)$, Gamma $\mathcal{G}(\nu,\theta)$ with Gamma $\mathcal{G}(\alpha,\beta)$, Binomial $\mathcal{B}(n,\theta)$ with Beta $\mathcal{B}e(\alpha,\beta)$, Multinomial $\mathcal{M}_k(\theta_1,\ldots,\theta_k)$ with Dirichlet $\mathcal{D}(\alpha_1,\ldots,\alpha_k)$, and $\mathcal{N}(\mu,1/\theta)$ with Gamma $\mathcal{G}(\alpha,\beta)$. An extension of (1.8.1) which also allows for conjugate priors contains exponential type densities with parameter dependent support, like the uniform or the Pareto distribution.

As mentioned above, conjugate priors provide linear estimators for a particular parameterization. If $\xi(\theta) = \mathbb{E}_\theta[x]$, which is equal to $\nabla\psi(\theta)$, the prior mean of $\xi(\theta)$ for the prior $\pi(\theta|\mu,\lambda)$ is $\frac{x_0}{\lambda}$ and, if $x_1,\ldots,x_n$ are i.i.d. $f(x|\theta)$,

$$\mathbb{E}^\pi[\xi(\theta)|x_1,\ldots,x_n] = \frac{x_0+n\bar{x}}{\lambda+n}.$$

1.8.2 Gray codes

CHAPTER 2

Random Variable Generation and Computational Methods

Version 1.1 February 27, 1998

The methods developed in this book mostly rely on the possibility of producing (with a computer) a supposedly endless flow of iid random variables for well-known distributions. This generation is, in turn, based on the production of uniform random variables. We thus provide in this chapter a particular uniform generator, along with standard generation methods, to produce random variables from both standard and nonstandard distributions. We also give an introduction to approximation methods for densities and their connections with simulation.

2.1 Simulating Uniform Random Variables

2.1.1 Introduction

Methods of simulation are based on the production of random variables, often independent random variables, that are distributed according to a distribution f that is not necessarily explicitly known (see, for example, Examples 1.1.1, 1.1.2 and 1.1.3). The type of random variable production is formalized below in the definition of a *pseudo-random number generator*. In this chapter, we concentrate on the generation of random variables that are uniform on the interval $[0,1]$, because the uniform distribution $\mathcal{U}_{[0,1]}$ provide the basic probabilistic representation of randomness. In fact, in describing the structure of a space of random variables, it is always possible to represent the generic probability triple $(\Omega, \mathcal{F}, P)$ (where Ω represents the whole space, $\mathcal{F}$ represents a σ-algebra on Ω, and P is a probability measure) as $([0,1], \mathcal{B}([0,1]), \mathcal{U}_{[0,1]})$ (where $\mathcal{B}$ are the Borel sets on $[0,1]$) and therefore equate the variability of $\omega \in \Omega$ with that of a uniform variable in $[0,1]$ (see for instance Billingsley, 1995). The random variables X are then functions from $[0,1]$ to $\mathcal{X}$, transformed by the *generalized inverse.*

Definition 2.1.1 For a function F on $\mathrm{I\!R}$, the *generalized inverse* of F, F^-, is the function defined by

$$F^-(u) = \inf\{x;\ F(x) \geq u\}\ . \tag{2.1.1}$$

We then have the following lemma, sometimes known as the *Probability Integral Transform*, which gives us a representation of any random variable as a transform of a uniform random variable.

Lemma 2.1.2 *If $U \sim \mathcal{U}_{[0,1]}$, then the random variable $F^-(U)$ has the distribution F.*

Proof. For all $u \in [0,1]$, the generalized inverse satisfies

$$F(F^-(u)) \geq u \quad and \quad F^-(F(x)) \leq x \ , \qquad \text{for all } x \text{ in } F^-([0,1]) \ .$$

Therefore,

$$\{(u,x);\ F^-(u) \leq x\} = \{(u,x);\ F(x) \geq u\}$$

and

$$P(F^-(U) \leq x) = P(U \leq F(x)) = F(x) \ .$$

□

Thus, formally, in order to generate a random variable $X \sim F$, it suffices to generate U according to $\mathcal{U}_{[0,1]}$ and then take the transform $x = F^-(u)$. The generation of uniform random variables is therefore a key determinant in the behavior of simulation methods for other probability distributions, since those distributions can be represented as a deterministic transformation of uniform random variables. (Although, in practice, we often use methods other than that of Lemma 2.1.2, this basic representation is usually a good way to think about things, while it somehow clarifies the usual introduction of random variables as *measurable functions*.) More importantly, Lemma 2.1.2 shows that a bad choice of a uniform random number generator can invalidate the resulting simulation procedure.

Before presenting a reasonable uniform random number generator, we first digress a bit to discuss what we mean by a "bad" random number generator. The logical paradox[1] associated with the generation of "random numbers" is the problem of producing a *deterministic* sequence of values in $[0,1]$ which imitates a sequence of *iid* uniform random variables $\mathcal{U}_{[0,1]}$. (Techniques based on the physical imitation of a "random draw" using the internal clock of the machine have been ruled out. This is because, first, there is no guarantee on the *uniform* nature of numbers thus produced and, second, there is no reproducibility of such samples.) But here we really do not want to enter into the philosophical debate on the notion of "random", and whether it is, indeed, possible to "reproduce randomness" (see, for example, Dellacherie, 1978).

For our purposes, there are methods that use a fully deterministic process to produce a random sequence in the following sense: Having generated

[1] Von Neumann (1951) summarizes this problem very clearly by writing *"Any one who considers arithmetical methods of reproducing random digits is, of course, in a state of sin. As has been pointed out several times, there is no such thing as a random number—there are only methods of producing random numbers, and a strict arithmetic procedure of course is not such a method."*

$(X_1, \cdots, X_n)$, knowledge of X_n [or of $(X_1, \cdots, X_n)$] imparts no discernible knowledge of the value of X_{n+1}. Of course, given the initial value X_0, the sample $(X_1, \cdots, X_n)$ is always the same. Thus, the "pseudo-randomness" produced by these techniques is limited since two samples $(X_1, \cdots, X_n)$ and $(Y_1, \cdots, Y_n)$ produced by the algorithm will not be independent, nor identically distributed, nor comparable in any probabilistic sense. This limitation should not be forgotten: the validity of a random number generator is based on a single sample $X_1, \cdots, X_n$ when n tends to $+\infty$ and not on replications $(X_{11}, \cdots, X_{1n}), (X_{21}, \cdots, X_{2n}), \ldots (X_{k1}, \cdots, X_{kn})$ where n is fixed and k tends to infinity. In fact, the distribution of these n-tuples depends on the manner in which the initial values X_{r1} $(1 \leq r \leq k)$ were generated.

It is also the case that the random number generation methods discussed here are not directly related to the *Markov Chain* methods discussed in Chapters 6 and 7, because the variables (X_n) always form a trivial Markov chain, in the sense that the transition kernel is equal to a Dirac mass. In particular, there is neither ergodicity nor convergence of the distribution of X_n to the uniform distribution, since the distribution of X_n given X_0 does remain a Dirac mass for every n. (Surprisingly, it is still possible to speak of stationary distributions in these deterministic setups, as shown in Example 2.1.4.) With these limitations in mind, we can now introduce the following operational definition, which avoids the difficulties of the philosophical distinction between a deterministic algorithm and the reproduction of a random phenomenon.

Definition 2.1.3 *A uniform pseudo-random number generator* is an algorithm which, starting from an initial value u_0 and a transformation D, produces a sequence $(u_i) = (D^i(u_0))$ of values in $[0, 1]$. For all n, the values $(u_1, \cdots, u_n)$ reproduce the behavior of an *iid* sample $(V_1, \cdots, V_n)$ of uniform random variables when compared through a usual set of tests.

This definition is clearly restricted to *testable* aspects of the random variable generation, which are connected through the deterministic transformation $u_i = D(u_{i-1})$. Thus, the validity of the algorithm consists in the verification that the sequence $U_1, \cdots, U_n$ leads to acceptance of the hypothesis

$$\text{H} : U_1, \cdots, U_n \quad \text{are iid} \quad \mathcal{U}_{[0,1]}.$$

The set of tests used is generally of some consequence. There are classical tests of uniformity, such as the Kolmogorov-Smirnov test. Many generators will be deemed adequate under such examination. In addition, and perhaps more importantly, one can use methods of *time series* to determine the degree of correlation between between U_i and $(U_{i-1}, \cdots, U_{i-k})$, by using an ARMA$(p, q)$ model, for instance. One can also use nonparametric tests, like those of Lecoutre and Tassi (1987), applying them on arbitrary decimals of U_i. Marsaglia has also assembled a set of tests called *Die Hard.* Dellacherie (1978) gives a more mathematical treatment of this subject plus a historical review of successive notions of random sequences and the corresponding formal tests of randomness, such as those of Martin-Löef (1966).

Definition 2.1.3 is therefore *functional*: An algorithm that generates uniform numbers is acceptable if it is not rejected by a set of tests. This methodology is not without problems, however. Consider, for example, particular applications that might demand a large number of iterations, as the theory of large deviations (Bucklew, 1990), or particle physics, where algorithms resistant to standard tests may exhibit fatal faults. In particular, algorithms having hidden periodicities (see below) or which are not uniform for the smaller digits may be difficult to detect. Ferrenberg, Landau and Wang (1992) show, for instance, that an algorithm of Wolff (1989), reputed to be "good", results in systematic biases in the processing of Ising models (see Example 5.2.5), due to long term correlations in the generated sequence.

The notion that a deterministic system can imitate a random phenomenon may also suggest the use of *chaotic* models to create random number generators. These models, which result in complex deterministic structures (see Ruelle, 1976, Bergé, Pommeau and Vidal, 1984, Gleick, 1989, or Guégan, 1994) are based on dynamic systems of the form $X_{n+1} = D(X_n)$ which are very sensitive to the initial condition X_0.

Example 2.1.4 –The Logistic Function– The logistic function $D_\alpha(x) = \alpha x(1-x)$ produces, for some values of $\alpha \in [3.57, 4.00]$, chaotic configurations. In particular, the value $\alpha = 4.00$ yields a sequence (X_n) in $[0,1]$ that, theoretically, has the same behavior as a sequence of random numbers (or random variables) distributed according to the *arcsine distribution* with density $1/\pi\sqrt{x(1-x)}$. In a similar manner, the "tent" function",

$$D(x) = \begin{cases} 2x & \text{if } x \le 1/2, \\ 2(1-x) & \text{if } x > 1/2, \end{cases}$$

produces a sequence (X_n) that tends to $\mathcal{U}_{[0,1]}$ (see Problem 2.1). ||

Although the limit (or stationary) distribution associated with a dynamic system $X_{n+1} = D(X_n)$ is sometimes defined and known, the chaotic features of the system are not guarantees for an acceptable behavior (in the probabilistic sense) of the associated generator. In particular, the second generator of Example 2.1.4 has a disastrous behavior. Given the finite representation of real numbers in the computer, the sequence (X_n) sometimes will converge to a fixed value. (For instance, the tent function progressively eliminates the last decimals of X_n.) Moreover, even when these functions give a good approximation of randomness in the unit square $[0,1] \times [0,1]$ (see Example 2.1.5), the hypothesis of randomness is rejected by many standard tests. Classic examples from the theory of chaotic functions do not lead to acceptable pseudo-random number generators.

Example 2.1.5 (Continuation of Example 2.1.4) Figure 2.1.1 illustrates the properties of the generator based on D_α. The histogram of transforms $Y_n = 0.5 + \arcsin(X_n)/\pi$ of a sample of successive values $X_{n+1} =$

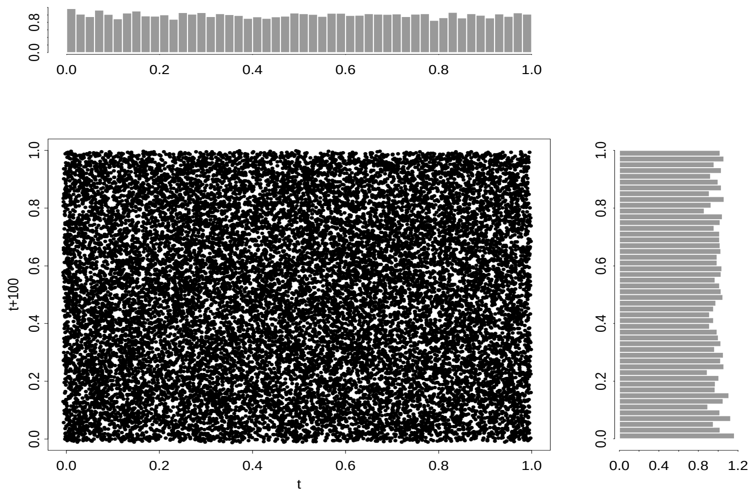


Figure 2.1.1. *Plot of the sample* (Y_t, Y_{t+100}) $(t = 1, \ldots, 9899)$ *for the sequence* $X_{t+1} = 4x_t(1 - X_t)$ *and* $Y_t = F(X_t)$, *along with the (marginal) histograms of* Y_t *and* Y_{t+100}.

$D_\alpha(X_n)$ fits the uniform density extremely well. Moreover, while the plots of (Y_n, Y_{n+1}) and (Y_n, Y_{n+10}) do not display characteristics of uniformity, Figure 2.1.1 shows that the sample of (Y_n, Y_{n+100}) satisfactorily fills the unit square. But the 100 calls to D_α between two generations are prohibitive in terms of computing time. $\|$

We have presented in this introduction some necessary basic notions to now describe a very good pseudo-random number generator, the algorithm *Kiss*[2] of Marsaglia and Zaman (1993). To keep our presentation simple, we only present a single generator, instead of a catalog of usual generators. For those, the books of Knuth (1981), Rubinstein (1981), Ripley (1987) and Fishman (1996) are excellent sources.

2.1.2 The Kiss Generator

As we have remarked above, the finite representation of real numbers in a computer can radically modify the behavior of a dynamic system. Preferred generators are those that take into account the specifics of this representation and provide a uniform sequence. It is important to note that such a sequence does not really take values in the interval $[0, 1]$ but rather on the integers $\{0, 1, \cdots, M\}$, where M is the largest integer accepted by the

[2] The name is an acronym of the saying *Keep it simple, stupid!*, and not reflective of more romantic notions.... After all, this is a statistics text!

(d) Let $X_1, \ldots, X_n$ be iid from the Pareto $\mathcal{P}a(\alpha, \beta)$ distribution with known lower limit α. The corresponding density is

$$f(x|\beta) = \frac{\beta\alpha^\beta}{x^{\beta+1}},$$

$x > \alpha$. Show that

$$S = -\sum \log X_i \ , \qquad K(\beta) = -\log(\beta\alpha^\beta),$$

the saddlepoint is given by

$$\hat{t} = \frac{-n}{s + n\log(\alpha)} \ ,$$

and the saddlepoint approximation is

$$f(\hat{\beta}|\beta) \approx \left[\frac{n}{2\pi}\right]^{1/2} \left(\frac{\beta}{\hat{\beta}}\right)^n e^{(1-\beta/\hat{\beta})} \frac{1}{\hat{\beta}} \ .$$

Show that the renormalized version of this approximation is exact.

2.55 Show that the quasi-Monte Carlo methods introduced in §2.6.1 lead to standard Riemann integration for the equidistributed sequences in dimension 1.

2.56 Establish (2.6.1) and show that the divergence $D(x_1, \ldots, x_n)$ leads to the Kolmogorov-Smirnov test in nonparametric Statistics.

2.6 Notes

2.6.1 Quasi-Monte Carlo methods

Quasi-Monte Carlo methods were proposed in the 1950's to overcome some drawbacks of regular Monte Carlo methods by replacing probabilistic bounds on the errors with deterministic bounds. The idea at the core of quasi-Monte Carlo methods is to substitute the randomly (or pseudo-randomly) generated sequences used in regular Monte Carlo methods with a deterministic sequence (x_n) in order to minimize the so-called *divergence*

$$D(x_1, \ldots, x_n) = \sup_u \left| \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{[0,u]}(x_i) - u \right| .$$

This is also the *Kolmogorov–Smirnov distance* between the empirical cdf and that of the uniform distribution, used in non-parametric tests. For fixed n, the solution is obviously $x_i = \frac{2i-1}{2n}$ in dimension 1 but the goal here is to get a *low-discrepancy* sequence (x_n) which provides small values of $D(x_1, \ldots, x_n)$ for all n's, i.e. such that $x_1, \ldots, x_{n-1}$ do not depend on n, and can thus be updated sequentially.

As shown in Niederreiter (1992), there exist such sequences, which ensure a divergence rate of order $O(n^{-1}\log(n)^{d-1})$, where d is the dimension of the integration space. Since it can be shown that the divergence is related to the overall approximation error by

$$\left| \frac{1}{n} \sum_{i=1}^n h(x_i) - \int f(x)dx \right| \leq V(h)D(x_1, \ldots, x_n) \qquad (2.6.1)$$

(see Niederreiter 1992), where $V(h)$ is the total variation of h,

$$V(f) = \lim_{N\to\infty} \sup_{x_0=1\leq\cdots\leq x_N=1} \sum_{j=1}^{N} |h(x_j) - h(x_{j-1})| \,,$$

the gain over standard Monte Carlo methods can be substantial since standard methods lead to order $O(n^{-1/2})$ errors (see Chapter **??**, §3.4). The advantage over standard integration techniques such as Riemann sums is also important when the dimension d increases since the later are of order n^{d-2} (see Yakowitz et al. 1978).

The true comparison with regular Monte Carlo methods is however more delicate than a simple assessment of the order of convergence. Construction of these sequences, although independent from h, can be quite involved. More importantly, the construction requires that the functions to be integrated have bounded support, which can be a hindrance in practice. See Niederreiter (1992) for extensions in optimization setups.

CHAPTER 3

Monte Carlo Integration

Version 1.2 February 27, 1998

3.1 Introduction

There are two major classes of numerical problems that arise in statistical inference, *optimization* problems and *integration* problems. (An associated problem, that of *implicit equations* can often be reformulated as an optimization problem.) Although optimization is generally associated with the likelihood approach, and integration with the Bayesian approach, these are not strict classifications, as shown by Examples 1.2.1, 1.4.4, and Examples 3.1.1 - 3.1.3.

Examples 1.1.1 – 1.4.4 have also shown that it is not always possible to derive explicit probabilistic models and even less possible to analytically compute the estimators associated with a given paradigm (maximum likelihood, Bayes, method of moments, etc.). Moreover, other statistical methods, such as *bootstrap* methods, although unrelated to the Bayesian approach, may involve the integration of the empirical cdf (see §1.5). Similarly, alternatives to standard likelihood, such as *marginal* likelihood, may require the integration of the nuisance parameters (Barndorff-Nielsen and Cox 1994).

On the other hand, Bayes estimators are not always posterior expectations. In general, the Bayes estimate under the loss function $\mathrm{L}(\theta,\delta)$ and the prior π is the solution of the minimization program

$$\min_{\delta} \int_{\Theta} \mathrm{L}(\theta,\delta)\ \pi(\theta)\ f(x|\theta)\ d\theta\ . \tag{3.1.1}$$

Only when the loss function is the quadratic function $||\theta-\delta||^2$ will the Bayes estimator be a posterior expectation. While some other loss functions lead to general solutions $\delta^\pi(x)$ of (3.1.1) in terms of $\pi(\theta|x)$ (see for instance Robert 1994a or 1996c for the case of *intrinsic losses*), a specific setup where the loss function is constructed by the decision-maker almost always precludes analytical integration of (3.1.1). This necessitates an approximate solution of (3.1.1) either by numerical methods or by simulation.

Thus, whatever the type of statistical inference, we are led to consider numerical solutions. The previous chapter has illustrated a number of methods for the generation of random variables with any given distribution, and hence provides a basis for the construction of solutions to our statistical problems. Thus, just as the search for stationary state in a dynamical system in physics or in economics can require one or several simulations of successive states of the system, statistical inference on complex models will often require the use of simulation techniques. We now look at a number of examples illustrating these situations, before embarking on a description of simulation based integration methods. Chapter 5 deals with the corresponding simulation based optimization methods, which often rely on Markov chain tools, described in Chapter 4.

Example 3.1.1 –L_1 **loss**– For $\theta \in \mathbb{R}$ and $\mathrm{L}(\theta,\delta) = |\theta - \delta|$, the Bayes estimator associated with π is the posterior median of $\pi(\theta|x)$, $\delta^\pi(x)$, which is the solution to the equation

$$\int_{\theta \le \delta^\pi(x)} \pi(\theta)\ f(x|\theta)\ d\theta = \int_{\theta \ge \delta^\pi(x)} \pi(\theta)\ f(x|\theta)\ d\theta\ . \tag{3.1.2}$$

In the setup of Example 1.3.1, that is when $\lambda = ||\theta||^2$ and $X \sim \mathcal{N}_p(\theta, I_p)$, this equation is quite complex, since

$$\pi(\lambda|x) \propto \lambda^{p-1/2} \int e^{-||x-\theta||^2/2} \prod_{i=1}^{p-2} \sin(\varphi_i)^{p-i-1}\ d\varphi_1 \ldots d\varphi_{p-1}\ ,$$

where $\lambda, \varphi_1, \ldots, \varphi_{p-1}$ are the polar coordinates of θ, that is, $\theta_1 = \lambda\cos(\varphi_1)$, $\theta_2 = \lambda\sin(\varphi_1)\cos(\varphi_2)$, ... ||

Example 3.1.2 –**Piecewise linear and quadratic loss functions**– Consider a loss function which is piecewise quadratic,

$$\mathrm{L}(\theta,\delta) = w_i(\theta-\delta)^2 \quad \text{when} \quad \theta - \delta \in [a_i, a_{i+1}), \quad \omega_i > 0. \tag{3.1.3}$$

Differentiating (3.1.3) shows that the associated Bayes estimator satisfies

$$\sum_i w_i \int_{a_i}^{a_{i+1}} (\theta - \delta^\pi(x))\ \pi(\theta|x)\ d\theta = 0\ ,$$

that is

$$\delta^\pi(x) = \frac{\sum_i\ w_i\ \int_{a_i}^{a_{i+1}} \theta\ \pi(\theta)\ f(x|\theta)\ d\theta}{\sum_i\ w_i\ \int_{a_i}^{a_{i+1}} \pi(\theta)\ f(x|\theta)\ d\theta}\ .$$

Although formally explicit, the computation of $\delta^\pi(x)$ requires the computation of the posterior means restricted to the intervals $[a_i, a_{i+1})$, and of the posterior probabilities of these intervals.

Similarly, consider a piecewise linear loss function,

$$\mathrm{L}(\theta,\delta) = w_i|\theta-\delta| \quad \text{if} \quad \theta - \delta \in [a_i, a_{i+1}),$$

or Huber's (1972) loss function,

$$\mathrm{L}(\theta,\delta) = \begin{cases} \rho(\theta-\delta)^2 & \text{if } |\theta-\delta| < c, \\ 2\rho c\{|\theta-\delta| - c/2\} & \text{otherwise,} \end{cases}$$

where ρ and c are specified constants. Although a specific type of prior distribution leads to explicit formulas, most priors result only in integral forms of δ^π. Some of these may be quite complex. ||

Inference based on *classical decision theory* evaluates the performance of estimators (maximum likelihood estimator, best unbiased estimator, moment estimator, etc.) through the loss imposed by the decision-maker or by the setting. Estimators are then compared through their expected losses, also called risks. In most cases it is impossible to obtain an analytical evaluation of the risk of a given estimator, or even to establish that a new estimator (uniformly) dominates a standard estimator.

It may seem that the topic of *James-Stein* estimation is an exception to this observation, given the abundant literature on the topic. In fact, for some families of distributions (such as exponential or spherically symmetric) and some types of loss functions (such as quadratic or concave), it is possible to analytically establish domination results over the maximum likelihood estimator or unbiased estimators (see Robert 1994 Chapter 8, or Lehmann and Casella 1997, Chapter 5). Nonetheless, in these situations, estimators such as *empirical Bayes estimators*, which are quite attractive in practice, will rarely allow for analytic expressions. This makes their evaluation under a given loss problematic.

Given a sampling distribution $f(x|\theta)$ and a conjugate prior distribution $\pi(\theta|\lambda,\mu)$, the empirical Bayes method estimates the *hyperparameters* λ, μ from the *marginal distribution*

$$m(x|\lambda,\mu) = \int f(x|\theta)\ \pi(\theta|\lambda,\mu)\ d\theta$$

by maximum likelihood. The estimated distribution $\pi(\theta|\hat{\lambda},\hat{\mu})$ is then used as in a standard Bayesian approach (that is, without taking into account the effect of the substitution), to derive a point estimator. See Maritz and Lwin (1989) or Searle et al. (1992, Chapter 9) for a more detailed discussion on this approach. The following example illustrates some difficulties encountered in evaluating empirical Bayes estimators (see also Example 3.5.1).

Example 3.1.3 –**Empirical Bayes estimator**– Let X have the distribution $X \sim \mathcal{N}_p(\theta, I_p)$. The corresponding conjugate prior is $\mathcal{N}_p(\mu, \lambda I_p)$ where the hyperparameter μ is generally fixed, for instance $\mu = 0$. In the empirical Bayes approach, the scale hyperparameter λ is replaced by the maximum likelihood estimator, $\hat{\lambda}$, based on the marginal distribution $X \sim \mathcal{N}_p(0, (\lambda+1)I_p)$. This leads to the maximum likelihood estimator $\hat{\lambda} = (||x||^2 - p + 1)^+$. Since the posterior distribution of θ given λ is $\mathcal{N}_p(\lambda x/(\lambda+1), \lambda I_p/(\lambda+1))$, the empirical Bayes inference is based on the pseudo-posterior $\mathcal{N}_p(\hat{\lambda}x/(\hat{\lambda}+1), \hat{\lambda}I_p/(\hat{\lambda}+1))$. If, for instance, $||\theta||^2$ is the quantity of interest, and if it is evaluated under a quadratic loss, the empirical Bayes estimator is

$$\begin{aligned}\delta^{eb}(x) &= \left(\frac{\hat\lambda}{\hat\lambda+1}\right)^2 ||x||^2 + \frac{\hat\lambda p}{\hat\lambda+1} \\ &= \left[\left(1-\frac{p}{||x||^2}\right)^+\right]^2 ||x||^2 + p\left(1-\frac{p}{||x||^2}\right)^+ \\ &= (||x||^2-p)^+ .\end{aligned}$$

This estimator dominates both the best unbiased estimator, $||x||^2-p$, and the maximum likelihood estimator based on $||x||^2 \sim \chi^2_p(||\theta||^2)$ (see Saxena and Alam 1982 and Example 1.3.2). However, since the proof of this second domination result is quite involved, one might first check for domination through a simulation experiment which evaluates the risk function,

$$R(\theta,\delta) = \mathbb{E}_\theta[(||\theta||^2-\delta)^2] ,$$

for the three estimators. This quadratic risk is often normalized by $1/(2||\theta||^2 + p)$ (which does not affect domination results but ensures the existence of a minimax estimator; see Robert 1994), and then the resulting Bayes estimator

$$\delta^\pi(x) = \mathbb{E}^\pi\left[\frac{||\theta||^2}{2||\theta||^2+p}\,\middle|\, x,\lambda\right] \Big/ \mathbb{E}^\pi\left[\frac{1}{2||\theta||^2+p}\,\middle|\, x,\lambda\right]$$

does not have an explicit form. ||

A general solution to the different computational problems contained in the previous examples and in those of §1.1 is to use simulation, of either the true or approximate distributions to calculate the quantities of interest. In the setup of Decision Theory, whether it is classical or Bayesian, this solution is natural, since risks and Bayes estimators involve integrals with respect to probability distributions. We will see in Chapter 5 why this solution also applies in the case of maximum likelihood estimation. Note that the possibility of producing an almost infinite number of random variables distributed according to a given distribution gives us access to the use of *frequentist* and *asymptotic* results much more easily than in usual inferential settings (see Serfling 1987, or Lehmann and Casella 1997, Chap. IT) where the sample size is most often fixed. One can therefore apply probabilistic results such as the Law of Large Numbers or the Central Limit Theorem, since they allow for a control of the convergence of simulation methods (which is equivalent to the deterministic bounds used by numerical approaches.)

3.2 Classical Monte Carlo integration

Before applying our simulation techniques to more practical problems, we first need to develop their properties in some detail. This is more easily accomplished by looking at the generic problem of evaluating the integral

$$\mathbb{E}_f[h(X)] = \int_{\mathcal{X}} h(x)\ f(x)\ dx\ . \tag{3.2.1}$$

Based on previous developments, it is natural to propose using a sample $(X_1, \ldots, X_m)$ generated from the density f to approximate (3.2.1) by the empirical average

$$\overline{h}_m = \frac{1}{m} \sum_{j=1}^{m} h(x_j)\ ,$$

since $\overline{h}_m$ converges almost surely to $\mathbb{E}_f[h(X)]$ by the Strong Law of Large Numbers. Moreover, when h^2 has a finite expectation under f, the speed of convergence of $\overline{h}_m$ can be assessed since the variance

$$\text{var}(\overline{h}_m) = \frac{1}{m} \int_{\mathcal{X}} \left(h(x) - \mathbb{E}_f[h(X)]\right)^2 f(x)dx$$

can also be estimated from the sample $(X_1, \ldots, X_m)$ through

$$v_m = \frac{1}{m^2} \sum_{j=1}^{m} [h(x_j) - \overline{h}_m]^2\ .$$

For m large,

$$\frac{\overline{h}_m - \mathbb{E}_f[h(X)]}{\sqrt{v_m}}$$

is therefore approximately distributed as a $\mathcal{N}(0,1)$ variable, and this leads to the construction of a convergence test and of confidence bounds on the approximation of $\mathbb{E}_f[h(X)]$.

Example 3.2.1 (Continuation of Example 3.1.3) Consider the evaluation of δ^π for $p = 3$ and $x = (0.1, 1.2, -0.7)$. Since

$$\delta^\pi(x) = \frac{1}{2}\ \left\{\mathbb{E}^\pi[(2||\theta||^2 + 3)^{-1}|x]^{-1} - 3\right\}\ ,$$

this requires the computation of

$$\mathbb{E}^\pi[(2||\theta||^2 + 3)^{-1}|x] = \int_{\mathbb{R}^3} (2||\theta||^2 + 3)^{-1}\ \pi(\theta|x)\ d\theta\ .$$

If $\pi(\theta)$ is the non-informative prior distribution proportional to $||\theta||^{-2}$ (see Example 1.3.2), we would need a sample $(\theta_1, \ldots, \theta_m)$ from the posterior distribution

$$\pi(\theta|x) \propto ||\theta||^{-2}\ \exp\left\{-||x - \theta||^2/2\right\}\ . \tag{3.2.2}$$

The simulation of (3.2.2) can be done by representing θ in polar coordinates $(\rho, \varphi_1, \varphi_2)$ $(\rho > 0, \varphi_1 \in [0, 2\pi], \varphi_2 \in [-\pi/2, \pi/2])$, with $\theta = (\rho\cos\varphi_1,$ $\rho\sin\varphi_1\cos\varphi_2,\ \rho\sin\varphi_1\sin\varphi_2)$, which yields

$$\pi(\rho, \varphi_1, \varphi_2|x) \propto \exp\left\{\rho x \cdot (\theta/\rho) - \rho^2/2\right\}\ \sin(\varphi_1)\ .$$

If we denote $\xi = \theta/\rho$, which only depends on (φ_1, φ_2), then $\rho|\varphi_1, \varphi_2 \sim \mathcal{N}(x \cdot \xi, 1)$ truncated to $\mathbb{R}^+$. The integration of ρ then leads to

$$\pi(\varphi_1, \varphi_2|x) \propto \Phi(-x \cdot \xi) \exp\{(x \cdot \xi)^2/2\} \sin(\varphi_1),$$

where $x \cdot \xi = x_1 \cos(\varphi_1) + x_2 \sin(\varphi_1) \cos(\varphi_2) + x_3 \sin(\varphi_1) \sin(\varphi_2)$. Unfortunately, the marginal distribution of (φ_1, φ_2) is not directly available since it involves the cdf of the normal distribution, Φ.

For simulation purposes, we can modify the polar coordinates in order to remove the positivity constraint on ρ. The alternative constraint becomes $\varphi_1 \in [-\pi/2, \pi/2]$, which ensures identifiability for the model. Since ρ now varies in $\mathbb{R}$, the marginal distribution of (φ_1, φ_2) is

$$\pi(\varphi_1, \varphi_2|x) \propto \exp\{(x \cdot \xi)^2/2\} \sin(\varphi_1),$$

which can be simulated by an accept-reject algorithm using the instrumental function $\sin(\varphi_1) \exp\{||x||^2/2\}$. One can therefore simulate (φ_1, φ_2) based on a uniform distribution on the half unit sphere. The algorithm corresponding to this decomposition is the following

Algorithm A.19 ***Polar Simulation***

```
1. Simulate $(\varphi_1, \varphi_2)$ from the uniform distribution
   on the half unit sphere and $U$ from $\mathcal{U}_{[0,1]}$
   until
          $U \leq \exp\{x \cdot \xi - ||x||^2/2\}$          $[A.19]$
2. Generate $\rho$ from the normal distribution
          $\mathcal{N}(x_1 \cos(\varphi_1) + x_2 \sin(\varphi_1) \cos(\varphi_2) + x_3 \sin(\varphi_1) \sin(\varphi_2), 1)$
```

The sample resulting from $[A.19]$ provides a subsample $(\rho_1, \ldots, \rho_m)$ in step 2. and an approximation of $\mathbb{E}^\pi[(2\rho^2 + 3)^{-1}|X]$,

$$T_m = \frac{1}{m} \sum_{j=1}^{m} (2\rho_j^2 + 3)^{-1} .$$

Figure 3.2.1 gives a realization of a sequence of T_m, the envelope being constructed from the normal approximation through the 95% confidence interval $T_m \pm 1.96\sqrt{v_m}$. ||

The approach followed in the above example can be successfully utilized in many cases, even though it is often possible to achieve greater efficiency through numerical methods (Riemann quadrature, Simpson method, etc.) in dimension 1 or 2. The scope for application of this Monte Carlo integration method is obviously not only limited to the Bayesian paradigm, since the performances of complex procedures can be measured as in Example ?? in any setting where the distributions involved in the model can be simulated. We mentioned in §3.1 the potential of this approach in evaluating estimators based on a decision-theoretic derivation. The same applies for testing (which is formally a branch of Decision Theory) where the level of

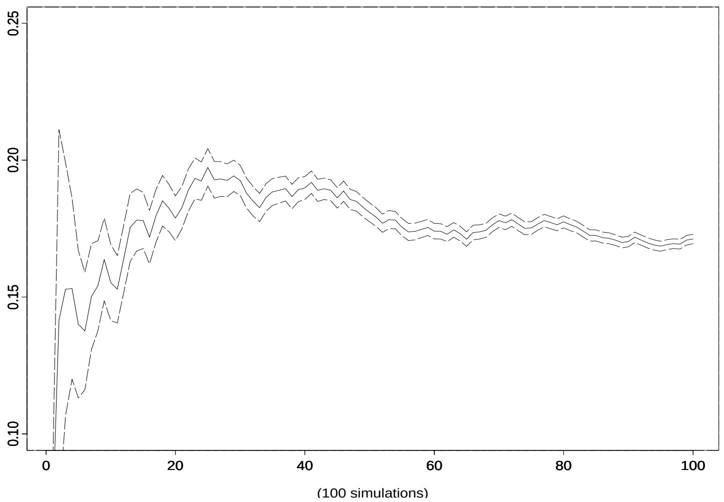


Figure 3.2.1. *Convergence of the Bayes estimator of* $||\theta||^2$ *under normalized quadratic loss for the reference prior* $\pi(\theta) = ||\theta||^{-2}$ *and the observation* $x = (0.1, 1.2, -0.7)$*. The envelope provides a nominal* 95% *confidence interval on* $||\theta||^2$*.*

significance of a test, and its power function, can be easily computed, and simulation thus can provide a useful improvement over asymptotic approximations when explicit computations are impossible.

Example 3.2.2 –**Normal cdf**– Since the normal cdf cannot be written in an explicit form, a possible way to construct normal distribution tables is to use simulation. Consider thus the generation of a sample of size n, $(X_1, \ldots, X_n)$, based on the Box-Muller algorithm $[A_1]$ of Example 2.2.1.

The approximation of

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

by the Monte Carlo method is thus

$$\hat{\Phi}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{x_i \le t},$$

with (exact) variance $\Phi(t)(1-\Phi(t))/n$ (as the variables $\mathbb{I}_{x_i \le t}$ are independent Bernoulli with success probability $\Phi(t)$). For values of t around $t = 0$ the variance is thus approximately $1/4n$ and to achieve a precision of four decimals the approximation requires on average $\sqrt{n} = \sqrt{2}\, 10^4$ simulations, that is, 200 million iterations. Table ?? gives the evolution of this approximation for several values of t and shows an accurate evaluation for 100 million iterations. Note that greater accuracy is achieved in the tails. ||

$$= \frac{1}{n}\left(h(Z_{n+t}) + \sum_{i=1}^{n+t-1} b(Z_i)h(Z_i)\right)$$

with

$$b(Z_i) = \left(1 + \frac{t(g(Z_i) - \rho f(Z_i))}{(n-1)(1-\rho)f(Z_i)}\right)^{-1} .$$

3.26 **(Continuation of Problem 3.25)** If $S_n = \sum_1^{n+t-1} b(z_i)$, show that

$$\delta = \frac{1}{n}\left(h(Z_{n+t}) + \frac{n-1}{S_n}\sum_{i=1}^{n+t-1} b(Z_i)h(Z_i)\right)$$

asymptotically dominates the usual Monte Carlo approximation, conditional on the number of rejected variables t, under quadratic loss. (*Hint:* Show that the sum of the weights S_n can be replaced by $(n-1)$ in δ and assume $\mathbb{E}_f[h(X)] = 0$.)

3.27 (Berger, Philippe and Robert 1997) For Σ a $p{\times}p$ positive definite symmetric matrix, consider the distribution

$$\pi(\theta) \propto \frac{\exp\left(-(\theta-\mu)^t\Sigma^{-1}(\theta-\mu)/2\right)}{||\theta||^{p-1}} .$$

(a) Show that the distribution is well-defined, that is, that

$$\int_{\mathbb{R}^p} \frac{\exp\left(-(\theta-\mu)^t\Sigma^{-1}(\theta-\mu)/2\right)}{||\theta||^{p-1}} d\theta < \infty.$$

(b) Show that an importance sampling implementation based on the normal instrumental distribution $\mathcal{N}_p(\mu, \Sigma)$ is not satisfactory from both theoretical and practical points of view.

(c) Examine the alternative based on a gamma distribution $\mathcal{G}a(\alpha, \beta)$ on $\eta = ||\theta||^2$ and a uniform distribution on the angles.

3.28 Given a binomial experiment $x_n \sim \mathcal{B}(n, p)$ with $p = 10^{-6}$, determine the minimum sample size n for

$$P\left(\left|\frac{x_n}{n-p}\right| \le \epsilon p\right)$$

when $\epsilon = 10^{-1}, 10^{-2}, 10^{-3}$.

3.29 When the Y_i's are generated from (3.7.1), show that $J^{(m)}$ is distributed from $\lambda(\theta_0)^{-n}\exp(-n\theta J)$. Deduce that (3.7.2) is unbiased.

3.30 Show that

$$\int_0^t (\xi(s) - \xi(0))d\xi(s) = \frac{1}{2}(\xi(t)^2 - t)$$

3.31 Show that, in (3.7.4), $b(\cdot)$ is defined by $b(x) = b_0(x) + \frac{1}{2}\sigma(x)\sigma'(x)$ in dimension $\alpha = 1$. Extend to the general case by introducing $\partial\sigma_j$, which is a matrix with (i, k) element $\partial\sigma_{ij}(x)/\partial x_k$.

3.32 Show that the solution to (3.7.3) can also be expressed through a *Stratonovich integral*,

$$X(t) = X(0) + \int_0^t b_0(X(s))ds + \int_0^t \sigma(X(s)) \circ d\xi(s),$$

where the second integral is defined as the limit

$$\lim_{\Delta \to 0} \sum_{i=1}^{n} \frac{\sigma(X(s_i)) + \sigma(X(s_{i-1}))}{2} (\xi(s_i) - \xi(s_{i-1})),$$

where $0 = s_0 < s_1 < \ldots < s_n = t$ and $\max_i(b_i - s_{i-1}) \leq \Delta$. (*Note:* The integral above cannot be defined in the usual sense because the Wiener process has almost surely unbounded variations. See Talay 1996, p.56.)

3.33 Show that the solution to the Ornstein-Uhlenbeck equation

$$dX(t) = -X(t)dt + \sqrt{2}d\xi(t) \tag{3.6.3}$$

is a stationary $\mathcal{N}(0,1)$ process.

3.34 Show that, if $(X(t))$ is solution to (3.6.3), $Y(t) = \operatorname{atan} X(t)$ is solution of a SDE with $b(x) = \frac{1}{4}\sin(4x) - \sin(2x)$, $\sigma(x) = \sqrt{2}\cos^2(x)$.

3.7 Notes

3.7.1 Large deviations techniques

When we introduced importance sampling methods in §3.3, we showed in Example 3.3.1 that alternatives to direct sampling were preferable where sampling from the tails of a distribution f. When the event A is particularly rare, say $p(A) \leq 10^{-6}$, methods such as importance sampling (§3.3) are definitely necessary to get an acceptable approximation. Since the optimal choice given in Theorem 3.3.4 is formal, in the sense that it involves the unknown constant I, more practical choices have been proposed in the literature. In particular, Bucklew (1990) indicates how the *theory of large deviations* may help in devising proposal distributions in this purpose.

Very sketchily, the theory of large deviations is concerned with the approximation of tail probabilities $P(|S_n - \mu| > \varepsilon)$ when S_n is a sum of i.i.d. random variables,

$$S_n = \frac{X_1 + \ldots + X_n}{n},$$

n goes to infinity, and ε is large. (When ε is small, the normal approximation based on the Central Limit Theorem works well enough.) If $M(\theta) = \mathbb{E}[\exp(\theta X_1)]$ is the moment generating function of X_1, and if $I(x) = \sup_\theta\{\theta x - \log M(\theta)\}$, the large deviation approximation is

$$\frac{1}{n}\log P(S_n \in F) \approx -\inf_F I(x).$$

This result is sometimes called *Cramer's theorem.*

The simulation device based on this approximation is called *twisted simulation.* If the problem is to evaluate

$$I = P\left(\frac{1}{n}\sum_{i=1}^{n} h(x_i) \geq 0\right),$$

3.7.2 Simulation of stochastic differential equations

Given a differential equation in $\mathbb{R}^d$,

$$dX(t) = b_0(X(t))dt,$$

where b_0 is a function from $\mathbb{R}^d$ to $\mathbb{R}^d$, it is often of interest to consider the perturbation of this equation by a random noise, $\xi(t)$,

$$\frac{dX(t)}{dt} = b_0(X(t)) + \sigma(X(t))\xi(t), \tag{3.7.3}$$

which is called a *stochastic differential equation* (SDE),[6] σ being the variance factor taking values in the space of $d \times d$ matrices. Applications of SDE's abound in fluid mechanics, random mechanics and particle Physics.

The perturbation ξ in (3.7.3) is often chosen to be a *Wiener process*, that is such that $\xi(t)$ is a Gaussian vector with mean 0, independent components and correlation

$$\mathbb{E}[\xi_i(t)\xi_j(s)] = \delta_{ij}\min(t,s) \ ,$$

where $\delta_{ij} = \mathbb{I}_{i=j}$.

The solution of (3.7.3) can also be represented through an *Itô integral*,

$$X(t) = X(0) + \int_0^t b(X(s))ds + \int_0^t \sigma(X(s))d\xi(s), \tag{3.7.4}$$

where b is derived from b_0 and σ (see Problems 3.31 and 3.32), and the second integral is the limit

$$\lim_{\Delta \to 0} \sum_{i=1}^{n} \sigma(X(s_i))(\xi(s_i) - \xi(s_{i-1})),$$

where $0 = s_0 < s_1 < \ldots < s_n = t$ and $\max_i(b_i - s_{i-1}) \leq \Delta$. This limit exists whenever $\sigma(X)$ is square-integrable, that is when

$$\mathbb{E}\left[\int_0^t |\sigma(X(s))|^2 ds\right] < \infty.$$

See, e.g., Ikeda and Watanabe (1981) for details.

In this setup, simulations are necessary to produce an approximation of the trajectory of $(X(t))$, given the Wiener process $(\xi(t))$, or to evaluate expectations of the form $\mathbb{E}[h(X(t))]$. It may also be of interest to compute the expectation $\mathbb{E}[h(X)]$ under the stationary distribution of $(X(t))$, when this process is ergodic, a setting we will encounter again in the MCMC method (see Chapters 4 and 7).

A first approximation to the solution of (3.7.4) is based on the discretization

$$X(t) \approx X(0) + b(X(0))t + \sigma(X(0))(\xi(t) - \xi(0)) \ ,$$

[6] The material in this section is of a more advanced mathematical level than the remainder of the book and it will not be used in the sequel. This description of simulation methods for SDE's borrows heavily from the expository paper of Talay (1996) which presents in much deeper details the use of simulation techniques in this setup.

CHAPTER 4

Markov Chains

Version 1.1 February 27, 1998

In this chapter we introduce fundamental notions of Markov chains, and state the results that are needed to establish the convergence of various MCMC algorithms and , more generally, to understand the literature on this topic. Thus this chapter, along with basic notions of probability theory, will provide enough foundation for the understanding of the following chapters. It, unfortunately, is necessarily a brief and therefore incomplete introduction to Markov chains, and we refer the reader to Meyn and Tweedie (1993), on which this chapter is based, for a thorough introduction to Markov chains. Other perspectives can be found in Doob (1953), Chung (1967), Feller (1970, 1971), Billingsley (1995) for general treatments, and Nummelin (1984), Revuz (1984) and Resnick (1994) for books entirely dedicated to Markov chains. Given the purely utilitarian goal of this chapter, its style and presentation differ from those of other chapters, especially with regard to the plethora of definitions and theorems and to the rarity of examples and proofs. In order to make the book accessible to those who are more interested in the implementation aspects of MCMC algorithms than in their theoretical foundations, we include a preliminary section on the essential facts about Markov chains that are necessary for the next chapters.

Before formally introducing the notion of a Markov chain, note that we do not deal in this chapter with Markovian models in *continuous time* (also called *Markovian processes*) since the very nature of simulation leads[1] us to consider only discrete time stochastic processes, $(X_n)_{n\in\mathbb{N}}$. Indeed, Hastings (1970) notes that the use of pseudo-random generators and the representation of numbers in a computer imply that the Markov chains related with Markov chain Monte Carlo methods are, in fact, finite state space Markov chains. However, we also consider arbitrary state space Markov chains to allow for continuous support distributions and to avoid addressing the problem of approximation of these distributions with discrete support distributions, since such an approximation depends on both material and algorithmic specifics of a given technique (see Roberts, Rosenthal and Schwartz,

[1] Some Markov chain Monte Carlo algorithms still employ a diffusion representation to speed up convergence to the stationary distribution (see for instance §??, Roberts and Tweedie 1995 or Phillips and Smith 1996).

1995, for a study of the influence of discretization on the convergence of Markov chains associated with Markov chain Monte Carlo algorithms).

4.1 Essentials for MCMC

For those familiar with the properties of Markov chains, this first section provides a brief survey of the properties of Markov chains that are contained in the chapter, and are essential for the study of MCMC methods. Starting with Section 4.2, the theory of Markov chains is developed from first principles.

In the setup of MCMC algorithms, Markov chains are constructed from a *transition kernel* K (Definition 4.2.1), which is a conditional probability density, as $X_{n+1} \sim K(X_n, X_{n+1})$. A typical example is provided by the *random walk* process, defined as $X_{n+1} = X_n + \epsilon_n$, where ϵ_n is generated independently of $X_n, X_{n-1}, \ldots$ (see Example 4.9.4). The chains encountered in MCMC settings enjoy a very strong stability property, namely a *stationary probability distribution* exists by construction (Definition 4.5.1). That is, a distribution π such that, if $X_n \sim \pi$, $X_{n+1} \sim \pi$, if the kernel K allows for free moves all over the state space (this freedom is called *irreducibility* in the theory of Markov chains and is formalized in Definition 4.3.1 as the existence of $n \in \mathbb{N}$ such that $P(X_n \in A|X_0) > 0$ for every A such that $\pi(A) > 0$). This property also ensures that most of the chains involved in MCMC algorithms are *recurrent*, that is that the average number of visits to an arbitrary set A is infinite (Definition 4.4.5), or even Harris recurrent, that is such that the probability of an infinite number of returns to A is 1 (Definition 4.4.8), which ensures that the chain has the same limiting properties for every starting value instead of *almost* every starting value. (Therefore, this appears as an equivalent for Markov chains of the notion of continuity for functions.)

This latter point is quite important in the context of MCMC algorithms. Since most algorithms are started from some arbitrary point x_0, we are in effect starting the algorithm from a set of measure zero (under a continuous dominating measure). Thus, insuring that the chain converges for almost every starting point is not enough, and we need Harris recurrence to guarantee convergence from every starting point.

The *stationary probability* is also a *limiting distribution* in the sense that the limiting distribution of X_{n+1} is π under the total variation norm (see Proposition 4.6.2), notwithstanding the initial value of X_0. Stronger forms of convergence are also encountered in MCMC settings, like *geometric* and *uniform* convergence (see Definitions 4.6.8 and 4.6.11). In a simulation setup, a most interesting consequence of this convergence property is that the average

$$\frac{1}{N} \sum_{n=1}^{N} h(X_n) \tag{4.1.1}$$

converges to the expectation $\mathbb{E}_\pi[h(X)]$ almost surely. When the chain is *reversible*, that is when the transition kernel is symmetric, a Central Limit theorem also holds for this average.

In Chapter 8, diagnoses will be based on a minorization condition. that is the existence of a set C such that there also exists $m \in \mathbb{N}$, $\epsilon_m > 0$, and a probability measure ν_m such that

$$P(X_m \in A|X_0) \geq \epsilon_m \nu_m(A)$$

when $X_0 \in C$. The set C is then called a *small set* (Definition 4.3.7) and visits of the chain to this set can be exploited to create independent batches in the sum (4.1.1), since, with probability ϵ_m the next value of the Markov chain is generated from the minorizing measure ν_m, which is independent of X_m.

4.2 Basic notions

A Markov chain is a sequence of random variables that can be thought of as evolving over time, with probability of a transition depending on the particular set that the chain is in. It therefore seems natural, and in fact is mathematically somewhat cleaner, to define the chain in terms of its *transition kernel*, the function that determines these transitions.

Definition 4.2.1 A *transition kernel* is a function K defined on $\mathcal{X} \times \mathcal{B}(\mathcal{X})$ such that

(i) $\forall x \in \mathcal{X}$, $K(x, \cdot)$ is a probability measure;

(ii) $\forall A \in \mathcal{B}(\mathcal{X})$, $K(\cdot, A)$ is measurable.

When $\mathcal{X}$ is *discrete*, the transition kernel simply is a (transition) matrix K defined by

$$P_{xy} = P(X_n = y|X_{n-1} = x)\ , \qquad x, y \in \mathcal{X}.$$

In the continuous case, the *kernel* also denotes the conditional density $K(x, x')$ of the transition $K(x, \cdot)$. That is, $P(X \in A|x') = \int_A K(x', x)dx$.

Example 4.2.2 – **Bernoulli-Laplace Model**– Consider $\mathcal{X} = \{0, 1, \cdots, M\}$ and (X_n) such that X_n represents the state, at time n, of a tank which contains exactly M particles and is connected to another identical tank. Two types of particles are introduced in the system, and there are M of each type. If X_n denotes the number of particles of the first kind in the first tank at time n, and the moves are restricted to a single exchange of particles between both tanks at each instant, the transition matrix is given by (for $0 < x, y < M$)

$$\begin{aligned} P_{xy} &= 0 \quad \text{if} \quad |x - y| > 1\ , \\ P_{xx} = 2\,\frac{x(M-x)}{M^2}\ ,\ P_{x(x-1)} &= \left(\frac{x}{M}\right)^2\ ,\ P_{x(x+1)} = \left(\frac{M-x}{M}\right)^2 \end{aligned}$$

and $P_{01} = P_{M(M-1)} = 1$. (This model is the *Bernoulli–Laplace* model, see Feller, 1970, Chap. XV) $\|$

The chain (X_n) is usually defined for $n \in \mathbb{N}$ rather than for $n \in \mathbb{Z}$. Therefore, the distribution of X_0, the initial state of the chain, plays an important role. In the discrete case, given an initial distribution $\mu = (\omega_1, \omega_2, \ldots)$, the marginal probability distribution of X_1 is then

$$\mu_1 = \mu K \tag{4.2.1}$$

and, by repeated multiplication, $X_n \sim \mu_n = \mu K^n$. Similarly, in the continuous case, if μ denotes the initial distribution of the chain, namely if

$$X_0 \sim \mu \, , \tag{4.2.2}$$

then we let P_μ denote the probability distribution of (X_n) under condition (4.2.2). When X_0 is fixed, in particular for μ equal to the Dirac mass δ_{x_0}, we use the alternative notation P_{x_0}.

Definition 4.2.3 Given a transition kernel K, a sequence $X_0, X_1, \cdots, X_n$, $\cdots$ of random variables is a *Markov chain*, denoted by (X_n), if, for any t, the conditional distribution of X_t given $x_{t-1}, x_{t-2}, \ldots, x_0$ is the same as the distribution of X_t given x_{t-1}. That is,

$$\begin{aligned} P(X_{k+1} \in A | x_0, x_1, x_2, \cdots, x_k) &= P(X_{k+1} \in A | x_k) \\ &= \int_A K(x_k, dx) \end{aligned} \tag{4.2.3}$$

The chain is *time-homogeneous* if the distribution of $(X_{t_1}, \cdots, X_{t_k})$ given x_{t_0} is the same as the distribution of $(X_{t_1-t_0}, X_{t_2-t_0}, \cdots, X_{t_k-t_0})$ given x_0 for every k and every $(k+1)$-uplet $t_0 \le t_1 \le \cdots \le t_k$.

So in the case of a Markov chain, if the initial distribution or the initial state is known, the construction of the Markov chain (X_n) is entirely determined by its *transition*, namely by the distribution of X_n conditionally on x_{n-1}.

The study of Markov chains is almost always restricted to the time-homogeneous case and we omit this designation in the following. It is, however, important to note here that an incorrect implementation of Markov Chain Monte Carlo algorithms can easily produce time-heterogeneous Markov chains for which the standard convergence properties do not apply. (See also the case of the ARMS algorithm in §6.3.3)

Example 4.2.4 –**Simulated Annealing**– The *simulated annealing* algorithm (see §5.2.3) is often implemented in a time-heterogeneous form and studied in time-homogeneous form. Given a finite state space with size K, $\Omega = \{1, 2, \cdots, K\}$, an energy function E and a temperature T, the simulated annealing Markov chain $X_0, X_1, \ldots$ is represented by the following transition operator: Conditionally on X_n, Y is generated from a fixed probability distribution $(\pi_1, \cdots, \pi_K)$ on Ω and the new value of the chain is given by

$$X_{n+1} = \begin{cases} Y & \text{with probability } \exp\{(E(Y) - E(X_n))/T\} \wedge 1, \\ X_n & \text{otherwise.} \end{cases}$$

If the temperature T depends on n, the chain is time-heterogeneous. ||

Example 4.2.5 –**AR**(1) **Models**– AR(1) models provide a simple illustration of Markov chains on continuous state space. If

$$X_n = \theta X_{n-1} + \varepsilon_n \ , \qquad \theta \in \mathbb{R},$$

with $\varepsilon_n \sim N(0,\sigma^2)$, and if the ε_n's are independent, X_n is indeed independent from $X_{n-2}, X_{n-3}, \ldots$ conditionally on X_{n-1}. The Markovian properties of an AR(q) process can be derived by considering the vector $(X_n, \cdots, X_{n-q+1})$. On the contrary, ARMA(p,q) models do not fit in the Markovian framework (see Problem 4.2). ||

In the general case, the fact that the kernel K determines the properties of the chain (X_n) can be deduced from the relations

$$\begin{array}{rcl} P_x(X_1 \in A_1) & = & K(x, A_1) \ , \\ P_x((X_1, X_2) \in A_1 \times A_2) & = & \displaystyle\int_{A_1} K(y_1, A_2)\ K(x, dy_1) \\ & \cdots & \\ P_x((X_1, \cdots, X_n) \in A_1 \times \cdots \times A_n) & = & \displaystyle\int_{A_1} \cdots \int_{A_{n-1}} K(y_{n-1}, A_n) \\ & & \times K(x, dy_1) \cdots K(y_{n-2}, dy_{n-1}) \ . \end{array}$$

In particular, the relation $P_x(X_1 \in A_1) = K(x, A_1)$ indicates that $K(x_n, dx_{n+1})$ is a *version* of the conditional distribution of X_{n+1} given X_n. However, as we have defined a Markov chain by first specifying this kernel, we do not need to be concerned with different versions of the conditional probabilities. This is why we noted that constructing the Markov chain through the transition kernel was mathematically "cleaner". (Moreover, in the following chapters, we will see that the objects of interest are often these conditional distributions, and it is important that we need not worry about different versions. Nonetheless, the properties of a Markov chain considered in this chapter are independent of the version of the conditional probability chosen.)

If we denote $K^1(x, A) = K(x, A)$, the kernel for n transitions is given by $(n > 1)$

$$K^n(x, A) = \int_{\mathcal{X}} K^{n-1}(y, A)\ K(x, dy).$$

The following result provides convolution formulas of the type $K^{m+n} = K^m \star K^n$, which are called *Chapman–Kolmogorov equations.*

Lemma 4.2.6 ***Chapman-Kolmogorov equations*** *For every* $(m, n) \in \mathbb{N}^2$, $x \in \mathcal{X}$, $A \in \mathcal{B}(\mathcal{X})$,

$$K^{m+n}(x, A) = \int_{\mathcal{X}} K^n(y, A)\ K^m(x, dy) \ .$$

(In a very informal sense, the Chapman-Kolmogorov equations are stating that to get from x to A in $m + n$ steps, you must pass through some

y on the n^{th} step.) In the discrete case, Lemma 4.2.6 is simply interpreted as a matrix product, and follows directly from (4.2.1). In the general case, we need to consider K as an operator on the space of integrable functions, i.e.

$$Kh(x) = \int h(y)\ K(x, dy)\ , \qquad h \in \mathcal{L}_1(\lambda)\ ,$$

λ being the dominating measure of the model; K^n is then the n-th composition of P, namely $K^n = K \circ K^{n-1}$.

Definition 4.2.7 A *resolvant* associated with the kernel P is a kernel of the form

$$K_\varepsilon(x, A) = (1 - \varepsilon) \sum_{i=0}^{\infty} \varepsilon^i K^i(x, A), \qquad 0 < \epsilon < 1,$$

and the chain with kernel K_ε is said to be a *K_ε-chain.*

Given an initial distribution μ, we can associate with the kernel K_ε a chain $\{X_n^\epsilon\}$ which formally corresponds to a sub-chain of the original chain (X_n), where the indices in the subchain are generated from a geometric distribution with parameter $1 - \varepsilon$. Thus K_ϵ is indeed a kernel, and we will see that the resulting Markov chain (X_n^ϵ) enjoys much stronger regularity. This will be used later to establish many properties of the original chain.

If $\mathbb{E}_\mu[\ \cdot\ ]$ denotes the expectation associated with the distribution P_μ, the *(weak) Markov property* can be written as the following result, which just rephrases the limited memory properties of a Markov chain:

Proposition 4.2.8 ***Weak Markov property*** *For every initial distribution μ and every $(n+1)$ sample $(X_0, \ldots, X_n)$,*

$$\mathbb{E}_\mu[h(X_{n+1}, X_{n+2}, \cdots)|x_0, \cdots, x_n] = \mathbb{E}_{x_n}[h(X_1, X_2, \cdots)], \tag{4.2.4}$$

provided that the expectations exist.

Note that if h is the indicator function then this definition is exactly the same as 4.2.3. However, (4.2.4) can be generalized to other classes of functions—hence the terminology "weak"— and it becomes particularly useful with the notion of *stopping time*, in the convergence control of Markov Chain Monte Carlo algorithms in Chapter 8.

Definition 4.2.9 Consider $A \in \mathcal{B}(\mathcal{X})$. The first n for which the chain enters the set A is denoted by

$$\tau_A = \inf\{n \geq 1; X_n \in A\}, \tag{4.2.5}$$

and is called the *stopping time* at A with, by convention, $\tau_A = +\infty$ if $x_n \notin A$ for every n. More generally, a function $\zeta(x_1, x_2, \ldots)$ is called a *stopping rule* if the set $\{\zeta = n\}$ is measurable for the σ-algebra induced by $(X_0, \cdots, X_n)$. Associated with the set A, we also define

$$\eta_A = \sum_{t=1}^{\infty} \mathbb{I}_A(X_t), \tag{4.2.6}$$

the *number of passages* of (X_n) in A.

is lumpable for $A_1 = \{1,2\}$, $A_2 = \{3,4\}$ and $A_3 = \{5\}$.

4.56 Consider the random walk on $\mathbb{R}^+$, $X_{n+1} = (X_n + \epsilon_n)^+$, with $\mathbb{E}[\epsilon_n] = \beta$.

(a) Establish Lemma 4.9.1. (*Hint:* Consider an alternative V to V^* and show by recurrence that

$$\begin{aligned} V(x) &\geq \int_C K(x,y)V(y)dy + \int_{C^c} K(x,y)V(y)dy \\ &\geq \ldots \geq V^*(x)\,.) \end{aligned}$$

(b) Establish Theorem 4.9.3 by assuming that there exists x^* such that $P_{x^*}(\tau_C < \infty) < 1$, choosing M such that $M \geq V(x^*)/[1 - P_{x^*}(\tau_C < \infty)]$, and establishing that $V(x^*) \geq M[1 - P_{x^*}(\tau_C < \infty)]$.

(c) Show that, if an irreducible Markov chain has a σ-finite invariant measure, this measure is unique up to a multiplicative factor. (*Hint:* Use Theorem 4.7.3.)

(d) (Kemeny and Snell 1960) Show that, for an aperiodic irreducible Markov chain with finite state space and with transition matrix $\mathbb{P}$, there always exists a stationary probability distribution which satisfies

$$\pi = \pi \mathbb{P}.$$

(e) Show that, if $\beta < 0$, the random walk is recurrent. (*Hint:* Use the drift function $V(x) = x$.)

(f) Show that, if $\beta = 0$ and $\text{var}(\epsilon_n) < \infty$, (X_n) is recurrent. (*Hint:* Use $V(x) = \log(1+x)$ for $x > R$ and $V(x) = 0$, otherwise, for an adequate bound R.)

(g) Show that, if $\beta > 0$, the random walk is transient.

4.57 Show that, if there exist a finite potential function V and a small set C such that V is bounded on C and satisfies (4.9.3), the corresponding chain is Harris positive.

4.58 Show that the random walk on $\mathbb{Z}$ is transient when $\mathbb{E}[W_n] \neq 0$. (*Hint:* Use $V(x) = 1 - \varrho^x$ for $x > 0$ and 0 otherwise when $\mathbb{E}[W_n] > 0$.)

4.59 Show that the chains defined by the kernels (4.9.9) and (4.9.11) are either both recurrent or both transient.

4.60 (Chan and Geyer 1994) Prove that the following Central Limit Theorem can be considered a corollary to Theorem 4.9.12 (see Note 4.9.3).

Corollary 4.8.1 *Suppose that the stationary Markov chain (X_n) is geometrically ergodic with $M^* = \int |M(x)| f(x)dx < \infty$, and satisfies the moment conditions of Theorem 4.9.12. Then $\sigma^2 = \lim_{n\to\infty} n\,\text{var}\bar{X}_n < \infty$ and, if $\sigma^2 > 0$, $\sqrt{n}\bar{X}_n/\sigma$ tends in law to $\mathcal{N}(0,\sigma^2)$.*

(*Hint*: Integrate (with respect to f) both sides of Definition 4.6.8 to conclude that the chain is exponentially fast α-mixing, and apply Theorem 4.9.12.)

4.9 Notes

4.9.1 Drift conditions

Besides atoms and small sets, Meyn and Tweedie (1993) rely on another tool to check or establish various stability results, namely the *drift criteria* which can be traced back to Lyapunov. Given a function V on $\mathcal{X}$, the *drift of* V is defined by

$$\Delta V(x) = \int V(y)\ P(x,dy) - V(x)\ .$$

This notion is also used in the following chapters to verify the convergence properties of some MCMC algorithms (see, e.g., Theorem 6.3.7 or Mengersen and Tweedie 1996).

The following lemma is instrumental in deriving drift conditions for the transience or the recurrence of a chain (X_n).

Lemma 4.9.1 *If $C \in \mathcal{B}(\mathcal{X})$, the smallest positive function which satisfies the conditions*

$$\Delta V(x) \leq 0 \ \ \textit{if}\ \ x \notin C,\ \ V(x) \geq 1 \ \ \textit{if}\ \ x \in C \tag{4.9.1}$$

is given by

$$V^*(x) = P_x(\sigma_C < \infty)\ ,$$

when σ_C denotes

$$\sigma_C = \inf\{n \geq 0; x_n \in C\}\ .$$

Note that, if $x \notin C$, $\sigma_C = \tau_C$, while $\sigma_C = 0$ on C. We then have the following necessary and sufficient condition:

Theorem 4.9.2 *The ψ-irreducible chain (X_n) is transient if, and only if, there exist a bounded positive function V and a real number $r \geq 0$ such that every x for which $V(x) > r$, we have*

$$\Delta V(x) > 0\ . \tag{4.9.2}$$

Proof. If $C = \{x; V(x) \leq r\}$ and M is a bound on V, the conditions (4.9.1) are satisfied by

$$\tilde{V}(x) = \begin{cases} (M - V(x))/(M-r) & \text{if } x \in C^c, \\ 1 & \text{if } x \in C. \end{cases}$$

Since $\tilde{V}(x) < 1$ for $x \in C^c$, $V^*(x) = P_x(\tau_C < \infty) < 1$ on C^c, and this implies the transience of C, therefore the transience of (X_n). The converse can be deduced from a (partial) converse to Proposition 4.4.7 (see Meyn and Tweedie 1993, p.190). □

Condition (4.9.2) describes an average increase of $V(x_n)$ once a certain level has been attained, and therefore does not allow a sure return to 0 of V. The condition is thus incompatible with the stability associated with recurrence. On the other hand, if there exists a potential function V "attracted" to 0, the chain is recurrent.

Theorem 4.9.3 *Consider (X_n) a ψ-irreducible Markov chain. If there exist a small set C and a function V such that*

$$C_V(n) = \{x; V(x) \leq n\}, \qquad \forall n$$

is a small set, the chain is recurrent if

$$\Delta V(x) \leq 0 \ \textit{on}\ C^c.$$

Theorem 4.9.8 *If the ergodic chain (X_n) with invariant distribution π satisfies conditions (4.9.8), for every function g such that $|g| \leq f$, then*

$$\begin{aligned}\gamma_g^2 &= \lim_{n\to\infty} n\mathbb{E}_\pi[S_n^2(\overline{g})] \\ &= \mathbb{E}_\pi[\overline{g}^2(x_0)] + 2\sum_{k=1}^{\infty} \mathbb{E}_\pi[\overline{g}(x_0)\overline{g}(x_k)]\end{aligned}$$

is nonnegative and finite. If $\gamma_g > 0$, the Central Limit Theorem holds for $S_n(\overline{g})$. If $\gamma_g = 0$, $\sqrt{n}S_n(\overline{g})$ almost surely goes to 0.

This theorem is definitely relevant for convergence control of Markov Chain Monte Carlo algorithms since, when $\gamma_g^2 > 0$, it is possible to assess the convergence of the ergodic averages $S_n(g)$ to the quantity of interest $\mathbb{E}^\pi[g]$. Theorem 4.9.8 also suggests how to implement this control through renewal theory, as discussed in detail in Chapter 8.

4.9.2 Eaton's Admissibility Condition

Eaton (1992) exhibits interesting connections, similar to Brown (1971), between the admissibility of an estimator and the recurrence of an associated Markov chain. The problem considered by Eaton (1992) is to determine whether, for a *bounded* function $g(\theta)$, a generalized Bayes estimator associated with a prior measure π is admissible under quadratic loss. Assuming that the posterior distribution $\pi(\theta|x)$ is well defined, he introduces the transition kernel

$$K(\theta,\eta) = \int_{\mathcal{X}} \pi(\theta|x) f(x|\eta)\, dx, \tag{4.9.9}$$

which is associated with a Markov chain $(\theta^{(n)})$ generated as follows: the transition from $\theta^{(n)}$ to $\theta^{(n+1)}$ is done by generating first $x \sim f(x|\theta^{(n)})$ and then $\theta^{(n+1)} \sim \pi(\theta|x)$. (Most interestingly, this is also a kernel used by Markov Chain Monte Carlo methods, as shown in Chapter 7.) Note that the prior measure π is an invariant measure for the chain $(\theta^{(n)})$. For every measurable set C such that $\pi(C) < +\infty$, consider

$$V(C) = \left\{h \in \mathcal{L}^2(\pi); h(\theta) \geq 0 \text{ and } h(\theta) \geq 1 \text{ when } \theta \in C\right\}$$

and

$$\Delta(h) = \int\int \{h(\theta) - h(\eta)\}^2\, K(\theta,\eta)\pi(\eta)\, d\theta\, d\eta.$$

The following result then characterizes admissibility for *all bounded functions* in terms of Δ and $V(C)$, i.e., independently of the estimated functions g:

Theorem 4.9.9 *If, for every C such that $\pi(C) < +\infty$,*

$$\inf_{h\in V(C)} \Delta(h) = 0, \tag{4.9.10}$$

then the Bayes estimator $\mathbb{E}^\pi[g(\theta)|x]$ is admissible under quadratic loss for every bounded function g.

This result is obviously quite general but only mildly helpful in the sense that the practical verification of (4.9.10) for every set C can be overwhelming. Note also that (4.9.10) always holds when π is a proper prior distribution since $h \equiv 1$ belongs to $\mathcal{L}^2(\pi)$ and $\Delta(1) = 0$ in this case. The extension then considers approximations of 1 by functions in $V(C)$. Eaton (1992) exhibits a connection with the Markov chain $(\theta^{(n)})$ which gives a condition equivalent to Theorem 4.9.9. First, for a given set C, a stopping rule σ_C is defined as the first integer $n > 0$ such that $(\theta^{(n)})$ belongs to C (and $+\infty$ otherwise).

Theorem 4.9.10 *For every set C such that $\pi(C) < +\infty$,*

$$\inf_{h \in V(C)} \Delta(h) = \int_C \left\{1 - P(\sigma_C < +\infty | \theta^{(0)} = \eta)\right\} \pi(\eta)\, d\eta.$$

Therefore, the generalized Bayes estimators of bounded functions of θ are admissible if, and only if, the associated Markov chain $(\theta^{(n)})$ is recurrent.

Again, we refer to Eaton (1992) for extensions, examples, and comments on this result. Note, however, that the verification of the recurrence of the Markov chain $(\theta^{(n)})$ is much easier to operate than the determination of the lower bound of $\Delta(h)$. Hobert and Robert (1997) consider the potential use of the dual chain based on the kernel

$$K'(x,y) = \int_\Theta f(y|\theta)\pi(\theta|x) d\theta \tag{4.9.11}$$

(see Problem 4.59) and derive admissibility results for various distributions of interest.

4.9.3 Mixing Conditions and Central Limit Theorems

In §4.7.2 we established a central limit theorem using regeneration, which allowed us to use a typical independence argument. Other conditions, known as *mixing conditions*, can also result in a Central Limit Theorem. These mixing conditions guarantee that the dependence in the Markov chain decreases fast enough, and variables that are far enough apart are close to being independent. Unfortunately, these conditions are usually quite difficult to verify. Consider the property of *α-mixing* (Billingsley 1995, Section 27).

Definition 4.9.11 A sequence $X_0, X_1, X_2, \cdots$ is *α-mixing* if

$$\alpha_n = \sup_{A,B} |P(X_n \in A, X_0 \in B) - P(X_n \in A)P(X_0 \in B)| \tag{4.9.12}$$

goes to 0 when n goes to infinity.

So we see that an α-mixing sequence will tend to "look independent" if the variables are far enough apart. As a result, we would expect that Theorem 4.7.3 is a consequence of α-mixing. This is in fact the case, as every positive recurrent aperiodic Markov chain is α-mixing (Rosenblatt 1971, Section VII.3), and if the Markov chain is stationary and α-mixing, the covariances go to zero (Billingley 1995, Section 27).

However, for a Central Limit Theorem we need even more. Not only must the Markov chain be α-mixing, we need the coefficient α_n to go to 0 fast enough. That is, we need the dependence to go away fast enough. One version of a Markov chain Central Limit Theorem is the following (Billingsley 1995, Section 27):

Theorem 4.9.12 *Suppose that the Markov chain (X_n) is stationary and α-mixing with $\alpha_n = O(n^{-5})$ and that $\mathbb{E}[X_n] = 0$ and $\mathbb{E}[X_n^{12}] < \infty$. Then $\sigma^2 = \lim_{n\to\infty} n \operatorname{var}\bar{X}_n < \infty$ and, if $\sigma^2 > 0$, $\sqrt{n}\bar{X}_n$ tends in law to $\mathcal{N}(0,\sigma^2)$.*

CHAPTER 5

Monte Carlo Optimization

5.1 Introduction

Similar to the problem of integration, differences between the numerical approach and the simulation approach to the problem

$$\max_{\theta \in \Theta} h(\theta) \tag{5.1.1}$$

lie in the treatment of the function[1] h. (Note that (5.1.1) also covers minimization problems by considering $-h$.) In approaching a minimization problem using deterministic numerical methods, the analytical properties of the target function (convexity, boundedness, smoothness) are often paramount. For the simulation approach, we are more concerned with h from a probabilistic (rather than analytical) point of view. Obviously, this dichotomy is somewhat artificial, as there exist simulation approaches where the probabilistic interpretation of h is not used. Nonetheless, the use of the analytical properties of h plays a lesser role in the simulation approach.

By comparison with numerical methods, which enjoy a longer history than simulation methods (see, for instance Kennedy and Gentle 1980, Ciarlet and Thomas 1982, Sakarovitch 1984, Thisted 1988), the appeal of simulation can be found in the lack of constraints on both the regularity of the domain Θ and on the function h. Obviously, there may exist an alternative numerical approach which provides an exact solution to (5.1.1), a property rarely achieved by a stochastic algorithm, but simulation has the advantage of bypassing the preliminary steps of devising an algorithm and studying whether some regularity conditions on h hold . This is particularly true when the function h is very costly to compute.

Example 5.1.1 –**Signal processing**– Ó Ruanaidh and Fitzgerald (1996) study signal processing data, of which a simple model is $(i = 1, \ldots, N)$

$$x_i = \alpha \cos(\omega t_i) + \beta \sin(\omega t_i) + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2),$$

[1] Although we use θ as the running parameter, and h typically corresponds to a possibly penalized transform of the likelihood function, this setup applies to many other inferential problems than just likelihood or posterior maximization. As noted in the introduction to Chapter 3, complex loss functions but also confidence region also require optimization procedures.

with unknown parameters $\alpha, \beta, \omega, \sigma$ and observation times $t_1, \ldots, t_N$. The likelihood function is then of the form

$$\sigma^{-N} \exp\left(-\frac{(x - G\binom{\alpha}{\beta})^t (x - G\binom{\alpha}{\beta})}{2\sigma^2}\right),$$

with $x = (x_1, \ldots, x_N)$ and

$$G = \begin{pmatrix} \cos(\omega t_1) & \sin(\omega t_1) \\ \vdots & \vdots \\ \cos(\omega t_N) & \sin(\omega t_N) \end{pmatrix}.$$

The prior $\pi(\alpha, \beta, \omega, \sigma) = \sigma^{-1}$ then leads to the marginal distribution

$$(5.1.2) \quad \pi(\omega|x) \propto \left(x^t x - x^t G (G^t G)^{-1} G^t x\right)^{(2-N)/2} (\det G^t G)^{-1/2},$$

which, although explicit in ω, is not particularly simple to compute. This setup is also illustrative of functions with many modes, as shown by Ó Ruanaidh and Fitzgerald (1996). $\|$

Following Geyer (1996), we want to consider two approaches to Monte Carlo optimization. The first is an *exploratory* approach, in which the goal is to optimize the function h by describing its entire range. The actual properties of the function play a lesser role here, with the Monte Carlo aspect more closely tied to the exploration of Θ, even though the slope of h can be used to speed up the exploration. (Such a technique can be useful in describing functions with multiple modes, for example.) The second approach is based on a probabilistic *approximation* of the objective function h and is rather a preliminary step to the optimization *per se*. Here, the Monte Carlo aspect exploits the probabilistic properties of the function h to come up with an acceptable approximation, and is less concerned with exploring Θ. We will see that this approach can be tied to *missing data methods*, such as the EM algorithm. We note also that Geyer (1996) only considers the second approach to be "Monte Carlo optimization". Obviously, even though we are considering these two different approaches separately, they might be combined in a given problem. In fact, methods like the EM (5.3.3) or the Robbins-Monro (5.3.5) algorithms take advantage of the Monte Carlo approximation to enhance their particular optimization technique.

5.2 Stochastic Exploration

5.2.1 A basic solution

There are a number of cases where the exploration method is particularly well-suited. First, if Θ is bounded, which can often be achieved by a reparameterization, a first approach to the resolution of (5.1.1) is to simulate from a uniform distribution on Θ, $u_1, \ldots, u_m \sim \mathcal{U}_\Theta$, and to use the approximation $h_m^* = \max(h(u_1), \ldots, h(u_m))$. This method is convergent (as

m goes to ∞) but it may be very slow since it does not take into account any specific features of h. Distributions other than the uniform which are related with h may then do better. In particular, in setups where the likelihood function is extremely costly to compute, the number of evaluations should be kept to a minimum.

Therefore, a second, and more fruitful, direction consists in relating h to a probability distribution. For instance, if h is positive and if

$$\int_\Theta h(\theta)\, d\theta < +\infty ,$$

the resolution of (5.1.1) amounts to finding the *modes* of the density h. More generally, if these conditions are not satisfied, the function $h(\theta)$ can be transformed into a positive and integrable function $H(\theta)$ in such a way that the solutions to (5.1.1) are those which maximize $H(\theta)$ on Θ. For example, we can take $H(\theta) = \exp(h(\theta)/T)$ or $H(\theta) = \exp\{h(\theta)/T\}/(1 + \exp\{h(\theta)/T\})$ and choose T to accelerate convergence or to avoid local maxima (as in simulated annealing, see §sec:2.3.3). When the problem is expressed in statistical terms, it becomes natural to then generate a sample $(\theta_1, \ldots, \theta_m)$ from h (or H) and to apply a standard mode estimation method (or to simply compare the $h(\theta_i)$'s). In some cases, it may be more useful to decompose $h(\theta)$ into $h(\theta) = h_1(\theta)h_2(\theta)$ and to simulate from h_1.

Example 5.2.1 –**Minimization**– Consider the function in $\mathbb{R}^2$

$$\begin{aligned} h(x,y) &= (x\sin(20y) + y\sin(20x))^2 \cosh(\sin(10x)x) \\ &+ (x\cos(10y) - y\sin(10x))^2 \cosh(\cos(20y)y) \end{aligned}$$

to be minimized. Since this function has many local minima, as shown by Figure 5.2.1, it does not satisfy the conditions under which standard minimization methods are guaranteed to provide the local minimum. On the other hand, the distribution on $\mathbb{R}^2$ with density proportional to $\exp(-h(x,$ $y))$ can be simulated, even though this is not a standard distribution, via Markov Chain Monte Carlo techniques, and a convergent approximation of the minimum of $h(x,y)$ can be derived from the minimum of the resulting $h(x_i, y_i)$'s. An alternative is to simulate from the density proportional to

$$h_1(x,y) = \exp\{-(x\sin(20y) + y\sin(20x))^2 - (x\cos(10y) - y\sin(10x))^2\},$$

which eliminates the computation of both cosh and sinh in the simulation step. ||

Exploration may be particularly difficult when the space Θ is not convex (or perhaps not even connected), and the simulation of the sample $(\theta_1, \ldots, \theta_m)$ can be much faster than a numerical method applied to (5.1.1). The appeal of simulation is even clearer in the case when $h(\theta) = \mathbb{E}[H(x,\theta)]$, for $X \sim f(x|\theta)$. If it is possible to simulate from the *density* $H(x,\theta)$, the solution of (5.1.1) is the mode of the marginal distribution of θ. (This setting may sound very contrived or even artificial but we will see in §5.3.1 that it includes the case of missing data models.)

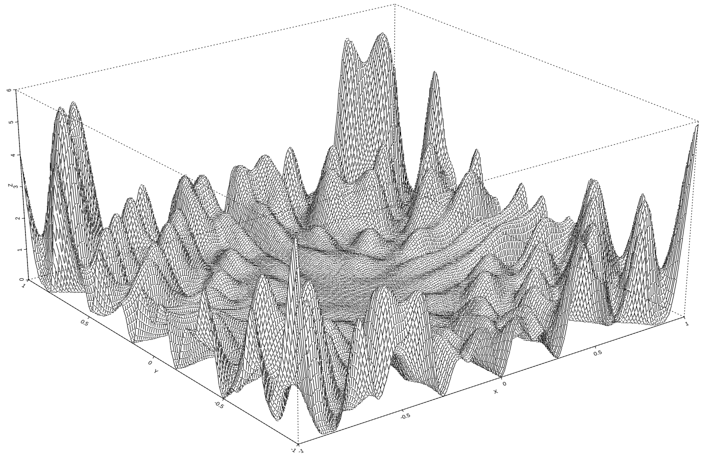


Figure 5.2.1. *Grid representation of the function $h(x,y)$ of Example 5.2.1 on* $[-1,1]^2$.

We now look at several methods to find maxima that can be classified as *exploratory methods.*

5.2.2 Gradient Methods

As mentioned in §1.6, the *gradient method* is a deterministic approach, in numerical analysis, to the problem (5.1.1) which produces a sequence (θ_j) converging to the exact solution of (5.1.1), θ^*, when the domain $\Theta \subset \mathrm{I\!R}^d$ and the function $(-h)$ are convex. The sequence (θ_j) is constructed in a recursive manner through

$$\theta_{j+1} = \theta_j + \alpha_j \nabla h(\theta_j) \ , \qquad \alpha_j > 0 \ , \tag{5.2.1}$$

where ∇h is the gradient of h. For various choices of the sequence (α_j) (see Ciarlet and Thomas 1982), the algorithm converges to the (unique) maximum.

In more general setups, that is when the function or the space is less regular, (5.2.1) can be modified by stochastic perturbations to achieve again convergence properties, as described in detail in Rubinstein (1981) or Duflo (1996, pp. 61–63). One of these stochastic modifications is to choose a second sequence (β_j) to define the chain (θ_j) by

$$\theta_{j+1} = \theta_j + \frac{\alpha_j}{2\beta_j} \, \Delta h(\theta_j, \beta_j \zeta_j) \, \zeta_j \ , \tag{5.2.2}$$

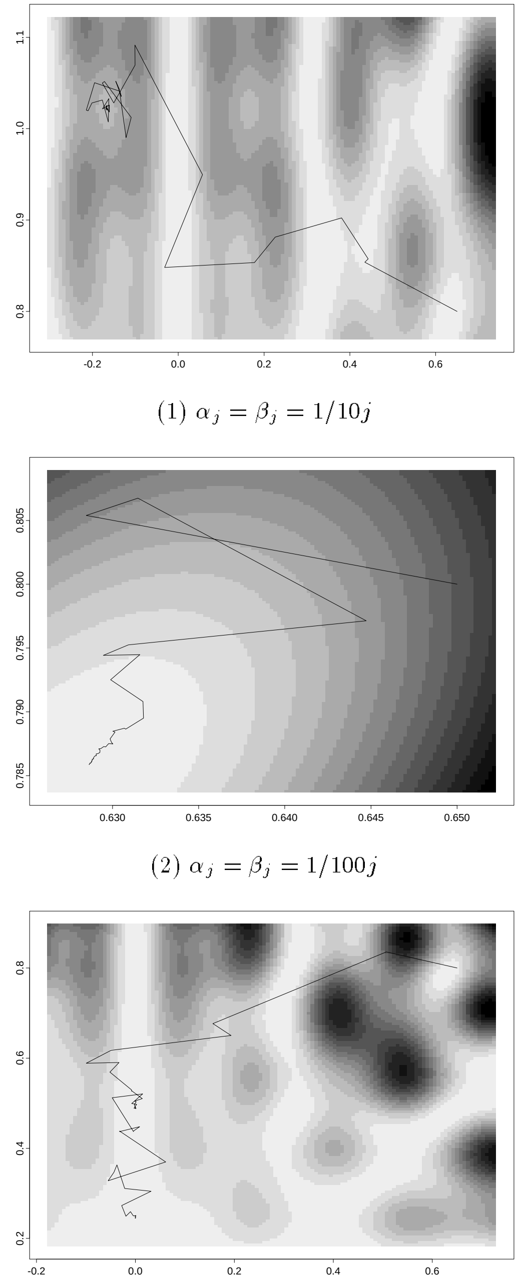
with ζ_j uniformly distributed on the unit sphere $||\zeta|| = 1$ and $\Delta h(x,y) = h(x+y) - h(x-y)$ which approximates $2||y|| \nabla h(x)$. Contrary to the deterministic approach, this method does not necessarily proceed along the steepest slope in θ_j but this property is a *plus* in the sense that it may avoid traps in local maxima or in saddlepoints of h. The convergence of (θ_j) to the solution θ^* again depends on the choice of (α_j) and (β_j). Note at this stage that (θ_j) can be seen as a *non-homogeneous Markov chain* which almost surely converges to a given value; the study of these chains is particularly arduous given their ever-changing transition kernel (see Winkler 1996, for some results in this direction). However, sufficiently strong conditions such as the decrease of α_j towards 0 and of α_j/β_j to a non-zero constant are enough to guarantee the convergence of the sequence (θ_j).

Example 5.2.2 (Continuation of Example 5.2.1) We can apply the iterative construction (5.2.2) to the multi-modal function $h(x, y)$ with different sequences of α_j's and β_j's. Figure 5.2.2 and Table 5.2.2 illustrate the convergence of the algorithm to different local minima of the function h, with occurrences where the sequence $h(\theta_j)$ increases and avoids other local minima. The solutions are quite distinct for the three different sequences, both in location and values. As shown by Table 5.2.2, the number of iterations needed to achieve stability of θ_T also varies with the choice of (α_j, β_j). Note that Case 1 ends up with a very poor evaluation of the minimum, due to a fast decrease of (α_j) associated with big jumps in the first iterations, while Case 2 converges to the closest local minima. Case 3 illustrates a general feature of the stochastic gradient method, namely that slower decrease rates of the sequence (α_j) tend to achieve better minima. The final convergence along a valley of h after some initial big jumps is also noteworthy. $\|$

Table 5.2.1. *Results of three stochastic gradient runs for the minimisation of the function h in Example 5.2.1 with different values of (α_j, β_j) and starting point $(.65, .8)$. The iteration T is obtained by the stopping rule $\|\theta_T - \theta_{T-1}\| < 10^{-5}$.*

α_j	β_j	θ_T	$h(\theta_T)$	$\min_t h(\theta_t)$	iteration
$1/10j$	$1/10j$	$(-0.166, 1.02)$	1.287	0.115	50
$1/100j$	$1/10j$	$(0.629, 0.786)$	0.00013	0.00013	93
$1/10\log(1+j)$	$1/j$	$(0.0004, 0.245)$	$4.24e-06$	$2.163e-07$	58

This approach is still (too) close to numerical methods in that it requires a precise knowledge on the function h, which is not necessarily available.



(1) $\alpha_j = \beta_j = 1/10j$

(2) $\alpha_j = \beta_j = 1/100j$

(3) $\alpha_j = 1/10 \log(1+j),\ \beta_j = 1/j$

Figure 5.2.2. *Stochastic gradient paths for three different choices of the sequences* (α_j) *and* (β_j) *and starting point* $(.65, .8)$ *for the same sequence* (ζ_j) *in (5.2.2). The grey levels are such that darker shades mean higher elevations. The function* h *to minimize is defined in Example5.2.1.*

5.2.3 Simulated Annealing

The simulated annealing algorithm[2] has been introduced by Metropolis (1953) et al. to minimize a criterion function on a finite set with very large size[3], but it also applies to optimization on a continuous set and to simulation (see Ackley et al. 1985 and Neal 1994).

The fundamental idea at the core of simulated annealing methods is that a change of scale, called *temperature*, allows for faster moves on the surface of the function h to maximize, whose negative is called *energy*. Therefore, rescaling partially avoids the trapping attraction of local maxima. Given a temperature parameter $T > 0$, a sample $\theta_1^T, \theta_2^T, \ldots$ is generated from the distribution

$$\pi(\theta) \propto \exp(h(\theta)/T)$$

and can be used as in §5.2.1 to come up with an approximate maximum of h. As T decreases towards 0, the values simulated from this distribution become concentrated in a narrower and narrower neighborhood of the local maxima of h (see Theorem 5.2.7, Problem 5.9, and Winkler 1996). The fact that this approach has a moderating effect on the attraction of the local maxima of h becomes more apparent when we consider the simulation method proposed by Metropolis et al. (1953). Starting from θ_0, ζ is generated from a uniform distribution on a neighborhood $\mathcal{V}(\theta_0)$ of θ_0 or, more generally, from a distribution with density $g(|\zeta - \theta_0|)$, and the new value of θ is generated as follows:

$$\theta_1 = \begin{cases} \zeta & \text{with probability } \rho = \exp(\Delta h/T) \wedge 1, \\ \theta_0 & \text{with probability } 1 - \rho, \end{cases}$$

where $\Delta h = h(\zeta) - h(\theta_0)$. (This method is in fact *the Metropolis algorithm*, which simulates the density proportional to $\exp\{h(\theta)/T\}$, described and justified in Chapter 6.) Therefore, if $h(\zeta) \geq h(\theta_0)$, ζ is accepted with probability 1. On the other hand, if $h(\zeta) < h(\theta_0)$, ζ may still be accepted with probability $\rho \neq 0$ and θ_0 is then changed into ζ. This property allows the algorithm to escape the attraction of θ_0 if θ_0 is a local maximum of h, with a probability which depends on the choice of the scale T, compared with the range of the density g.

In its most usual implementation, the simulated annealing algorithm modifies the temperature T at each iteration; it is then of the form

Algorithm A.20 –***Simulated Annealing***–

[2] This name is borrowed from the metallurgy vocabulary: A metal manufactured by a slow decrease of the temperature (*annealing*) is stronger than a metal manufactured by a fast decrease of the temperature. The vocabulary also relates to Physics, since the function to minimize is called *energy* and the variance factor T which controls convergence *temperature*. We will try to keep these idiosyncrasies to a minimal level, but they are quite common in the literature.

[3] This paper is also the originator of the *Markov Chain Monte Carlo methods* developed in the following chapters. The potential of these two simultaneous innovations has been discovered much latter by statisticians (Hastings 1970; Geman and Geman 1984) than by of physicists (see also Kirkpatrick et al. 1983).

where

$$t_i(\zeta,\sigma^2) = \mathbb{E}[Z_i|X_i,\zeta,\sigma^2] \quad \text{and} \quad v_i(\zeta,\sigma^2) = \mathbb{E}[Z_i^2|X_i,\zeta,\sigma^2]$$

(d) Show that

$$\begin{aligned} \mathbb{E}[Z_i|X_i,\zeta,\sigma^2] &= \zeta + \sigma H_i\left(\frac{u-\zeta}{\sigma}\right) \\ \mathbb{E}[Z_i^2|X_i,\zeta,\sigma^2] &= \zeta^2+\sigma^2+\sigma(u+\zeta)H_i\left(\frac{u-\zeta}{\sigma}\right) \end{aligned}$$

where

$$H_i(t) = \begin{cases} \frac{\varphi(t)}{1-\Phi(t)} & \text{if } X_i = 1 \\ -\frac{\varphi(t)}{\Phi(t)} & \text{if } X_i = 0. \end{cases}$$

(e) Show that $\hat{\zeta}_{(j)}$ converges to $\hat{\zeta}$ and that $\hat{\sigma}^2_{(j)}$ converges to $\hat{\sigma}^2$, the ML estimates of ζ and σ^2.

5.18 The EM algorithm can also be implemented in a Bayesian hierarchical model to find a posterior mode. Suppose that we have the hierarchical model

$$\begin{aligned} X|\theta &\sim f(x|\theta)\ , \\ \theta|\lambda &\sim \pi(\theta|\lambda)\ , \\ \lambda &\sim \gamma(\lambda)\ , \end{aligned}$$

where interest would be in estimating quantities from $\pi(\theta|x)$. Since

$$\pi(\theta|x) = \int \pi(\theta,\lambda|x)d\,\lambda,$$

where $\pi(\theta,\lambda|x) = \pi(\theta|\lambda,x)\pi(\lambda|x)$, the EM algorithm is a candidate method for finding the mode of $\pi(\theta|x)$, where λ would be used as the augmented data.

(a) Define $k(\lambda|\theta,x) = \pi(\theta,\lambda|x)/\pi(\theta|x)$, and show that

$$\log \pi(\theta|x) = \int \log \pi(\theta,\lambda|x)k(\lambda|\theta^*,x)d\,\lambda - \int \log k(\lambda|\theta,x)k(\lambda|\theta^*,x)d\,\lambda.$$

(b) If the sequence $(\hat{\theta}_{(j)})$ satisfies

$$\max_\theta \int \log \pi(\theta,\lambda|x)k(\lambda|\theta_{(j)},x)d\,\lambda = \int \log \pi(\theta_{(j+1)},\lambda|x)k(\lambda|\theta_{(j)},x)d\,\lambda,$$

show that $\log \pi(\theta_{(j+1)}|x) \geq \log \pi(\theta_{(j)}|x)$. Under what conditions will the sequence $(\hat{\theta}_{(j)})$ converge to the mode of $\pi(\theta|x)$?

(c) For the hierarchy

$$\begin{aligned} X|\theta &\sim N(\theta,1)\ , \\ \theta|\lambda &\sim N(\lambda,1)\ , \end{aligned}$$

with $\pi(\lambda) = 1$, show how to use the EM algorithm to calculate the posterior mode of $\pi(\theta|x)$.

CHAPTER 6

The Metropolis-Hastings Algorithm

> ”What’s changed, except what needed changing?” And there was something in that, Cadfael reflected. What was changed was the replacement of falsity by truth, and however hard the assimilation might be, it must be for the better. Truth can be costly, but in the end, it never falls short of value for the price paid.
>
> —Ellis Peter, *The Confession of Brother Haluin*—

> ‘Have you any thought’, resumed Valentin, ‘of a tool with which it could be done?’
>
> ‘Speaking within modern probabilities, I really haven’t,’ said the doctor.
>
> —G.K. Chesterton, *The Innocence of Father Brown*—

6.1 Monte Carlo Methods based on Markov Chains

Chapter 3 has shown that it is not necessary to use a sample from the distribution f to approximate the integral

$$\int h(x)f(x)dx \ ,$$

since *importance sampling* techniques can be used. This chapter develops this possibility in a different way and shows that it is possible to obtain a sample $x_1, \cdots, x_n$ distributed from f without simulating from f. The basic principle underlying the methods described in this chapter and the following chapters is *to use an ergodic Markov chain with stationary distribution* f: for an arbitrary starting value $x^{(0)}$, a chain $(X^{(t)})$ is generated from a transition kernel with stationary distribution f, which moreover ensures the convergence in distribution of $(X^{(t)})$ to f. (Given that the chain is ergodic, the starting value $x^{(0)}$ is, in principle, unimportant.) For instance, for a “large enough” T, $X^{(T)}$ can thus be considered as distributed from f and the methods studied in this chapter produce a sample $X^{(T)}, X^{(T+1)}, \ldots$, which is generated from f, even though the $X^{(T+t)}$’s are not independent.

Definition 6.1.1 A *Markov chain Monte Carlo (MCMC) method* for the simulation of a distribution f is any method producing an ergodic Markov chain $(X^{(t)})$ whose stationary distribution is f.

Why should we resort to such a convoluted approach to simulate from f? In comparison with the techniques developed in Chapter 3, this involved strategy may indeed sound suboptimal, as it relies on asymptotic convergence properties. As a corollary, the number of iterations required to obtain a good approximation of f is *a priori* important. The call for Markov chains is nonetheless justified from at least two points of view: First, Chapter 5 has shown that stochastic optimization algorithms as those of Robbins-Monro or the SEM algorithm naturally produce Markov chain structures which should be generalized and, if possible, optimized. Second, the call to Markov chains allows for a much greater generality than the methods presented in Chapter 2, where we have produced only one single general method of simulation, the ARS algorithm (§2.2.4), which moreover only applies for log-concave densities. Markov Chain Monte Carlo methods achieve a "universal" dimension in the sense that they (and not only formally) validate the use of positive densities g for the simulation of arbitrary distributions of interest f. Moreover, even when an accept-reject algorithm is available, it is sometimes more efficient to use the pair (f, g) through a Markov chain, as detailed in §6.3.1.

Thus, even if the extension proposed in Definition 6.1.1 may seem purely formal at this stage, the (re-)discovery of Markov Chain Monte Carlo methods by the statisticians in the 1990's has undoubtedly induced considerable progress in simulation-based inference and, in particular, in Bayesian inference, since it has opened access to the analysis of models which were too complex to be satisfactorily processed by previous schemes[1]. This chapter covers the most general MCMC method, namely the *Metropolis–Hastings algorithm*, while Chapter 7 specializes in *the Gibbs sampler* which, although a particular case of Metropolis–Hastings algorithm (see Theorem 7.1.11), has fundamentally different methodological and historical motivations.

How should we implement the principle brought forward by Definition 6.1.1? Despite its formal aspect, this definition implies that the use of a chain $(X^{(t)})$ resulting from a Markov Chain Monte Carlo algorithm with stationary distribution f is similar to the use of an iid sample from f in the sense that the ergodic theorem (Theorem 4.7.3) guarantees the convergence of the empirical average

$$\frac{1}{T} \sum_{t=1}^{T} h(x^{(t)}) \tag{6.1.1}$$

to the quantity $\mathbb{E}_f[h(X)]$. A sequence $(X^{(t)})$ produced by a Markov Chain Monte Carlo algorithm can thus be employed just as an iid sample.

Therefore, if there is no particular requirement on independence but if the incentive for the simulation study is rather on the properties of the distribution f, there is no need for the generation of n independent chains

[1] For example, Chapter 9 considers the case of latent variable models, which prohibit both analytic processing and numerical approximation in both classical (maximum likelihood) and Bayesian setups (see also Example 1.1.2).

$(X_i^{(t)})$ $(i = 1, \ldots, n)$ where the "terminal" values $X_i^{(T)}$ only are kept.[2] In other words, a single Markov chain is enough to ensure a proper approximation through estimates like (6.1.1) of $\mathbb{E}_f[h(X)]$ for the functions h of interest (and sometimes even of the density f, as detailed in Chapter 7). Obviously, handling this sequence is rather more arduous than in the iid case because of the dependence structure, and some approaches to the convergence control of (6.1.1) are given in §6.4 and in Chapter 8.

Which transition should we use, then? Given the principle stated in Definition 6.1.1, one can propose a infinite number of practical implementations based, for instance, on methods used in statistical physics. The Metropolis–Hastings algorithms proposed in this chapter have the definite advantage of imposing minimal requirements on the study of the density f, while allowing for a wide choice of possible implementations, in sharp contrast with the Gibbs sampler given in Chapter 7, which is much more restrictive. Introduced by Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953) in a setup of optimization on a discrete state space, these methods have later been generalized by Hastings (1970) and Peskun (1973) to statistical simulation. Despite several later attempts in specific settings (see for example Geman and Geman 1984, Tanner and Wong 1987, Besag 1989), the starting point for an intensive use of these methods by the statistical community can be traced to the presentation of the *Gibbs sampler* by Gelfand and Smith (1990). This gap of more than thirty years can be partially attributed to the lack of appropriate computing power since most of the examples now processed by Markov Chain Monte Carlo algorithms could not have been treated previously.

6.2 The Metropolis–Hastings algorithm

Before illustrating the universality of Metropolis–Hastings algorithms and demonstrating their straightforward implementation, we first evacuate the (important) issue of their theoretical validity. Since the results presented below are valid for all types of Metropolis–Hastings algorithms, we do not include examples in this section, but rather wait for §6.3, which presents a typology of these types.

6.2.1 Definition

The Metropolis–Hastings algorithm starts with a conditional density $q(y|x)$, defined with respect to the dominating measure for the model (see §6.3.4 for a non-trivial example). It can only be implemented in practice when $q(\cdot|x)$ is easy to simulate and is either explicitly available (up to a multiplicative constant *independent from* x), or *symmetric*, that is such that $q(x|y) = q(y|x)$.

[2] For one thing, the determination of the "proper" length T is still under debate, as we will see in Chapter 8. For another, this approach induces a considerable waste of $n(T-1)$ simulations out of nT.

The Metropolis–Hastings algorithm associated with the objective distribution f and the conditional distribution q produces a Markov chain $(X^{(t)})$ through the following transition:

Algorithm A.24 ***Metropolis–Hastings algorithm***

```
Given $x^{(t)}$,
1. Generate  $Y_t \sim q(y|x^{(t)})$.
2. Take
       $X^{(t+1)} = \begin{cases} Y_t & \text{with probability } \rho(x^{(t)}, Y_t), \\ x^{(t)} & \text{with probability } 1 - \rho(x^{(t)}, Y_t), \end{cases}$
   where                                                          $[A.24]$
       $\rho(x, y) = \min\left\{ \frac{f(y)}{f(x)} \frac{q(x|y)}{q(y|x)}, 1 \right\}$ .
```

The distribution q will be called the *instrumental* (or *proposal*) *distribution.*

This algorithm always accepts values y_t such that the "likelihood ratio" $f(y_t)/q(y_t|x^{(t)})$ is increased compared with the previous value. It is only in the symmetric case that the acceptance is driven by the ratio $f(y_t)/f(x^{(t)})$. An important feature of the algorithm $[A.24]$ is that it may also accept values y_t such that the ratio is decreasing, similar to stochastic optimization methods (see §5.2.2). Like the accept-reject method, the Metropolis–Hastings algorithm only depends on the ratios $f(y_t)/f(x^{(t)})$ and $q(x^{(t)}|y_t)/q(y_t|x^{(t)})$ and is therefore independent of normalizing constants.

Obviously, the probability $\rho(x^{(t)}, y_t)$ is only defined when $f(x^{(t)}) > 0$. However, if the chain starts with a value $x^{(0)}$ such that $f(x^{(0)}) > 0$, it follows that $f(x^{(t)}) > 0$ for every $t \in \mathbb{N}$ since the values of y_t such that $f(y_t) = 0$ lead to $\rho(x^{(t)}, y_t) = 0$, and are therefore rejected by the algorithm.

There are similarities between $[A.24]$ and the accept-reject methods of §2.2.2 and it is possible to use the algorithm $[A.24]$ as an alternative to an accept-reject algorithm, when given a pair (f, g). Both approaches are compared in §6.3.1. Nonetheless, a sample produced by $[A.24]$ obviously differs from an iid sample. For one thing, it involves repeated occurrences of the same value, since to reject Y_t leads to repeat $x^{(t)}$ at time $t+1$, while this is an impossible feature in continuous iid settings, The y_t's generated by the algorithm $[A.24]$ are thus associated with weights of the form m_t/T ($m_t = 0, 1, \cdots$) in the average (6.1.1), depending on how many times the subsequent values have been rejected, and the comparison with importance sampling is somehow more relevant, as discussed in §6.4.

It is obviously necessary to impose minimal conditions on the conditional distribution q for f to be the limiting distribution of the chain $(X^{(t)})$ produced by $[A.24]$. For instance, $\mathcal{E}$, the support of f, is supposed to be *connected.* (This assumption is often omitted in the literature but the lack of connectedness of $\mathcal{E}$ can deeply invalidate the Metropolis–Hastings algorithm. In such a case, it is necessary to proceed on one connected component at a time and show that the different connected components of $\mathcal{E}$ are linked by the kernel of $[A.24]$. See Hobert, Robert and Goutis 1997 for a treatment of the non-connected Gibbs sampler.) If the support of $\mathcal{E}$ is truncated by q, that is, if there exists $A \subset \mathcal{E}$ such that

$$\int_A f(x)dx > 0 \quad \text{and} \quad \int_A q(y|x)dy = 0 \;, \qquad \forall x \in \mathcal{E} \;,$$

the algorithm $[A.24]$ does not have f as a limiting distribution since, for $x^{(0)} \notin A$, the chain $(X^{(t)})$ never visits A. In full generality, we still have the following result:

Theorem 6.2.1 *For every conditional distribution q, whose support includes $\mathcal{E}$, f is a stationary distribution of the chain $(X^{(t)})$ produced by $[A.24]$.*

Proof. The transition kernel associated with $[A.24]$ can be written

$$K(x,x') = \rho(x,x')q(x'|x) + (1-\rho(x,x'))\delta_x(x') \;,$$

where δ_x denotes the Dirac mass in x. Therefore, for every measurable set A,

$$\begin{aligned}
\int K(x,A)f(x)dx &= \int\int \mathbb{I}_A(x')\rho(x,x')q(x'|x)f(x)dxdx' \\
&\quad + \int\int (1-\rho(x,x'))q(x'|x)dx'\, \mathbb{I}_A(x)f(x)dx \\
&= \int\int_D \mathbb{I}_A(x')\, \frac{f(x')}{f(x)}\, \frac{q(x|x')}{q(x'|x)}\; q(x'|x)f(x)dxdx' \\
&\quad + \int\int_{D^c} \mathbb{I}_A(x')\; q(x'|x)\; f(x)dxdx' \\
&\quad + \int\int_D \mathbb{I}_A(x)\; \left(1 - \frac{f(x')}{f(x)}\, \frac{q(x|x')}{q(x'|x)}\right)\; q(x'|x)\; f(x)dxdx'
\end{aligned}$$

for $D = \{(x,x'); f(x')\; q(x|x') \le f(x)\; q(x'|x)\}$. Thus,

$$\begin{aligned}
\int K(x,A)f(x)dx &= \int\int_D \mathbb{I}_A(x')f(x')q(x|x')dxdx' \\
&+ \int\int_D \mathbb{I}_A(x)f(x)q(x'|x)dxdx' \\
&+ \int\int_{D^c} \mathbb{I}_A(x')\; f(x)q(x'|x)dxdx' \\
&- \int\int_D \mathbb{I}_A(x)\; f(x')\; q(x|x')\; dxdx'
\end{aligned}$$

where, in the second and fourth integrals on the right hand side, we make the change of variables (x',x) to (x,x'). This also transforms the set D into

D^c, and vice versa. Therefore, we have

$$\int K(x,A)f(x)dx = \int\int \mathbb{I}_A(x')\, f(x')\, q(x|x')dxdx' = \int_A f(x')dx' \;,$$

establishing the stationarity of f. □

The stationarity of f is therefore established for almost any conditional distribution q, a fact which indicates how universal Metropolis–Hastings algorithms are. As shown by Hastings (1970), Metropolis–Hastings algorithms are only a special case of a more general class of algorithms whose transition is associated with the acceptance probability

$$\varrho(x,y) = \frac{s(x,y)}{1 + \dfrac{f(x)q(y|x)}{f(y)q(x|y)}} \;, \tag{6.2.1}$$

where s is an arbitrary positive symmetric function such that $\varrho(x,y) \le 1$ (see also Winkler 1995). The particular case $s(x,y)=1$ is also known as the *Boltzman algorithm* and used in simulation for particle Physics, although Peskun (1973) has shown that, in the discrete case, the performances of this algorithm are always suboptimal when compared with the Metropolis–Hastings algorithm (see Problem 6.4). Tierney (1995) and Mira, Geyer and Tierney (1998) propose extensions to the continuous case.

6.2.2 Convergence Properties

In order to establish the ergodicity of $(X^{(t)})$ and therefore to validate $[A.24]$ as a Markov Chain Monte Carlo algorithm, we need to prove both the *f-irreducibility* and the *aperiodicity* of $(X^{(t)})$. Tierney (1994) has indeed established the following result:

Theorem 6.2.2 *If the chain $(X^{(t)})$ is f-irreducible, it is positive recurrent. If $(X^{(t)})$ is in addition aperiodic, it is ergodic.*

Proof. Suppose $(X^{(t)})$ is not recurrent, then the space $\mathcal{X}$ can be written as $\cup E_i$ with $P(X^{(t)} \in E_i)$ converging to 0 (see Definition 4.4.5) for every value $x^{(0)}$. This is impossible, given the f-irreducibility of $(X^{(t)})$. The chain $(X^{(t)})$ is therefore positive recurrent, since f is stationary. When $(X^{(t)})$ is aperiodic, ergodicity follows from Theorem 4.3.3. □

Theorem 6.2.3 *If $(X^{(t)})$ is f-irreducible, it is Harris recurrent.*

Proof. This result can be established by using the fact that the only bounded harmonic functions are constant. If h is an harmonic function, it satisfies

$$h(x_0) = \mathbb{E}[h(X^{(1)})|x_0] = \mathbb{E}[h(X^{(t)})|x_0]$$

and therefore h is f-almost everywhere constant and equal to $\mathbb{E}_f[h(X)]$. Since

$$\mathbb{E}[h(X^{(1)})|x_0] = \int \rho(x_0,x_1)\; q(x_1|x_0)\; h(x_1)dx_1 + (1-\overline{\rho}(x_0))\; h(x_0) \;,$$

where $c(b)$ denotes the number of *clusters*, i.e. the number of sites connected by active bonds.

(b) Show that the *Swendson-Wang* algorithm

```
1. Take $b_{st} = 0$ if $x_s \neq x_t$ and, for $x_s = x_t$,
   $b_{st} = \begin{cases} 1 & \text{with probability } 1 - q_{st}, \\ 0 & \text{otherwise;} \end{cases}$
2. For every cluster, choose a color at random on $G$.
```

leads to simulations from π (*Note:* This algorithm is acknowledged as accelerating convergence in image processing.)

6.6 Notes

6.6.1 Geometric Convergence of Metropolis-Hastings algorithms

The sufficient condition (6.2.3) of Theorem 6.2.4 for the irreducibility of the Metropolis–Hastings Markov chain is particularly well adapted to *random walks*, with transition densities $q(y|x) = g(y - x)$. It is indeed enough that g be non empty in a neighborhood of 0 to ensure the ergodicity of $[A.24]$ (see §6.3.2 for a detailed study of these methods). On the other hand, it is quite difficult to obtain a convergence stronger than the simple ergodic convergence of (6.1.1) or than the (total variation) convergence of $\|P^n_{x^{(0)}} - f\|_{TV}$ without introducing additional conditions on f and q. For instance, it is impossible to establish *geometric convergence* without a restriction to the discrete case or without considering particular transition densities, since Roberts and Tweedie (1996) have come up with chains which are not geometrically ergodic. Defining, for every measure ν and every ν-measurable function h, the *essential supremum*

$$\text{ess}_\nu \sup\ h(x) = \inf\ \{w; \nu(h(x) > w) = 0\}\ ,$$

they have indeed established the following result:

Theorem 6.6.1 *If the marginal probability of acceptance satisfies*

$$\text{ess}_f \sup\ (1 - \overline{\rho}(x)) = 1,$$

the algorithm $[A.24]$ *is not geometrically ergodic.*

Therefore, if $\overline{\rho}$ is not bounded from below on a set of measure 1, a geometric speed of convergence cannot be guaranteed for $[A.24]$. This result is important as it characterizes Metropolis–Hastings algorithms which are weakly convergent (see the extreme case of Example 8.2.8); however, it cannot be used as a criterion for geometric convergence, since the function $\overline{\rho}$ is almost always intractable.

In the particular case when $\mathcal{E}$ is a small set (see Chapter 4), Roberts and Polson (1995) note that the chain $(X^{(t)})$ is uniformly ergodic. This condition, however, is rather restrictive besides the $\mathcal{E}$ finite case. It is in fact equivalent to *Doeblin's condition*, as stated by Theorem 4.6.12. Chapter 7 exhibits continuous examples of Gibbs samplers when uniform ergodicity holds (see Examples 7.1.18 and 7.1.6) but §6.3.2 has shown that in the particular case of random walks, geometric convergence is almost never guaranteed, while being a natural choice for the instrumental distribution q.

6.6.2 A Reinterpretation of Simulated Annealing

Consider a function E defined on a finite set $\mathcal{X}$ with such a large cardinal that a minimization of E based on the comparison of the values of $E(\mathcal{X})$ is not feasible. The simulated annealing technique (see §5.2.3) is based on a conditional density q on $\mathcal{X}$ such that $q(i|j) = q(j|i)$ for every $(i, j) \in \mathcal{X}^2$. For a given value $T > 0$, it produces a Markov chain $(X^{(t)})$ on $\mathcal{X}$ by the following transition:

```
1. Generate $\zeta_t$ according to $q(\zeta|x^{(t)})$.
2. Take
```

$$X^{(t+1)} = \begin{cases} \zeta_t & \texttt{with probability} \exp\left(\{E(x^{(t)}) - E(\zeta_t)\}/T\right) \wedge 1, \\ x^{(t)} & \texttt{otherwise.} \end{cases}$$

As noted in §5.2.3, the simulated value ζ_t is automatically accepted when $E(\zeta_t) \leq E(x^{(t)})$. The fact that the simulated annealing algorithm may give a value of $E(X^{(t+1)})$ larger than $E(x^{(t)})$ is a very positive feature of the method, since it allows for escapes from the attraction zone of local minima of E when T is large enough. Comparing [A.24] and the simulated annealing algorithm, the later appears as a particular case of Metropolis–Hastings algorithm, given the symmetry in the conditional distribution. The chain $(X^{(t)})$ is therefore associated with the stationary distribution $f(x) \propto \exp(-E(x)/T)$ provided that the matrix of the $q(i|j)$'s generates an irreducible chain. Note, however, that the theory of homogeneous Markov chains, presented in Chapter 4, does not cover the extension to the case when T varies with t and converges to 0 "slowly enough" (typically in $\log t$).

6.6.3 Reference Acceptance Rates

Gelman, Gilks and Roberts (1996) recommend the use of instrumental distributions such that their *acceptance rate is close to* $1/4$ *for models of high dimension and is equal to* $1/2$ *for the models of dimension* 1 *or* 2. This heuristic rule is based on the asymptotic behavior of an *efficiency criterion* equal to the ratio of the variance of an estimator based on an iid sample and the variance of the estimator (3.1.1), that is

$$\left[1 + 2\sum_{k>0} \operatorname{cov}\left(X^{(t)}, X^{(t+k)}\right)\right]^{-1}$$

in the case $h(x) = x$. In the particular case when f is the density of the $\mathcal{N}(0,1)$ distribution and when g is the density of a Gaussian random walk with variance σ, Gelman *et al.* (1996) have shown that the optimal choice of σ is 2.4, with moreover a dissymmetry in the efficiency in favor of large values of σ. The corresponding acceptance rate is

$$\rho = \frac{2}{\pi} \arctan\left(\frac{2}{\sigma}\right) ,$$

equal to 0.44 for $\sigma = 2.4$. A second result by Gelman et al. (1996), based on an approximation of $x^{(t)}$ by a Langevin diffusion process (see §6.3.5) when the dimension of the problem goes to infinity, is that the acceptance probability converges to 0.234, i.e. approximately 1/4. An equivalent version of this empirical rule is to *take the scale factor in* g *equal to* $2.38/\sqrt{d}\,\Sigma$, where d is the

CHAPTER 7

The Gibbs Sampler

He sat, continuing to look down the nave, when suddenly the solution to the problem just seemed to present itself. It was so simple, so obvious he just started to laugh, the echoes pealing around the deserted church. [...] He remembered the voice of his old 'Dominus' Father Benedict, telling him that there was a solution to every problem. "It's just a matter of perspective, my dear boy," he used to boom out. "Just a matter of perspective."
—P.C. Doherty, *Satan in St Mary's*—

In this place he was found by Gibbs, who had been sent for him in some haste. He got to his feet with promptitude, for he knew no small matter would have brought Gibbs in such a place at all.
—G.K. Chesterton, *The Innocence of Father Brown*—

7.1 General Principles

7.1.1 Definition

The previous chapter has developed simulation techniques which could be called "generic", since they only require a limited amount of information about the distribution to simulate. However, Metropolis–Hastings algorithms achieve higher levels of efficiency when they take into account the specifics of the distribution f, in particular through the calibration of the acceptance rate (see §6.4.1). For instance, as an example of a generic algorithm, ARMS (§6.3.3), aims at reproducing f in an automatic manner. In contrast, the properties and performance of the Gibbs sampling method presented in this chapter are very closely tied to the distribution f. This is because the choice of an instrumental distribution is essentially reduced to a choice between a *finite* number of possibilities.

Although the Gibbs sampler is, formally, a special case of Metropolis–Hastings algorithm (or rather a combination of Metropolis–Hastings algorithms on different components; see Theorem 7.1.10 below), the Gibbs sampling algorithm has a number of distinct features:

(i) The acceptance rate of the Gibbs sampler is uniformly equal to 1. Therefore, every simulated value is accepted and the results of §6.4.1 on the optimal acceptance rates are not valid in this setting. This also means that the convergence assessment for this algorithm must be treated differently than for Metropolis–Hastings techniques.

(ii) The use of the Gibbs sampler implies limitations on the choice of instrumental distributions and requires a prior knowledge of some analytical or probabilistic properties of f.

(iii) The Gibbs sampler is, by construction, multidimensional. Even though some components of the simulated vector may be artificial for the problem of interest, or unnecessary for the required inference, the construction is still at least two-dimensional.

(iv) The Gibbs sampler does not apply to problems where the number of parameters varies, as in §6.3.4, for obvious reasons of lack of irreducibility of the resulting chain.

These different points will be made clearer after a discussion of the properties of the Gibbs sampler. However, we will first define the specific features of this algorithm. Suppose that for some $p > 1$ the random variable $\mathbf{X} \in \mathcal{X}$ can be written as $\mathbf{X} = (X_1, \ldots, X_p)$, where the X_i's are either uni- or multidimensional. Moreover, suppose that we can simulate from the corresponding conditional densities $f_1, \ldots, f_p$, that is

$$X_i|x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p \sim f_i(x_i|x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p).$$

The associated *Gibbs sampling* algorithm (or *Gibbs sampler*) is given by the following transition from $X^{(t)}$ to $X^{(t+1)}$:

Algorithm A.31 –***The Gibbs sampler***– *Given* $\mathbf{x}^{(t)} = (x_1^{(t)}, \ldots, x_p^{(t)})$, *generate*

```
1. $X_1^{(t+1)} \sim f_1(x_1|x_2^{(t)}, \ldots, x_p^{(t)})$;

2. $X_2^{(t+1)} \sim f_2(x_2|x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)})$,        $[A.31]$
      ...

p. $X_p^{(t+1)} \sim f_p(x_p|x_1^{(t+1)}, \ldots, x_{p-1}^{(t+1)})$.
```

The densities $f_1, \ldots, f_p$ are called the *full conditionals*, and it is a particular feature of the Gibbs sampler that these are the densities used for simulation. So, even in a high dimensional problem, all of the simulations may be univariate, which is usually an advantage.

Example 7.1.1 –**Bivariate Gibbs sampler**– Let the random variables X and Y have joint density $f(x, y)$, and generate a sequence of observations according to the following: `Set` $X_0 = x_0$, `and for` $t = 1, 2, \ldots$, `generate`

$$\begin{aligned} Y_t &\sim f_{Y|X}(\cdot|x_{t-1}) \\ X_t &\sim f_{X|Y}(\cdot|y_t) \end{aligned} \tag{7.1.1}$$

where $f_{Y|X}$ and $f_{X|Y}$ are the conditional distributions. The sequence (X_t, Y_t), is a Markov chain, as is each sequence (X_t) and (Y_t) individually. For example, the chain (X_t) chain has transition kernel

$$K(x, x^*) = \int f_{X|Y}(x^*|y) f_{Y|X}(y|x) dy,$$

with invariant distribution $f_X(\cdot)$.

For the special case of the bivariate normal density,

$$(X,Y) \sim \mathcal{N}_2\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) , \tag{7.1.2}$$

the Gibbs sampler is

`Given` y_t`, generate`

$$\begin{aligned} X_{t+1} \mid y_t &\sim \mathcal{N}(\rho y_t,\ 1-\rho^2) \\ Y_{t+1} \mid x_t &\sim \mathcal{N}(\rho x_{t+1},\ 1-\rho^2). \end{aligned} \tag{7.1.3}$$

Obviously, this is a formal example since the bivariate normal density can be directly simulated by the Box-Muller algorithm (see Example 2.2.2). ||

Example 7.1.2 –**Auto-exponential model**– The *auto-exponential model* of Besag (1974) has been found useful in some aspects of spatial modelling. For the particular case when $y \in \mathbb{R}^3_+$, the corresponding density is

$$f(y_1,y_2,y_3) \propto \exp\{-(y_1+y_2+y_3+\theta_{12}y_1y_2+\theta_{23}y_2y_3+\theta_{31}y_3y_1)\} ,$$

with known $\theta_{ij} > 0$. The full conditional densities are exponential,

$$y_i|y_{j\neq i} \sim \mathcal{E}xp\left(1+\sum\nolimits_{j\neq i} \theta_{ij}y_j\right) ,$$

and so are very easy to simulate from. In contrast, the other conditionals, and the marginal distributions have forms such as

$$\begin{aligned} y_2|y_1 &\sim \int_0^{+\infty} f(y_1,y_2,y_3)\, dy_3 \\ &\qquad \propto \frac{\exp\{-(y_1+y_2+\theta_{12}y_{12})\}}{1+\theta_{23}y_2+\theta_{12}y_1} , \\ y_1 &\sim \pi(y_1) \propto \int_0^{+\infty} \frac{\exp\{-y_2-\theta_{12}y_1y_2\}}{1+\theta_{23}y_2+\theta_{12}y_1}\, dy_2\, e^{-y_1}, \end{aligned}$$

which cannot be simulated easily. ||

Example 7.1.3 –**Ising model**– For the *Ising model* of Example 5.2.5, where

$$f(s) \propto \exp\left\{-H\sum\nolimits_i s_i - J\sum\nolimits_{(i,j)\in\mathcal{N}} s_is_j\right\}, \qquad s_i \in \{-1,1\},$$

and where $\mathcal{N}$ denotes the neighborhood relation for the network, the full conditional distribution is

$$\begin{aligned} f(s_i|s_{j\neq i}) &= \frac{\exp\{-Hs_i - Js_i\sum_{j:(i,j)\in\mathcal{N}} s_j\}}{\exp\{-H-J\sum_j s_j\}+\exp\{H+J\sum_j s_j\}} \\ &= \frac{\exp\{-(H+\sum_j s_j)(s_i+1)\}}{1+\exp\{-2(H+\sum_j s_j)\}} , \end{aligned}$$

which is a logistic distribution on $(s_i + 1)/2$. It is therefore particularly easy to implement $[A.32]$ for these conditional distributions by updating successively each node of the network. ||

7.1.2 Completion

Following the mixture method proposed in §2.2.1, it is easy to generalize the Gibbs sampling algorithm by a "demarginalization" or *completion* construction.

Definition 7.1.4 Given a probability density f, a density g that satisfies

$$\int_Z g(x,z)\,dz = f(x)$$

is called a *completion* of f.

The density g is chosen so that the full conditionals of g are easy to simulate from, and the following Gibbs algorithm is implemented. For $p > 1$, write $y = (x, z)$ and denote the conditional densities of $g(y) = g(y_1, \ldots, y_p)$ by

$$\begin{array}{rcl} Y_1|y_2,\ldots,y_p & \sim & g_1(y_1|y_2,\ldots,y_p), \\ Y_2|y_1,y_3,\ldots,y_p & \sim & g_2(y_2|y_1,y_3,\ldots,y_p), \\ & & \ldots, \\ Y_p|y_1,\ldots,y_{p-1} & \sim & g_p(y_p|y_1,\ldots,y_{p-1}). \end{array}$$

$Y^{(t)}$ to $Y^{(t+1)}$.

Algorithm A.32 –*Completion Gibbs sampler*–

```
Given $(y_1^{(t)},\ldots,y_p^{(t)})$, simulate
1. $Y_1^{(t+1)} \sim g_1(y_1|y_2^{(t)},\ldots,y_p^{(t)})$,
2. $Y_2^{(t+1)} \sim g_2(y_2|y_1^{(t+1)},y_3^{(t)},\ldots,y_p^{(t)})$,        $[A.32]$
   ...
p. $Y_p^{(t+1)} \sim g_p(y_p|y_1^{(t+1)},\ldots,y_{p-1}^{(t+1)})$.
```

We need a simple example here–Dave Winfield? No baseball, please!!!

Example 7.1.5 –**Truncated Normal Distribution**– As in Example 2.2.12, consider a truncated normal distribution,

$$f(x) \propto e^{-(x-\mu)^2/2\sigma^2}\,\mathbb{I}_{x \geq \underline{\mu}}\ .$$

As mentioned in Example 2.2.12, the naive simulation of a normal $\mathcal{N}(0,1)$ till the outcome is above $\underline{\mu}$ is suboptimal for large values of $\underline{\mu}$. However, the devising and the optimization of the accept-reject algorithm constructed in Example 2.2.12 can be overly costly if the algorithm is only to be used a few times. An alternative is to use completion, as, for instance,

$$g(x,z) \propto \mathbb{I}_{x \geq \underline{\mu}}\,\mathbb{I}_{z \leq \exp\{-(x-\mu)^2/2\sigma^2\}}\ .$$

(This is a special case of *slice sampling*, see [A.33].) The corresponding implementation of [A.32] is then

```
Simulate
1. $X^{(t)}|z^{(t-1)} \sim \mathcal{U}([\underline{\mu}, \sqrt{-2\sigma^2 \log(z^{(t-1)})}])$,
2. $Z^{(t)}|x^{(t)} \sim \mathcal{U}([0, \exp\{-(x^{(t)} - \mu)^2/2\sigma^2\}])$.
```

Note that the initial value of z must be chosen so that $\sqrt{-2\sigma^2 \log(z^{(0)})} > \underline{\mu}$. $\|$

The *completion* of a density f into a density g such that f is the marginal density of g corresponds to one of the first (historical) appearances of the Gibbs sampling in Statistics, namely the introduction of *data augmentation* by Tanner and Wong (1987) (see Note §7.6.1).

In cases when such completions seem *necessary* (for instance, when every conditional distribution associated with f is not explicit), there indeed exists a wide choice among the infinite number of densities for which f is a marginal density. We will not discuss this choice in terms of *optimality*, first because there is no more results on this topic than on the choice of an optimal density g in the Metropolis–Hastings algorithm and, second, because there exists, in general, a *natural* completion of f in g and of x in y. *Missing data models*, treated in Chapter 9, provide a series of examples of natural completion (see also §5.3.1.)

Note the similarity of this approach with the EM algorithm for maximizing a missing data likelihood, which is described in §5.3.3, and even more with recent versions of EM such as ECM and MCEM (see Meng and Rubin 1991, 1992, or Liu and Rubin, 1994), which use maximizations of conditional parts of the likelihood like $g_1, g_2, \ldots, g_p$ to converge faster to the maximum of the likelihood function.

In principle, the Gibbs sampler by no means requires that the completion of f into g and of x in $y = (x, z)$ should be related to the problem of interest. There are therefore settings such that the vector z has no meaning from a statistical point of view and is only a useful device.

Example 7.1.6 –**Cauchy-normal posterior distribution**– As shown in Chapter 2 (§2.2.1), Student's t distribution can be generated as a mixture of a normal distribution by a χ^2 distribution. This kind of decomposition is also useful when the expression $[1+(\theta-\theta_0)^2]^{-\nu}$ appears in a more complex distribution. Consider for instance the density

$$f(\theta|\theta_0) \propto \frac{e^{-\theta^2/2}}{[1+(\theta-\theta_0)^2]^\nu}.$$

This is the posterior distribution of the location parameter in a Cauchy distribution (see Example 8.3.2) which also appears in the estimation of the parameter of interest in a linear calibration model (see Example 1.5.1).

The density $f(\theta|\theta_0)$ can be written as the marginal density

$$f(\theta|\theta_0) \propto \int_0^\infty e^{-\theta^2/2}\, e^{-[1+(\theta-\theta_0)^2]\,\eta/2}\, \eta^{\nu-1}\, d\eta,$$

7.53 †(Breslow and Clayton 1993) In the modelling of breast cancer cases y_i according to age x_i and year of birth d_i, an exchangeable solution is

$$\begin{aligned} Y_i &\sim \mathcal{P}(\mu_i), \\ \log(\mu_i) &= \log(d_i) + \alpha_{x_i} + \beta_{d_i}, \\ \beta_k &\sim \mathcal{N}(0, \sigma^2). \end{aligned}$$

(a) Derive the Gibbs sampler associated with almost flat hyperpriors on the parameters $\alpha_j, \beta_k, \sigma$.

(b) Breslow and Clayton (1993) consider a dependent alternative where ($k = 3, \ldots, 11$)

$$\beta_k | \beta_1, \ldots, \beta_{k-1} \sim \mathcal{N}(2\beta_{k-1} - \beta_{k-2}, \sigma^2), \tag{7.5.4}$$

while $\beta_1, \beta_2 \sim \mathcal{N}(0, 10^5 \sigma^2)$. Construct the associated Gibbs sampler and compare with the previous results.

(c) An alternative representation of (7.5.4) is

$$\beta_k | \beta_j, j \neq k \sim \mathcal{N}(\bar{\beta}_k, n_k \sigma^2).$$

Determine the value of $\bar{\beta}_k$ and compare the associated Gibbs sampler with the previous implementation.

7.54 †(Dobson 1983) The effect of a pesticide is tested against its concentration x_i on n_i beetles, $R_i \sim \mathcal{B}(n_i, p_i)$ of which are killed. Three generalized linear models are in competition:

$$\begin{aligned} p_i &= \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}, \\ p_i &= \Phi(\exp(\alpha + \beta x_i)), \\ p_i &= 1 - \exp(-\exp(\alpha + \beta x_i)), \end{aligned}$$

that is, the logit, probit and log-log models. For each of these models, construct a Gibbs sampler and compute the expected posterior *deviance*, that is, the posterior expectation of

$$D = 2 \left(\sum_{i=1}^{n} \hat{\ell}_i - \sum_{i=1}^{n} \ell_i \right),$$

where

$$\ell_i = r_i \log(p_i) + (n_i - p_i) \log(1 - p_i), \qquad \hat{\ell}_i = \max_{p_i} \ell_i.$$

7.55 †(Spiegelhalter et al. 1996) Consider a standard Anova model ($i = 1, \ldots, 4, j = 1, \ldots, 5$)

$$\begin{aligned} Y_{ij} &\sim \mathcal{N}(\mu_{ij}, \sigma^2), \\ \mu_{ij} &= \alpha_i + \beta_j, \\ \alpha_i &\sim \mathcal{N}(0, \sigma_\alpha^2), \\ \beta_j &\sim \mathcal{N}(0, \sigma_\beta^2), \\ \sigma^{-2} &\sim \mathcal{G}a(a, b), \end{aligned}$$

with $\sigma_\alpha^2 = \sigma_\beta^2 = 5$ and $a = 0$, $b = 1$. Gelfand et al. (1992) impose the constraints $\alpha_1 > \ldots > \alpha_4$ and $\beta_1 < \ldots < \beta_3 > \ldots > \beta_5$.

(a) Give the Gibbs sampler for this model. (*Hint:* Use the optimal truncated normal accept-reject algorithm of Example 2.2.12.)

7.6.2 Geometric Convergence

While[11] the geometric ergodicity conditions of Chapter 6 do not apply for Gibbs sampling algorithms, some sufficient conditions can also be found in the literature, although these are no so easy to implement in practice. The approach presented below is based on results by Schervish and Carlin (1992), Liu, Wong and Kong (1995), and Polson (1996) who work with a *functional representation* of the transition operator of the chain.

Schervish and Carlin (1992) define the measure μ as the measure with density $1/g$ with respect to the Lebesgue measure, where g is the density of the stationary distribution of the chain. They then define a scalar product on $\mathcal{L}_2(\mu)$

$$< r, s >= \int r(y)\ s(y)\ \mu(dy)$$

and define the operator T on $\mathcal{L}_2(\mu)$ by

$$(Tr)(y) = \int r(y')\ K(y', y)\ g(y')\ \mu(dy'),$$

where $K(y, y')$ is the transition kernel (7.1.6) with stationary measure g. In this setting, g is an *eigenvector associated with the eigenvalue* 1. The other eigenvalues of T are characterized by the following result:

Lemma 7.6.1 *The eigenvalues of T are all within the unit disk of $\mathbb{C}$.*

Proof. Consider an eigenvector r associated with the eigenvalue c, i.e. such that $Tr = cr$. Since

$$\int K(y', y)\ g(y')\ dy' = g(y),$$

r satisfies

$$\begin{aligned} |c| \int |r(y)|\ dy &= \int |Tr(y)|\ dy \\ &= \int \left| \int r(y')\ K(y', y)\ dy' \right| dy \\ &\leq \int \int |r(y')|\ K(y', y)\ dy'\ dy \\ &= \int |r(y')|\ dy' \end{aligned}$$

and therefore $|c| \leq 1$. □

The main requirement in Schervish and Carlin (1992) is the *Hilbert-Schmidt condition*

$$\int \int K^2(y, y')\ dy\ dy < \infty, \tag{7.6.1}$$

which ensures both the compacity of T and the geometric convergence of the chain $(Y^{(t)})$. The *adjoint operator* associated with T, T^*, is defined by

$$< Tr, s >=< r, T^* s >$$

[11] This section presents some results on the analysis of Gibbs sampling algorithms from a *functional analysis* point of view. It may be skipped on a first reading since it will not be used in the book and remains at a rather theoretical level.

a review of finer convergence properties. We still conclude with the warning that the practical consequences of this theoretical evaluation seem negligible. In fact, setups where eigenvalues of these operators are available almost always correspond to case where Gibbs sampling is not necessary (see, e.g., Example 7.6.4). The Hilbert-Schmidt condition (7.6.1) is moreover particularly difficult to check when $K(y, y')$ is not explicit. At last, we also consider that the eigenvalues of the operators T and F and the convergence of norms $\|g_t - g\|_{TV}$ to 0 are only marginally relevant in the study of the convergence of the average

$$\frac{1}{T} \sum_{t=1}^{T} h(y_t) \tag{7.6.6}$$

to $\mathbb{E}[h(Y)]$, even though there exists a theoretical connection between the asymptotic variance γ_h^2 of (7.6.6) and the spectrum (λ_k) of F through

$$\gamma_h^2 = \sum_{k \geq 2} w_k \frac{1+\lambda_k}{1-\lambda_k} \leq \frac{1+\lambda_2}{1-\lambda_2}$$

(see Besag and Green 1992 and Geyer 1992), where the weight w_k depend on h and F and the λ_k are the (decreasing) eigenvalues of F (with $\lambda_1 = 1$).

7.6.3 *The* BUGS *software*

The acronym BUGS stands for *Bayesian inference using Gibbs sampling.* This software has been elaborated by Spiegelhalter, Thomas, Best and Gilks (1995a,b,c) at the MRC Biostatistics Unit in Cambridge, England. As shown by its name, it has been designed to take advantage of the possibilities of the Gibbs sampler in Bayesian analysis. BUGS includes a language which is C or S-plus like and involves declarations about the model, the data and the prior specifications, for single or multiple levels in the prior modelling. For instance, for the benchmark nuclear pump failures dataset of Example 7.1.18, the model and priors are defined by

```
for (i in 1:N) {
    theta[i] ~ dgamma(alpha,beta)
    lambda[i]~ theta[i] * t[i]
    x[i]     ~ dpois(lambda[i])
}
alpha ~ dexp(1.0)
beta  ~ dgamma(0.1,1.0)
```

(see Spiegelhalter et al. 1995b, p.9). Most standard distributions are recognized by BUGS (21 are listed in Spiegelhalter et al. 1995a), which also allows for a large range of transforms. BUGS also recognizes a series of commands like `compile`, `data`, `out`, `stat`. The output of BUGS is a table of the simulated values of the parameters after an open number of warmup iterations, the batch size being also open.

A major restriction of this software is the use of the conjugate priors or at least log-concave distributions for the Gibbs sampler to apply. However, more complex distributions can be handled by discretization of their support and assessment of the sensitivity to the discretization step. In addition, improper priors are not accepted and must be replaced by proper priors with small precision, like `dnorm(0,0.0001)`, which represents a normal modelling with mean 0 and *precision* (inverse variance) 0.0001.

The BUGS manual (Spiegelhalter et al. 1995a) is quite informative and well-written.[14] In addition, the authors have written an extended and most helpful example manual (Spiegelhalter et al. 1995b, c), which exhibits the ability of BUGS to deal with an amazing number of models, including meta-analysis, latent variable, survival analysis, non-parametric smoothing, model selection and geometric modelling. (Some of these models are presented in Problems 7.45–7.56.) The BUGS software is also compatible with the convergence diagnosis software CODA presented in Note §8.6.4.

[14] At the present time, that is Spring 1998!, the BUGS software is available as a freeware on the Web site http://www.mrc_bsu.cam.ac.uk/bugs for a wide variety of platforms.

CHAPTER 8

Diagnosing Convergence

"Everything that is not what it seems and not what it reasonably should be", said Cadfael firmly, "must have significance. And until I know what that significance is, I cannot be content."

—Ellis Peter, *The Heretic's Apprentice*—

Leaphorn never counted on luck. Instead, he expected order–the natural sequence of behavior, the cause producing the natural effect, the human behaving in the way it was natural for him to behave. He counted on that and on his own ability to sort out the chaos of observed facts and find in them this natural order.

—Tony Hillerman, *The Blessing Way*—

8.1 When and why to stop?

The two previous chapters have laid the theoretical foundations of MCMC algorithms and showed that, under fairly general conditions, the chains produced by these algorithms are ergodic, or even geometrically ergodic. While such developments are obviously necessary, they are nonetheless insufficient from the point of view of the implementation of MCMC methods, since they do not directly induce methods to control *the* chain produced by an algorithm (in the sense of a *stopping rule* which guarantees that the number of iterations is sufficient). In other words, the general convergence results do not tell us when to stop the MCMC algorithm. The goal of this chapter is therefore to present, sometimes in an allusive mode, the numerous methods of control proposed in the literature, as in the reviews of Cowles and Carlin (1994) and Brooks and Roberts (1995), as well as Robert (1995) (The style of this chapter reflects the more recent and exploratory nature of these methods, by describing a sequence of non comparable techniques with widely varying degrees of theoretical justification.)

From a general point of view, there are three (increasingly stringent) types of convergence for which control is necessary:

(i) convergence of the chain $\theta^{(t)}$ to the stationary distribution f (or *stationarization*);

(ii) convergence of the empirical average

$$\frac{1}{T}\sum_{t=1}^{T} h(\theta^{(t)}) \tag{8.1.1}$$

to $\mathbb{E}_f[h(\theta)]$ for an arbitrary function h;

(iii) convergence of a sample $(\theta_1^{(t)}, \ldots, \theta_n^{(t)})$ to iid-ness.

First, (i) appears as a minimum requirement on a algorithm supposed to approximate simulation from f! For instance, the original implementation of the Gibbs sampler was based on the generation of n independent initial values $\theta_i^{(0)}$ $(i = 1, \cdots, n)$, and the storage of the last simulation $\theta_i^{(T)}$ in each chain. Strictly speaking, this approach thus requires a corresponding stopping rule for the correct determination of T (see for instance Tanner and Wong 1987). (Note that, in practice, this method induces a waste of resources, since it implies discarding most of the generated variables with little justification with regards to points (i) and (ii).)

On a general basis, it seems to us that this approach (i) to convergence issues is not particularly fruitful. In fact, from a theoretical point of view, f is only the *limiting* distribution of $\theta^{(t)}$; stationarity is therefore only achieved asymptotically and the $\theta_i^{(T)}$'s are distributed from $f_{\mu_0}^T$, if μ_0 is the (initial) distribution of $\theta^{(0)}$. If, on the opposite, we only consider a single realization (or *path*) of the chain $(\theta^{(t)})$, the question of *convergence* to the limiting distribution is not really relevant. Indeed, it is often possible[1] to consider the initial value $\theta^{(0)}$ as distributed from the distribution f, therefore to act as if the chain is already in its stationary regime from the start. (In cases where this assumption is unrealistic, there are methods like the exact simulation of Propp and Wilson (1996), discussed in §8.6.5, where stationarity can be rigorously achieved from the start.) This way of evacuating the first type of control may appear rather cavalier, but we do think that convergence to f *per se* is not the major issue for most MCMC algorithms, in the sense that the chain truly produced by the algorithm often behaves like a chain initialized from f. More precisely, slow exploration of the support of f and strong correlations between the $\theta^{(t)}$'s are rather the issues at stake. This is not to say that stationarity should not be tested at all, as we will see in §8.2.2, since, notwithstanding the starting distribution, μ_0, the chain may be slow to explore the different regions of the support of f, with lengthy stays in each of these regions (e.g. the modes of the distribution f), and a stationarity test may be instrumental in detecting such difficulties.

The second type (ii) of convergence is deemed to be the most relevant in the implementation of MCMC algorithms. Indeed, even when $\theta^{(0)} \sim f$, the exploration of the complexity of f by the chain $(\theta^{(t)})$ can be more or less lengthy, depending on the transition kernel chosen for the algorithm.

[1] We consider a standard statistical setup where the support of f is approximately known. This may not be the case for high dimensional setups or complex structures where the algorithm is initialized at random.

The purpose of the control is therefore to determine whether the chain has indeed exhibited all the facets of f (for instance, all the modes). Brooks and Roberts (1995) relate this convergence to the *mixing* speed of the chain, in the informal sense of a strong (or weak) dependence on initial conditions and of a slow (or fast) exploration of the support of f (see also Asmussen et al. 1992). A formal version of convergence control in this setup is the convergence assessment of §8.1.1. While the ergodic theorem guarantees the convergence of this average from a theoretical point of view, the relevant issue at this stage is to determine a minimal value for T which validates the approximation of $\mathbb{E}_f[h(\theta)]$ by (8.1.1).

The third type of convergence (iii) takes into account *independence* requirements for the simulated values. Rather than approximating integrals like $\mathbb{E}_f[h(\theta)]$, the goal is to produce variables θ_i which are (quasi-)independent. While the solution based on parallel chains mentioned above is not satisfactory, an alternative is to use *sub-sampling* (or *batch sampling*) to reduce correlation between the successive points of the Markov chain. This technique, which is customarily used in numerical simulation (see for instance Schmeiser 1989), subsamples the chain $(\theta^{(t)})$ with a batch size k, considering only the values $\eta^{(t)} = \theta^{(kt)}$. When the covariance $\text{cov}_f(\theta^{(0)}, \theta^{(t)})$ is decreasing monotonically with t (see §7.2.3), the motivation for subsampling is obvious. In particular, if the chain $(\theta^{(t)})$ satisfies an interleaving property (see §7.2.2), subsampling is justified. However, checking for the monotone decrease of $\text{cov}_f(\theta^{(0)}, \theta^{(t)})$—which also justifies Rao-Blackwellization (see §7.2.3)—is not always possible and, in some settings, the above covariance oscillates with t, which complicates the choice of k. Section §8.4.1 describes how Raftery and Lewis (1992a,b) estimate this batch size k. Note at this stage that sub-sampling necessarily leads to losses in efficiency with regards to the second convergence goal. In fact, as shown by MacEachern and Berliner (1994), it is always preferable to use the whole sample for the approximation of $\mathbb{E}_f[h(\theta)]$. Nonetheless, for convergence assessment, subsampling may be beneficial (see, e.g., Robert, Rydén and Titterington 1998).

Lemma 8.1.1 – *Consider $h \in \mathcal{L}^2(f)$. For every $k > 1$, if $(\theta^{(t)})$ is a Markov chain with stationary distribution f and if*

$$\delta_1 = \frac{1}{Tk} \sum_{t=1}^{Tk} h(\theta^{(t)}) \ , \quad \delta_k = \frac{1}{T} \sum_{\ell=1}^{T} h(\theta^{(k\ell)}),$$

the variance of δ_1 satisfies

$$\text{var}(\delta_1) \le \text{var}(\delta_k).$$

Proof. Define $\delta_k^1, \cdots, \delta_k^{k-1}$ as the drifted versions of $\delta_k = \delta_k^0$, that is

$$\delta_k^i = \frac{1}{T} \sum_{t=1}^{T} h(\theta^{(tk-i)}), \qquad i = 0, 1, \cdots, k-1 \ .$$

δ_1 can then be written under the form

$$\delta_1 = \frac{1}{k} \sum_{i=0}^{k-1} \delta_k^i \ .$$

Therefore

$$\begin{aligned} \operatorname{var}(\delta_1) &= \operatorname{var}\left(\frac{1}{k} \sum_{i=0}^{k-1} \delta_k^i \right) \\ &= \operatorname{var}(\delta_k^i)/k + \sum_{i \neq j} \operatorname{cov}(\delta_k^i, \delta_k^j)/k^2 \\ &\leq \operatorname{var}(\delta_k^1)/k + \sum_{i \neq j} \operatorname{var}(\delta_k^1)/k^2 \\ &= \operatorname{var}(\delta_k) \ , \end{aligned}$$

from the Cauchy–Schwarz inequality. □

In the remainder of the chapter, we consider independence issues only in cases where they have bearing on the control of the chain, as in renewal theory (see §8.2.3).

Besides distinguishing between convergence to stationarity (§8.2) and convergence of the average (§8.3), we also distinguish between the methods involving the simulation in parallel of M independent chains $(\theta_m^{(t)})$ $(1 \leq m \leq M)$ and those based on a single 'on-line' chain. The motivation of the former is intuitively sound: by simulating several chains, variability and dependence on the initial values are reduced and it should be easier to control convergence to the stationary distribution by comparing the estimations of quantities of interest on the different chains. The dangers of a naive implementation of this principle should be obvious, namely (a) that the slower chain governs convergence and (b) that the choice of the initial distribution is quintessential to guarantee that the different chains are well-dispersed. The elaborate developments of Gelman and Rubin (1992), Liu, Liu and Rubin (1995) and Johnson (1996) can be similarly criticized, even though they seem to propose a evaluation of convergence which is more robust that for single chain methods. For instance, Geyer (1992) points out that this robustness is illusory from several points of view. In fact, good performances of these parallel methods require a sufficient *a priori* knowledge on the distribution f in order to construct an initial distribution on $\theta_m^{(0)}$ which takes into account the specificities of f (modes, shape of high density regions, etc.). In other words, an initial distribution which is too concentrated around a local mode of f does not contribute significantly more than a single chain to the exploration of the specificities of f. Moreover, slow algorithms, like Gibbs sampling used in highly non linear setups, usually favor single chains, in the sense that a unique chain with MT observations and a slow rate of mixing is more likely to get closer to the stationary distribution than M chains of size T, which will presumably stay in the neighborhood of the starting point with higher probability. An additional practical drawback of parallel methods is that they require a modification of the original MCMC algorithm to deal with the processing of parallel outputs. (See Tierney 1994 and Raftery and Lewis 1996 for other criticisms.)On the other hand, single chain methods suffer more severely from the defect that "it only sees where it went", in the sense that the part of the support of f which has not been visited by the chain at time T is almost impossible to detect. Moreover, a single chain may present probabilistic pathologies which are more often avoided by parallel chains.

To conclude, let us agree with many authors that it is somehow illusory to aim at controlling the flow of a Markov chain and assessing its convergence behavior from a single realization of this chain. There always are settings (i.e. transition kernels) which, for most realizations, invalidate an arbitrary indicator, whatever its theoretical validation, and the randomness inherent to the nature of the problem prevents any categorical guarantee of performance. The crux of the difficulty is actually similar to Statistics, where the uncertainty due to the observations prohibits categorical conclusions and final statements. Far from being a failure acknowledgment, these remarks only aim at warning the reader[2] about the *relative* value of the indicators developed below. As noted by Cowles and Carlin (1994), it is simply inconceivable in the light of recent results to envision *automated stopping rules.* Brooks and Roberts (1995) also stress that the prevalence of a given control method strongly depends on the model and on the inferential problem under study. It is therefore even more crucial to develop robust and general evaluation methods which extend and complement the present battery of stopping criteria, the goal being nowadays to develop "convergence control spreadsheets", in the sense of computer graphical outputs which could present through several graphs different features of the convergence properties of the chain under study (see Cowles and Carlin 1996, Best, Cowles and Vines 1996, or Robert 1997a for illustrations). The criticisms presented in the wake of the techniques proposed below only highlight the incomplete aspect of each method and therefore do not aim at preventing their utilization but rather to warn against a selective interpretation of their results.

8.2 Monitoring Convergence to the Stationary Distribution

8.2.1 Graphical Methods

A natural empirical approach to convergence control is to draw pictures of the output of simulated chains, in order to detect deviant or non-stationary behaviors. For instance, as in Gelfand and Smith (1990), a first plot is to draw the sequence of the $\theta^{(t)}$'s against t. However, this plot is only useful

[2] To borrow from the injunction of Hastings (1970), "*even the simplest of numerical methods may yield spurious results if insufficient care is taken in their use (...) The setting is certainly no better for the Markov chain methods and they should be used with appropriate caution.*"

8.19 **(Problem 8.18 continued)** Define

$$I_t = \frac{1}{m(m-1)} \sum_{i \neq j} \chi_{ij}^t, \quad J_t = \frac{1}{m} \sum_{i=1}^m \chi_{ii}^t,$$

with

$$\chi_{ij}^t = \frac{K(\theta_i^{(0)}, \theta_j^{(2t-1)})}{\pi(\theta_j^{(2t-1)})},$$

based on m parallel chains $(\theta_j^{(0)})$ $(j = 1, \cdots, m)$.

(a) Show that, if $\theta_j^{(0)} \sim \pi$, $\mathbb{E}[I_t] = \mathbb{E}[J_t] = 1$, and that for every initial distribution on the $\theta_j^{(0)}$'s, $\mathbb{E}[I_t] \leq \mathbb{E}[J_t]$.

(b) Show that

$$\text{var}(I_t) \leq \text{var}(J_t)/(m-1)$$

8.20 (Raftery and Lewis 1992a) Deduce (8.4.3) from the normal approximation of δ_T. (*Hint:* Show that

$$\Phi\left(\sqrt{T}\frac{(\alpha+\beta)^{3/2}q}{\sqrt{\alpha\beta(2-\alpha-\beta)}}\right) \geq \frac{\varepsilon'+1}{2}\,.\Bigg)$$

8.21 (Raftery and Lewis 1996) Consider the logit model

$$\begin{aligned} \log\frac{\pi_i}{1-\pi_i} &= \eta + \delta_i \\ \delta_i &\sim \mathcal{N}(0, \sigma^2) \\ \sigma^{-2} &\sim \mathcal{G}a(0.5, 0.2). \end{aligned}$$

Study the convergence of the associated Gibbs sampler and the dependence on the starting value.

8.22 (Dellaportas 1994) Show that

$$\mathbb{E}_g\left[\min\left(1, \frac{f(x)}{g(x)}\right)\right] = \int |f(x) - g(x)| dx$$

Derive an estimator of the L_1 distance between the stationary distribution π and the distribution at time t, π^t.

8.23 (Tanner 1996) Show that, if $\theta^{(t)} \sim \pi^t$ and if the stationary distribution is the posterior density associated with $f(x|\theta)$ and $\pi(\theta)$, the weight

$$\omega_t = \frac{f(x|\theta^{(t)})\pi(\theta^{(t)})}{\pi^t(\theta^{(t)})}$$

converges to the marginal $m(x)$. Derive the *Gibbs stopper* of §8.6.1 by proposing an estimate of π^t.

8.24 For the witch hat distribution of Example 8.2.1, find a set of parameters (δ, σ, y) for which the mode at y takes many iterations to be detected. Apply the various convergence controls to this case.

8.25 Propose an estimator of the variance of S_T^P (when it exists) and derive a convergence diagnostic.

8.26 Check whether the importance sampling S_T^P is (a) available and (b) with finite variance for the Examples of Chapter 7.

8.27 For the model of Example 8.3.8, show that a small set is available in the $(\mu|\eta)$ space and derive the corresponding renewal probability.

8.28 Show that the sequence $(\xi^{(n)})$ defined in §8.4.2 is a Markov chain. (*Hint:* Show that

$$\begin{aligned} P &\quad (\xi^{(n)} = i|\xi^{(n-1)} = j, \xi^{(n-2)} = \ell, \ldots) \\ = &\quad \mathbb{E}_{\theta^{(0)}} \left[\mathbb{I}_{A_i} \left(\theta^{(\tau_{n-1}-1+\Delta_n)}\right) \middle| \theta^{(\tau_{n-1}-1)} \in A_j, \theta^{(\tau_{n-2}-1)} \in A_\ell, \ldots \right] \ , \end{aligned}$$

and apply the strong Markov property.)

8.6 Notes

8.6.1 Other Stopping Rules

Following Tanner and Wong (1987), who first approximated the distribution of $\theta^{(t)}$ to derive a convergence assessment, Ritter and Tanner (1992) propose a control method based on the distribution of the weight ω_t, which evaluate the difference between the limiting distribution f and an approximation of f at iteration t, $\hat{f}_t$,

$$w_t = \frac{f(\theta^{(t)})}{\hat{f}_t(\theta^{(t)})} \ .$$

Note that this evaluation aims more at controlling the convergence in distribution of $(\theta^{(t)})$ to f (first type of control) than at measuring the mixing speed of the chain (second type of control).

In some settings, the approximation $\hat{f}_t$ can be computed by Rao-Blackwellization as in (8.3.2) but Ritter and Tanner (1992) develop a general approach (called *Gibbs stopper*) where, if $K(\theta, \theta')$ denotes the transition kernel, the approximation of

$$f_t(\theta) = \int K(\theta', \theta) \ f_{t-1}(\theta') \, d\theta'$$

is

$$\hat{f}_t(\theta) = \frac{1}{m} \sum_{j=1}^{m} K(\theta^{(t-j)}, \theta) \ .$$

Brooks and Roberts (1995) cover the extension to Metropolis–Hastings algorithms (see also Zellner and Min 1995, for a similar approach).

The theoretical foundations of this approximation are limited since the $\theta^{(t-j)}$'s are *not* distributed from f_{t-1}. A different implementation can be based on parallel chains $\theta_j^{(t)}$, although it does not enjoy a stronger theoretical validity (see §8.4). An additional difficulty is related to the computation of K and of its normalizing constant, which can induce a considerable increase of the computation time.

Ritter and Tanner (1992) propose to use the weight w_t through a stopping rule based on the evolution of the histogram of the w_t's, until these weights are sufficiently concentrated near a constant which is the normalizing constant of f. However, the example treated in Tanner (1996) does not indicate how the quantitative assessment of concentration is operated, since it used characteristics of the distribution f which are usually unknown to calibrate this convergence of the weights w_t. Cowles and Carlin (1994) note moreover that the criterion is sensitive to the choice of m in $\hat{f}_t$, in the sense that large values of m lead to histograms which are much stabler and therefore indicate a seemingly faster convergence. Brooks and Roberts (1995) repeat these criticisms on the difficult interpretation of histograms and the influence of the size of windows. They propose to use the criterion differently via the empirical variance of the weights ω_t on parallel chains till it converges to 0.

8.6.2 Spectral Analysis

As already mentioned in Hastings (1970), the chain $(\theta^{(t)})$ or a transformed chain $(h(\theta^{(t)}))$ can be considered from a *time series* point of view (see Gouriéroux and Monfort 1990 or Brockwell and Davis 1996, for an introduction). For instance, under an adequate parameterization, we can model $(\theta^{(t)})$ as an ARMA(p,q) process, estimating the parameters p and q, and then use partially empirical convergence control methods. Geweke (1992) proposed to use the *spectral density* of $h(\theta^{(t)})$,

$$S_h(w) = \frac{1}{2\pi} \sum_{t=-\infty}^{t=\infty} \text{cov}\left(h(\theta^{(0)}), h(\theta^{(t)})\right) e^{inw},$$

where i denotes the complex square root of 1, that is

$$e^{inw} = \cos(nw) + i \sin(nw).$$

The spectral density relates to the asymptotic variance of (8.1.1) since the limiting variance γ_h of Proposition 4.9.8 is given by

$$\gamma_h^2 = S_h^2(0)\ .$$

Estimating S_h by non-parametric methods like *the kernel method* (see Bosq and Lecoutre 1988), Geweke (1992) takes the first T_A observations and the last T_B observations from a sequence of length T to derive

$$\delta_A = \frac{1}{T_A} \sum_{t=1}^{T_A} h(\theta^{(t)}), \quad \delta_B = \frac{1}{T_B} \sum_{t=T-T_B+1}^{T} h(\theta^{(t)})$$

and the estimates σ_A^2 and σ_B^2 of $S_h(0)$ based on both subsamples, respectively. Asymptotically (in T), the difference

$$\frac{\sqrt{T}(\delta_A - \delta_B)}{\sqrt{\dfrac{\sigma_A^2}{\tau_A} + \dfrac{\sigma_B^2}{\tau_B}}} \tag{8.6.1}$$

is a standard normal variable (with $T_A = \tau_A T$, $T_B = \tau_B T$ and $\tau_A + \tau_B < 1$). We can therefore derive a convergence diagnostic from (8.6.1), and a determination of the size t_0 of the training sample. The values suggested by Geweke (1992) are $\tau_A = 0.1$ and $\tau_B = 0.5$.

A global criticism of this spectral approach also applies to all the methods using a non-parametric intermediary step to estimate a parameter of the model, namely that they necessarily induces losses in efficiency in the processing of the problem (since they are based on a less constrained representation of the model). Moreover, the calibration of non-parametric estimation methods (as the choice of the *window* in the kernel method) is always delicate since it is not standardized. We therefore refer to Geweke (1992) for a more detailed study of this method, which is used in some softwares (see Best, Cowles and Vines 1995, for instance). Other approaches based on spectral analysis are given in Heidelberger and Welsh (1988) and Schruben, Singh and Tierney (1983) to test the stationarity of the sequence by Kolmogorov–Smirnov tests (see Cowles and Carlin 1994 and Brooks and Roberts 1995, for a discussion). Note that Heidelberger and Welch (1983) test stationarity via a Kolmogorov-Smirnov test, based on

$$B_T(S) = \frac{S_{[Ts]} - [Ts]\bar{\theta}}{(T\hat{\psi}(0))^{1/2}}\,, \qquad 0 \le s \le 1$$

where

$$S_t = \sum_{t=1}^{t} \theta^{(t)}, \quad \bar{\theta} = \frac{1}{T}\sum_{t=1}^{T} \theta^{(t)}\,,$$

and $\hat{\psi}(0)$ is an estimate of the spectral density. For large T, B_T is approximately a Brownian bridge and can be tested as such. Their method thus provide the theoretical background to Yu and Mykland's (1993) CUSUM criterion (see §8.3.1).

8.6.3 Further discretizations

Garren and Smith (1993) use the same discretization $z^{(t)}$ as Raftery and Lewis (1992a,b). They show that, under some conditions, there exist α and $1 > |\lambda_2| > |\lambda_3|$ such that the quantity

$$\begin{aligned} \varrho_t &= \mathbb{E}[z^{(t)}] \\ &= P(\theta^{(t)} \le \underline{\theta}) \\ &= \varrho + \alpha\lambda_2^t + O(|\lambda_3|^t), \end{aligned}$$

with ϱ the limiting value of ϱ_t. (These conditions are related with the eigenvalues of the functional operator associated with the transition kernel of the original chain $(\theta^{(t)})$ and with Hilbert-Schmidt conditions, see §7.6.2 and Brooks and Roberts 1995.) In their study of the convergence of ϱ_t to ϱ, Garren and Smith (1993) propose to use m parallel chains $(\theta_\ell^{(t)})$, with the same initial value $\theta_\ell^{(0)}$ $(1 \le \ell \le m)$. They then approximate ϱ_t by

$$\hat{\varrho}_t = \frac{1}{m}\sum_{\ell=1}^{m} \mathbb{I}_{\theta_\ell^{(t)} < \underline{\theta}}$$

and derive some estimations of ϱ, α and λ_2 from

$$\min_{(\rho,\alpha,\lambda_2)} \sum_{t=n_0+1}^{T} (\hat{\varrho}_t - \varrho + \alpha\lambda_2^t)^2,$$

where n_0 and T need to be calibrated. When T is too high, the estimators of α and λ_2 are instable, and Garren and Smith (1993) suggest to choose T such that $\hat{\alpha}$ and $\hat{\lambda}_2$ remain stable. When compared with the original binary control method, the approach of Garren and Smith (1993) does not require a preliminary evaluation of (α, β), but it is quite costly in simulations. Moreover, the expansion of ϱ_t around ϱ is only valid under conditions which cannot be verified in practice.

8.6.4 The `CODA` *software*

While the methods presented in this chapter are at various stages of their development, some of the most common techniques have been aggregated in an S-Plus software called CODA, developed by Best, Cowles and Vines (1996). While originally intended as an output processor for the BUGS software (see §7.6.3), this software can also be used to analyze the output of Gibbs sampling and Metropolis–Hastings algorithms. The techniques selected by Best et al. (1996) are mainly those described in Cowles and Carlin (1996), that is the convergence diagnostics of Gelman and Rubin (1992) (§8.3.4), Geweke (1992) (§8.6.2), Heidelberger and Welch (1983) (§8.6.1), Raftery and Lewis (1992a) (§8.4.1), plus plots of autocorrelation for each variable and of cross-correlations between variables. The MCMC output must however be presented in a very specific S-plus format to be processed by CODA.

8.6.5 Perfect simulation

Although the following imperative is rather in opposition with the theme of the previous chapters, in the sense that MCMC methods have been precisely introduced to overcome the difficulties of simulating directly from a given distribution, it is essential, in some settings, to start the Markov chain in its stationary regime, $\theta^{(0)} \sim f(x)$, because the bias caused by the initial value/distribution may be far from negligible. Moreover, if it becomes feasible to start from the stationary distribution, the convergence issues are reduced to the determination of an acceptable batch size k, so that $\theta^{(0)}, \theta^{(k)}, \theta^{(2k)}, \ldots$ are nearly independent, and to the accuracy of an ergodic average. In other cases, one needs to know, as put by Fill (1996), "how long is long enough", in order to evaluate the necessary computing time or the mixing properties of the chain.

Following Propp and Wilson (1996), several authors have proposed devices to sample directly from the stationary distribution f, i.e. algorithms such that $\theta^{(0)} \sim f$, at varying computational costs[11] and for specific distributions and/or transitions. The denomination of *perfect sampling* for such techniques was coined by Kendall (1996), replacing the *exact sampling* terminology of Propp and Wilson (1996) with a more triumphing qualification! The main bulk of the work on perfect sampling deals, so far, with finite state spaces; this is due, for one thing, to the greater simplicity of these spaces and, for another, to statistical physics motivations related to the Ising model (see Example 7.1.3). The appeal of these methods for mainstream statistical problems is yet unclear, from both points of view of (a) speeding up convergence and (b) controlling convergence, but Murdoch and Green (1997) have shown that some standard examples in continuous settings, like the nuclear pump failure model of Example 7.1.18 (see Problem 8.15), do allow for *perfect sampling.* Note also that in settings when the Duality Principle of §7.2 applies, the stationarity of the finite chain obviously transfers to the dual chain, even if the latter is continuous.

In a finite state space $\mathcal{X}$ of size k, the method proposed by Propp and Wilson (1996) is called *coupling from the past* (CFTP). It runs in parallel k chains corresponding to all possible starting points in $\mathcal{X}$ farther and farther back in time till all chains take the same value (or *coalesce*) at time 0 (or earlier).

[11] In most cases, the computation time required to produce $\theta^{(0)}$ exceeds by orders of magnitude the computation time of a $\theta^{(t)}$ from the transition kernel.

CHAPTER 9

Assimilation and Application: Missing Data Models

Version 1.2 February 27, 1998

9.1 Introduction

Missing data models (introduced in §5.3.1) seem to call naturally for simulation, in order for it to replace the missing data part so that one can proceed with a "classical" inference on the complete model. However, this intuition has taken a while to formalize correctly, and to go further than mere *ad hoc* solutions with no true theoretical justification. It is only with the EM algorithm that Dempster et al. (1977) (see §5.3.3) came up with a rigorous and general formulation of statistical inference though completion of missing data. This algorithm nonetheless requires a high degree of analyticity to compute the expectation (E) step and therefore cannot be used in all settings. As mentioned in §5.3.4 and §5.5.1, stochastic versions of EM (Broniatowski, Celeux and Diebolt 1983, Celeux and Diebolt 1985, Wei and Tanner 1990, Qian and Titterington 1991, Lavielle and Moulines 1997) have come closer to simulation goals by replacing the E step with a simulated completion of missing data, without however preserving the whole range of EM convergence properties.

This chapter mainly aims at illustrating the potential of Markov Chain Monte Carlo algorithms in the Bayesian analysis of missing data models, although it does not intend to provide the reader with an exhaustive treatment of these models. It must rather be understood as a sequence of examples on a common theme. See Everitt (1984), Little and Rubin (1987), Tanner (1991) or MacLachlan and Krishnan (1996) for deeper perspectives in this domain.

9.2 First examples

9.2.1 Discrete Data Models

Numerous settings (surveys, medical experiments, epidemiological studies, design of experiment, quality control, etc.) produce a *grouping* of the original observations in less informative categories, often for reasons beyond the control of the experimenter. Heitjan and Rubin (1991) (see also Rubin 1987) call the resulting process *Data coarsening* and study the effect of data aggregation on the inference. (In particular, they examine whether the grouping procedure has an effect on the likelihood and thus must be taken into account.)

Example 9.2.1 –**Rounding effect**– Heitjan and Rubin (1991) consider some random variables $y_i \sim \mathcal{E}xp(\theta)$ grouped in observations x_i $(1 \le i \le n)$ according to the procedure

$$g_i|y_i \sim \mathcal{B}(1, \Phi(\gamma_1 - \gamma_2 y_i)), \qquad x_i|g_i, y_i = \begin{cases} [y_i] & \text{if } g_i = 1, \\ 20[y_i/20] & \text{otherwise,} \end{cases}$$

where Φ is the cdf of the normal distribution $\mathcal{N}(0,1)$ and $[a]$ denotes the integral part of a. This model describes, for instance, the approximation bias to the inferior pack in a study on smoking habits, under the assumption that this bias increases with the daily consumption of cigarettes $[y_i]$. If the g_i's are known, the completion of the model is straightforward. Otherwise, the conditional distribution

$$\begin{aligned} \pi(y_i|x_i, \theta, \gamma_1, \gamma_2) \quad & \propto \quad e^{-\theta y_i} \left\{ \mathbb{I}_{[x_i, x_i+1]}(y_i) \Phi(\gamma_1 - \gamma_2 y_i) \right. \\ & \qquad \left. + \mathbb{I}_{[x_i, x_i+20]}(y_i) \Phi(-\gamma_1 + \gamma_2 y_i) \right\} \\ & = \quad e^{-\theta y_i} \left\{ \mathbb{I}_{[x_i, x_i+1]}(y_i) + \mathbb{I}_{[x_i+1, x_i+20]}(y_i) \Phi(-\gamma_1 + \gamma_2 y_i) \right\} \end{aligned}$$

is useful for the completion of the model through Gibbs sampling. This distribution can be simulated directly by an accept-reject algorithm (which requires a good approximation of Φ—see Example 2.2.1) or by introducing an additional artificial variable t_i such that

$$\begin{aligned} t_i|y_i \quad & \sim \quad \mathcal{N}_+(\gamma_1 - \gamma_2 y_i, 1, 0), \\ y_i|t_i \quad & \sim \quad \theta e^{-\theta y_i} \left\{ \mathbb{I}_{[x_i, x_i+1]}(y_i) + \mathbb{I}_{[x_i+1, x_i+20]}(y_i) \; \frac{e^{-t_i^2/2}}{\sqrt{2\pi}} \right\}, \end{aligned}$$

where $\mathcal{N}_+(\mu, \sigma^2, 0)$ denotes the normal distribution $\mathcal{N}(\mu, \sigma^2)$ truncated to $\mathbb{R}_+$ (see Example 2.2.12). The two distributions above can then be completed by

$$(\theta, \gamma_1, \gamma_2) \sim \theta^n \exp\left\{ -\theta \sum_{i=1}^n y_i \right\} \mathbb{I}_{\gamma_1 > \max(\gamma_2 y_i + t_i)} \; \pi(\theta, \gamma_1, \gamma_2)$$

to provide a Gibbs sampling algorithm. ||

When several variables are studied simultaneously in a sample, each corresponding to a grouping of individual data, the result is a *contingency table.* If the context is sufficiently informative to allow for a modeling of the individual data, the completion of the contingency table (by reconstruction of the individual data) may facilitate inference about the phenomenon under study.

		Diameter (inches)	
		≤ 4.0	> 4.0
Height (feet)	> 4.75	32	11
	≤ 4.75	86	35

Table 9.2.1. *Observation of two characteristics of the habitat of 164 lizards (Source: Schoener, 1968).*

Example 9.2.2 – **Lizard habitat** – Schoener (1968) studies the habitat of lizards, in particular the relation between height and diameter of the branches where they sleep. Table 9.2.1 provides the information available on these two parameters.

To test the independence between these factors, Fienberg (1977) proposes a classical solution based on a χ^2 test. A possible alternative is based to assume a parametric distribution on the individual observations y_{ijk} of diameter and of height $(i, j = 1, 2,\ k = 1, \ldots, n_{ij})$, for instance a multidimensional normal distribution with mean $\theta = (\theta_1, \theta_2)$ and covariance matrix

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} = \sigma^2 \Sigma_0 .$$

The likelihood associated with Table 9.2.1 is then implicit, even in the case $\rho = 0$. However, the model can be completed by simulation of the individual values and this allows for the approximation of the posterior distribution associated with the prior

$$\pi(\theta, \sigma, \rho) = \frac{1}{\sigma}\, \mathbb{I}_{[-1,1]}(\rho) ,$$

through an Markov Chain Monte Carlo algorithm. In fact, if $n_{11} = 32$, $n_{12} = 11$, $n_{21} = 86$ and $n_{22} = 35$, and if $\mathcal{N}_2^T(\theta, \Sigma; Q_{ij})$ represents the normal distribution restricted to one of the four quadrants Q_{ij} induced by $(\log(4.75), \log(4))$, the steps of the Gibbs sampler are

Algorithm A.46 –*Contingency table completion*–

```
1. Simulate $y_{ijk} \sim \mathcal{N}_2^T(\theta, \Sigma; Q_{ij})$      $(i, j = 1, 2,\ k = 1, \ldots, n_{ij})$;
2. Simulate $\theta \sim \mathcal{N}_2(\overline{y}, \Sigma/164)$;
3. Simulate $\sigma^2$ from the inverted gamma distribution                [A.46]
       $\mathcal{IG}\left(164, \frac{1}{2} \sum_{i,j,k} (y_{ijk} - \theta)^t\ \Sigma_0^{-1} (y_{ijk} - \theta)\right)$ ;
4. Simulate $\rho$ according to
   (9.2.1) $(1 - \rho^2)^{-164/2} \exp\left\{ -\frac{1}{2} \sum_{i,j,k} (y_{ijk} - \theta)^t\ \Sigma^{-1} (y_{ijk} - \theta) \right\}$ ,
```

where $\overline{y} = \sum_{i,j,k} y_{ijk}/164$. The normal truncated distribution can be simulated by the algorithm of Geweke (1991) or Robert (1995), but the distribution (9.2.1) requires a Metropolis–Hastings step based, for instance, on an inverse Wishart distribution ||

Another setup where grouped data appears in a natural fashion is made of *qualitative models*. See for instance Gouriéroux (1989) for an exhaustive processing of *probit* and *logit* models. The logit model being treated in Problems 7.15 and 9.1, we consider instead the probit model, where some binary variables y_i, taking values in $\{0, 1\}$ and associated with a vector $x_i \in \mathbb{R}^p$ of covariates, are modeled through a Bernoulli distribution

$$p_i = \Phi(x_i^t\beta) , \qquad \beta \in \mathbb{R}^p. \tag{9.2.2}$$

Even though the model (9.2.2) is generally defined in this form, a completed model can be introduced where the completed data y_i^* is grouped according to their sign, that is

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Given a conjugate distribution $\mathcal{N}_p(\beta_0, \Sigma)$ on β, the algorithm which approximates the posterior distribution $\pi(\beta|y_1, \ldots, y_n, x_1, \ldots, x_n)$ is then

Algorithm A.47 –***Probit posterior distribution***–

1. Simulate

$$y_i^* \sim \begin{cases} \mathcal{N}_+(x_i^t\beta, 1, 0) & \textit{if } y_i = 1, \\ \mathcal{N}_-(x_i^t\beta, 1, 0) & \textit{if } y_i = 0, \end{cases} \qquad (i = 1, \ldots, n) \qquad [A.47]$$

2. Simulate

$$\beta \sim \mathcal{N}_p((\Sigma^{-1} + XX^t)^{-1}(\Sigma^{-1}\beta_0 + \sum_i y_i^* x_i), (\Sigma^{-1} + XX^t)^{-1}).$$

where $\mathcal{N}_+(\mu, \sigma^2, \underline{u})$ and $\mathcal{N}_-(\mu, \sigma^2, \overline{u})$ denote the normal distribution truncated on the left in $\underline{u}$, and the normal distribution truncated on the right in $\overline{u}$, respectively, and where X is the matrix whose columns are made of the x_i's. See Albert and Chib (1993b) for an application of this model to longitudinal data for medical experiments and some details on the implementation of $[A.47]$.

9.2.2 Data missing at random

Numerous settings may lead to sets of incomplete observations. For instance, a survey with multiple questions may include non-answers to some personal questions; a calibration experiment may lack observations for some values of the calibration parameters; a pharmaceutical experiment on the aftereffects of a toxic product may skip some doses for a given patient, etc. The analysis of such structures is complicated by the fact that the failure to observe is not always explained. If these missing observations are entirely due to chance, it follows that the incompletely observed data only play a

	Men		Women	
Age	Single	Maried	Single	Maried
< 30	20.0	21.0	16.0	16.0
	24/1	5/11	11/1	2/2
> 30	30.0	36.0	18.0	–
	15/5	2/8	8/4	0/4

Table 9.2.2. *Average incomes and numbers of responses/non-responses to a survey on the income by age, sex and marital status. (Source: Little and Rubin, 1987.)*

role through their marginal distribution. However, these distributions are not always explicit and a natural approach leading to a Gibbs sampler algorithm is to replace the missing data by simulation.

Example 9.2.3 –**Non-ignorable non-response**– Table 9.2.2 describes the (fictious) results of a survey on the income depending on age, sex and family status. The observations are grouped by average, and we assume an exponential shape for the individual data,

$$y^*_{a,s,m,i} \sim \mathcal{E}xp(\mu_{a,s,m}) \quad \text{with} \quad \mu_{a,s,m} = \mu_0 + \alpha_a + \beta_s + \gamma_m \ ,$$

where $1 \leq i \leq n_{a,s,m}$ and where α_a $(a = 1, 2)$, β_s $(s = 1, 2)$ and γ_m $(m = 1, 2)$ correspond to the *age* (junior/senior), *sex* (fem./male) and *family* (single/married) effects, respectively. The model gets identifiable through constraints like $\alpha_1 = \beta_1 = \gamma_1 = 0$. An important difficulty appears when the lack of answer depends on the income, say in the shape of a logit model,

$$p_{a,s,m,i} = \exp\{w_0 + w_1 y^*_{a,s,m,i}\} \Big/ 1 + \exp\{w_0 + w_1 y^*_{a,s,m,i}\},$$

where $p_{a,s,m,i}$ denotes the probability of non-response.

If the prior distribution on the parameter set is the Lebesgue measure on $\mathbb{R}^2$ for (w_0, w_1) and on $\mathbb{R}_+$ for $\mu_0, \alpha_2, \beta_2, \gamma_2$, a direct analysis of the likelihood is not feasible analytically.

On the contrary, the likelihood of the complete model is much more explicit since it is

$$\prod_{\substack{a=1,2\\ s=1,2\\ m=1,2}} \prod_{i=1}^{n_{a,s,m}} \frac{\exp\{z^*_{a,s,m,i}(w_0 + w_1 y^*_{a,s,m,i})\}}{1 + \exp\{w_0 + w_1 y^*_{a,s,m,i}\}} (\mu_0 + \alpha_a + \beta_s + \gamma_m)^{r_{a,s,m}}$$

$$(9.2.3) \quad \times \quad \exp\left\{-r_{a,s,m}\overline{y}_{a,s,m}(\mu_0 + \alpha_a + \beta_s + \gamma_m)\right\}$$

where $z_{a,s,m,i} z^*_{a,s,m,i}$ represents the indicator of missing observation, while $n_{a,s,m}$, $r_{a,s,m}$ and $\overline{y}_{a,s,m}$ denote the number of persons covered by the survey, the number of responses and the average of these responses by category, respectively.

9.19 (Billio, Monfort and Robert 1998) A *dynamic desequilibrium* model is defined as the observation of

$$y_t = \min(y^*_{1t}, y^*_{2t}),$$

where the y^*_{it} are distributed from a parametric joint model, $f(y^*_{1t}, y^*_{2t})$.

a Give the distribution of (y^*_{1t}, y^*_{2t}) conditional on y_t.

b Show that a possible completion of the model is to first draw the regime (1 versus 2) and then draw the missing component.

c Show that when $f(y^*_{1t}, y^*_{2t})$ is Gaussian, the above steps can be implemented without approximation.

9.20 (Billio, Monfort and Robert 1998) Consider the *factor ARCH* model defined by $(t = 1, \dots, T)$

$$\begin{cases} y^*_t = (\alpha + \beta({}^*_{t-1})^2)^{1/2} \epsilon^*_t, \\ y_t = a y^*_t + \epsilon_t, \end{cases}$$

where the ϵ^*_t's are iid $\mathcal{N}(0,1)$ and the $epsilon_t$'s are $\mathcal{N}_p(0, \Sigma)$. The latent variables y^*_t are not observed.

a Propose a noninformative prior distribution on the parameter $\theta = (\alpha, \beta, a, \Sigma)$ which leads to a proper posterior distribution.

b Propose a completion step for the latent variables based on $f(y^*_t | y_t, y^*_{t-1}, \theta)$.

References

Aarts, E. and Kors, T.J. (1989) *Simulated Annealing and Boltzman Machines: a Stochastic Approach to Combinatorial Optimisation and Neural Computing.* J. Wiley, New York.

Abramowitz, M. and Stegun, I. (1964) *Handbook of Mathematical Functions.* Dover, New York.

Ackley, D.H., Hinton, G.E. and Sejnowski, T.J. (1985) A learning algorithm for Boltzmann machines. *Cognitive Science* **9**, 147–169.

Ahrens, J. and Dieter, U. (1974) Computer methods for sampling from gamma, beta, Poisson and binomial distributions. *Computing* **12**, 223–246.

Albert, J.H. (1988) Computational methods using a Bayesian hierarchical generalized linear model. *J. Amer. Statist. Assoc.* **83**, 1037–1044.

Albert, J.H. and Chib, S. (1993a) Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *J. Business Economic Statistics* **1**, 1–15.

Albert, J.H. and Chib, S. (1993b) Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88**, 669–679.

Asmussen, S. (1979) *Applied Probability and Queues.* J. Wiley, New York.

Asmussen, S., Glynn, P.W. and Thorisson, H. (1992) Stationarity detection in the initial transient problem. *ACM Trans. Modelling and Computer Simulations* **2**, 130–157.

Athreya, K.B. and Ney, P. (1978) A new approach to the limit theory of recurrent Markov chains. *Trans. Amer. Math. Soc.* **245**, 493–501.

Atkinson, A. (1979) The computer generation of Poisson random variables. *Appl. Statist.* **28**, 29–35.

Archer, G.E.B. and Titterington, D.M. (1995) Parameter estimation for hidden Markov chains. Tech. report, Dept. of Stat., U. of Glasgow.

Azencott, R. (1988) Simulated annealing. *Séminaire Bourbaki 40ième année, 1987–1988* **697**.

Barbe, P. and Bertail, P. (1994) *The Weighted Bootstrap.* Lecture Notes in Statistics **98**, Springer–Verlag.

Barbieri, M.M. and O'Hagan, T.J. (1996) A reversible jump MCMC sampler for Bayesian analysis of ARMA time series. Tech. report, Uni. "La Sapienza", Roma.

Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70**, 343-365.

Barndorff-Nielsen, O. (1991). Modified signed log likelihood ratio. *Biometrika* **78**, 557-563.

Barndorff-Nielsen, O.E. and Cox, D.R. (1989) *Asymptotic Techniques for Use in Statistics.* Chapman and Hall, London.

Barone, P. and Frigessi, A. (1989) Improving stochastic relaxation for Gaussian random fields. *Biometrics* **47**, 1473–1487.

Barry, D. and Hartigan, J. (1992) Product partition models for change point problems. *Ann. Statist.* **20**, 260–279.

Barry, D. and Hartigan, J. (1993) A Bayesian analysis of change point problems. *J. Amer. Statist. Assoc.* **88**, 309–319.

Bauwens, L. (1984) *Bayesian Full Information of Simultaneous Equations Models Using Integration by Monte Carlo.* Lecture Notes in Economics and Mathematical Systems 232, Springer-Verlag, New York.

Bauwens, L. and Richard, J.F. (1985) A 1-1 Poly-t random variable generator with application to Monte Carlo integration. *J. Econometrics* **29**, 19–46.

Bennett, J.E., Racine-Poon, A. and Wakefield, J.C. (1996) MCMC for nonlinear hierarchical models. In *Markov chain Monte-Carlo in Practice* (Ed. W.R. Gilks, S.T. Richardson and D.J. Spiegelhalter), 339–358. Chapman and Hall, London.

Bergé, P., Pommeau, Y. and Vidal, C. (1984) *Order Within Chaos.* J. Wiley, New York.

Berger, J.O. (1985) *Statistical Decision Theory and Bayesian Analysis* (2nd edition). Springer-Verlag, New York.

Berger, J.O. (1990) Robust Bayesian analysis: sensitivity to the prior. *J. Statist. Plann. Inference* **25**, 303–328.

Berger, J.O. (1994) An overview of of robust Bayesian analysis (with discussion). *TEST* **3**, 5–124.

Berger, J.O., Philippe, A. and Robert, C.P. (1996) Estimation of quadratuc functions: reference priors for non-centrality parameters. Tech. report #96-10C, Dept. of Statistics, Purdue Uni.

Berger, J.O. and Wolpert, R. (1988) *The Likelihood Principle* (2nd edition). IMS Lecture Notes — Monograph Series **9**, Hayward, California.

Bernardo, J.M. and Giròn, F.J. (1986) A Bayesian approach to cluster analysis. In *Second Catalan International Symposium on Statistics*, Barcelona, Spain.

Bernardo, J.M. and Giròn, F.J. (1988) A Bayesian analysis of simple mixture problems. In *Bayesian Statistics 3*, J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith (Eds.), 67–78. Oxford University Press, Oxford.

Bernardo, J.M. and Smith, A.F.M. (1994) *Bayesian Theory.* J. Wiley, New York.

Beran, R. (1977) Minimum Hellinger distance estimates for parametric models. *Ann. Statist.* **5**, 445–463.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society* (Series B) **36**, 192–326.

Besag, J. (1989) Towards Bayesian image analysis. *J. Applied Statistics* **16**, 395–407.

Besag, J.E. (1994) Discussion of "Markov chains for exploring posterior distributions". *Ann. Statist.* **22**, 1734-1741.

Besag, J. and Green, P.J. (1992) Spatial Statistics and Bayesian computation (with discussion). *Journal of the Royal Statistical Society* (Series B) **55**, 25–38.

Besag, J., Green, E., Higdon, D. and Mengersen, K.L. (1995) Bayesian computation and stochastic systems (with discussion). *Statistical Science* **10**, 3–66.

Besag, J. and Mengersen, K.L. (1993) Meta-Analysis via Markov Chain Monte-Carlo methods. Tech. report, Dept. of Statistics, Colorado State Univ.

Best, D.J. (1978) Letter to the editor. *Applied Statistics* (Ser. C) **27**, 181.