*Brazilian Journal of Probability and Statistics*
2013, Vol. 27, No. 4, 416–436
DOI: 10.1214/11-BJPS174

# Bayesian computation for statistical models with intractable normalizing constants

**Yves F. Atchadé[a], Nicolas Lartillot[b] and Christian Robert[c]**

[a]*University of Michigan*
[b]*Université de Montreal*
[c]*Université Paris-Dauphine and CREST, INSEE*

**Abstract.** This paper deals with a computational aspect of the Bayesian analysis of statistical models with intractable normalizing constants. In the presence of intractable normalizing constants in the likelihood function, traditional MCMC methods cannot be applied. We propose here a general approach to sample from such posterior distributions that bypasses the computation of the normalizing constant. Our method can be thought as a Bayesian version of the MCMC-MLE approach of Geyer and Thompson [*J. Roy. Statist. Soc. Ser. B* **54** (1992) 657–699]. We illustrate our approach on examples from image segmentation and social network modeling. We study as well the asymptotic behavior of the algorithm and obtain a strong law of large numbers for empirical averages.

## 1 Introduction

Statistical inference for models with intractable normalizing constants is a computationally challenging problem. This is a well-known issue and examples include the analysis of spatial point processes [Møller and Waagepetersen (2004)], image analysis [Ibanez and Simo (2003)], protein design [Kleinman et al. (2006)] and many others. Suppose we have a data set $x_0 \in \mathcal{X}$ generated from a statistical model with density

$$e^{E(x,\theta)}/Z(\theta),$$

depending on a parameter $\theta \in \Theta$, where the normalizing constant $Z(\theta) = \int_{\mathcal{X}} e^{E(x,\theta)}\,dx$ is not available in a closed form. Adopting a Bayesian perspective, we then take $\mu$ as the prior density of the parameter $\theta \in \Theta$. The posterior distribution of $\theta$ given $x_0$ is then given by

$$\pi(\theta|x_0) \propto \frac{1}{Z(\theta)} e^{E(x_0,\theta)} \mu(\theta), \tag{1.1}$$

where the proportionality sign means that $\pi(\theta|x_0)$ differs from the right-hand side expression only by a multiplicative constant (in $\theta$). Since $Z(\theta)$ cannot be easily

*Key words and phrases.* Monte Carlo methods, adaptive Markov Chain Monte Carlo, Bayesian inference, doubly-intractable distributions, Ising model.

evaluated, Monte Carlo simulation from this posterior distribution is particularly problematic even when using the Markov Chain Monte Carlo (MCMC). Murray et al. (2006) use the term *doubly intractable distribution* to refer to posterior distributions of the form (1.1). Indeed, state-of-the-art Monte Carlo sampling methods do not allow one to deal with such models in a Bayesian framework. For example, a Metropolis–Hastings algorithm with proposal kernel $Q$ and target distribution $\pi$ would have acceptance ratio

$$\min\left(1, \frac{e^{E(x_0,\theta')}}{e^{E(x_0,\theta)}} \frac{Z(\theta)}{Z(\theta')} \frac{\mu(\theta')}{\mu(\theta)} \frac{Q(\theta',\theta)}{Q(\theta,\theta')}\right),$$

which cannot be computed as it involves the intractable normalizing constant $Z$ evaluated both at $\theta$ and $\theta'$.

An early attempt to deal with this type of model in the frequentist framework is the pseudo-likelihood approximation of Besag (1974) which replaces the likelihood function $e^{E(x_0,\theta)}/Z(\theta)$ by a more tractable pseudo-likelihood function. But there are cases where this approximation is known to perform poorly [see, e.g., Marin et al. (2009)]. Maximum likelihood inference is equally possible: for instance, MCMC-MLE, a maximum likelihood approach using MCMC, was developed in the 90s [Geyer and Thompson (1992); Geyer (1994)]. Another related approach to find MLE estimates is Younes' algorithm [Younes (1988)] based on stochastic approximation. See also Ibanez and Simo (2003) for an interesting simulation study comparing the above three frequentist approaches.

Comparatively, little work has been done to develop computationally exact methods for the Bayesian approach to this problem. Indeed, various approximate algorithms exist in the literature, often based on path sampling [see, e.g., Gelman and Meng (1998)]. Recently, Møller et al. (2006) have shown that if exact sampling from $e^{E(x,\theta)}/Z(\theta)$ (as a density on the sample space $\mathcal{X}$) is possible, then a valid MCMC algorithm converging to (1.1) can be constructed. See also Murray et al. (2006) for some recent improvements. This approach relies on a clever auxiliary variable algorithm, but it requires a manageable perfect sampler and intractable normalizing constants often occur in models for which exact sampling of $X$ is either impossible or expensive.

In this paper, we develop an adaptive Monte Carlo approach that samples from (1.1). Our algorithm generates a process $\{\theta_n, n \geq 0\}$ (usually not Markov) such that as $n \to \infty$, the marginal distribution of $\theta_n$ converges to (1.1). In principle, any method to sample from (1.1) needs to deal with the intractable normalizing constant $Z(\theta)$. In the auxiliary variable method of Møller et al. (2006), computing the function $Z(\theta)$ is replaced by a perfect sampling step from $e^{E(x,\theta)}/Z(\theta)$ that is used to obtained an unbiased estimate of $Z(\theta)$. As mentioned above, this strategy works well as long as perfect sampling is feasible and inexpensive, but Marin et al. (2009) show that this is rarely the case for Potts models. In the present work, we follow a completely different approach building on the idea of estimating the function $Z$ on the fly, that is, during the simulation. The starting point of the method is importance sampling. Suppose that for some $\theta^{(0)} \in \Theta$, we can sample (possibly via MCMC) from the density $e^{E(x,\theta^{(0)})}/Z(\theta^{(0)})$ in $\mathcal{X}$. Based on such a sample, we can then estimate the ratio $Z(\theta)/Z(\theta^{(0)})$ for any $\theta \in \Theta$, for instance, via importance sampling. This idea is the same one behind the MCMC-MLE algorithm of Geyer and Thompson (1992). However, such estimates are typically poor when $\theta$ is taken too far from $\theta^{(0)}$. Instead of a single point $\theta^{(0)}$, suppose that we have a set of particles $\{\theta^{(i)}, i = 1, \ldots, d\}$ in $\Theta$ and sample from

$$\Lambda^\star(x, i) \propto e^{E(x,\theta^{(i)})}/Z\bigl(\theta^{(i)}\bigr)$$

on $\mathcal{X} \times \{1, \ldots, d\}$. Then, to the extent that the particles $\{\theta^{(i)}, i = 1, \ldots, d\}$ cover $\Theta$ well, a good estimate of $Z(\theta)$ (up to a multiplicative constant) can be obtained for any $\theta \in \Theta$. This is the strategy adopted in this work. We propose an algorithm that generates a process $\{(X_n, I_n, \theta_n), n \geq 0\}$ such that the marginal distribution of $(X_n, I_n)$ converges to $\Lambda^\star$ given above and the marginal distribution of $\theta_n$ converges to (1.1).

The paper is organized as follows. In Section 2.1, we describe a general adaptive Markov Chain Monte Carlo strategy to sample from target distributions of the form (1.1). A full description of the practical algorithm including practical implementation details is given in Section 2.2. We illustrate the algorithm with three examples, namely, the Ising model, a Bayesian image segmentation example and a Bayesian modeling of a social network. In this latter example, and to the best of our knowledge, exact sampling from the model is not feasible and Møller et al. (2006) cannot be applied. Those examples are presented in Section 4. Some theoretical aspects of the method are discussed in Section 3, while the proofs are postponed till Section 6. Some concluding remarks are gathered in Section 5.

## 2 Posterior distributions with intractable normalizing constants

### 2.1 A general adaptive approach

In this section, we outline a general strategy to sample from (1.1), which provides a unifying framework for a better understanding of the specific algorithm discussed in Section 2.2. We assume that $\Theta$ is a compact subset of a finite dimensional Euclidean space equipped with its Borel $\sigma$-algebra $\mathcal{B}(\Theta)$. Let $\mathcal{C}(\Theta)$ denote the set of all continuous functions $\zeta : \Theta \to \mathbb{R}$. We equip $\mathcal{C}(\Theta)$ with the supremum distance. Throughout, we assume that the function $\theta \to |E(x_0, \theta)|$ is bounded from above and for $\zeta \in \mathcal{C}(\Theta)$, $\pi_\zeta$ denotes the density on $\Theta$ defined by

$$\pi_\zeta(\theta) \propto \exp\bigl(E(x_0, \theta) - \zeta(\theta)\bigr). \tag{2.1}$$

We assume that for any $\theta \in \mathcal{C}(\Theta)$, we can construct a transition kernel $P_\zeta$ with invariant distribution $\pi_\zeta$ such that the maps $\zeta \to P_\zeta h(\theta)$ and $\zeta \to \pi_\zeta(h)$ are measurable maps for any bounded measurable function $h : \Theta \to \mathbb{R}$ and $\theta \in \Theta$.

In the above, $\pi_\zeta(h)$ and $P_\zeta h(\theta)$ are defined respectively as $\int \pi_\zeta(\vartheta)h(\vartheta)\,d\vartheta$ and $\int P_\zeta(\theta, d\vartheta)h(\vartheta)$. For instance, $P_\zeta$ may be a Metropolis–Hastings kernel with invariant distribution equal to $\pi_\zeta$. For a signed measure $\mu$, we define its total variation norm as $\|\mu\|_{\mathrm{TV}} := \sup_{|f|\le 1} |\mu(f)|$, while, for a transition kernel $P$, we define its iterates as $P^0(\theta, A) = \mathbf{1}_A(\theta)$ and $P^j(\theta, A) := \int P^{j-1}(\theta, dz)P(z, A)$ $(j > 0)$.

Let $\{\zeta_n, n \ge 0\}$ be a $\mathcal{C}(\Theta)$-valued stochastic process (random field) defined on some probability space $(\Omega, \mathcal{F}, \Pr)$ equipped with a filtration $\{\mathcal{F}_n, n \ge 0\}$. We assume that $\{\zeta_n, n \ge 0\}$ is $\mathcal{F}_n$-adapted. The sequence $\{\zeta_n, n \ge 0\}$ is interpreted as a sequence of estimators for $\log Z$. We will see below how to build such estimators in practice. At this stage, we make the following theoretical assumptions:

(A1) For any bounded measurable function $h : \Theta \to \mathbb{R}$,

$$\pi_{\zeta_n}(h) \longrightarrow \pi_z(h) \qquad \text{a.s.}$$

as $n \to \infty$, where $z$ is the function of log-normalizing constants: $z(\theta) = \log Z(\theta)$.

(A2) Moreover,

$$\sup_{\theta\in\Theta} \|P_{\zeta_n}(\theta, \cdot) - P_{\zeta_{n-1}}(\theta, \cdot)\|_{\mathrm{TV}} \longrightarrow 0, \qquad \text{a.s. as } n \to \infty;$$

(A3) and there exists $\rho \in (0, 1)$ such that for all integers $n \ge 0$,

$$\sup_{k\ge 0}\sup_{\theta\in\Theta} \|P^n_{\zeta_k}(\theta, \cdot) - \pi_{\zeta_k}(\cdot)\|_{\mathrm{TV}} \le \rho^n.$$

When such a sequence of random fields is available, we can construct a Monte Carlo process $\{\theta_n, n \ge 0\}$ on $(\Omega, \mathcal{F}, \Pr)$ to sample from $\pi$ as follows:

**Algorithm 2.1.** 1. *Initialize* $\theta_0 \in \Theta$ *arbitrarily.*
2. *Given* $\mathcal{F}_n$ *and* $\theta_n$, *generate* $\theta_{n+1}$ *from* $P_{\zeta_n}(\theta_n, \cdot)$.

This algorithm provides a converging approximation to $\pi$:

**Theorem 2.1.** *Assume that* (A1)–(A3) *hold and let* $\{\theta_n, n \ge 0\}$ *be given by Algorithm* 2.1. *For any bounded measurable function* $h : \Theta \to \mathbb{R}$,

$$n^{-1}\sum_{k=1}^{n} h(\theta_k) \longrightarrow \pi(h), \qquad \textit{a.s. as } n \to \infty.$$

**Remark 2.1.** 1. When $\Theta$ is indeed compact, assumptions (A1)–(A3) hold in most examples where intractable normalizing constants are an issue. But assuming that $\Theta$ is compact can be restrictive in some cases. This assumption can be relaxed in principle, but at the expense of greater technicalities that are beyond the scope of this paper.

2. Algorithm 2.1 gives a generic Monte Carlo approach for sampling from $\pi$, which can be implemented if we manage to obtain a consistent sequence of estimates of $z$. We show how to construct such estimates in Section 2.2.

## 2.2 Adaptive MCMC for doubly-intractable distributions

Let $\{\theta^{(i)}, i = 1, \dots, d\}$, $\theta^{(i)} \in \Theta$, be a set of $d$ particles, and let $\Lambda^\star$ be the probability density on the product space $\mathcal{X} \times \{1, \dots, d\}$ given by

$$\Lambda^\star(x, i) = \frac{1}{d} \exp\bigl(E\bigl(x, \theta^{(i)}\bigr) - z\bigl(\theta^{(i)}\bigr)\bigr), \qquad (x, i) \in \mathcal{X} \times \{1, \dots, d\}, \tag{2.2}$$

where $z(\theta^{(i)}) = \log Z(\theta^{(i)})$. The main motivation for introducing the distribution $\Lambda^\star$ is that sampling from this distribution allows to sample from the densities $x \to \exp(E(x, \theta^{(i)}) - z(\theta^{(i)}))$, $1 \le i \le d$, all at once. In general, exact sampling from $\Lambda^\star$ is not possible. Suppose that we can simulate a Markov chain $\{(X_n, I_n), n \ge 0\}$ on $\mathcal{X} \times \{1, \dots, d\}$ with target distribution $\Lambda^\star$. Then an estimate of $Z$ can be easily obtained. Indeed, for each $\theta \in \Theta$, let $\{\kappa_i(\theta), 1 \le i \le d\}$ be a probability mass function: $\kappa_i(\theta) \ge 0$, $\sum_{i=1}^d \kappa_i(\theta) = 1$. It is easy to see that, by importance sampling,

$$Z_n(\theta) = \sum_{i=1}^d \kappa_i(\theta) \left( \frac{1}{n+1} \sum_{k=1}^{n+1} e^{E(X_k, \theta) - E(X_k, \theta^{(i)})} \mathbf{1}_i(I_k) \right) \tag{2.3}$$

is a consistent estimate of $Z(\theta)$ for any $\theta \in \Theta$, and we can apply Algorithm 2.1 by setting $\zeta_n(\theta) = \log Z_n(\theta)$.

Unfortunately, even MCMC-sampling from $\Lambda^\star$ is intractable in general because the log-normalizing constants $z(\theta^{(i)})$ are rarely available. To deal with this difficulty, we propose an adaptive algorithm that estimate the constants $z(\theta^{(i)})$ adaptively during the simulation. For a vector of weights $c = (c(1), \dots, c(d)) \in \mathbb{R}^d$, let $\Lambda_c$ by the density on $\mathcal{X} \times \{1, \dots, d\}$ defined as

$$\Lambda_c(x, i) \propto \exp\bigl(E\bigl(x, \theta^{(i)}\bigr) - c(i)\bigr).$$

These distributions $\Lambda_c$ are the reweighting distributions of the Wang–Landau algorithm of Atchade and Liu (2010). Clearly, when $c(i) = z(\theta^{(i)}) = \log Z(\theta^{(i)})$, $1 \le i \le d$, then $\Lambda_c = \Lambda^\star$. Also, notice that the marginal distribution of $\Lambda_c$ on $\{1, \dots, d\}$ is proportional to $\{Z(\theta^{(i)}) e^{-c(i)}, 1 \le i \le d\}$ and is the uniform distribution on $\{1, \dots, d\}$ if and only if $c(i) = a + z(\theta^{(i)})$, for some finite constant $a$. This suggests the following adaptive strategy for updating the weight $c_n$. At time $n$, given $(c_n, X_n, I_n)$, we update $c_n(i)$ to $c_{n+1}(i) = c_n(i) + \gamma(\mathbf{1}_{\{i\}}(I_n) - \frac{1}{d})$, for some discount factor $\gamma > 0$. In this update, the weight $c_n(I_n)$ of the visited particle $\theta^{(I_n)}$ is increased to $c_n(I_n) + \gamma(1 - \frac{1}{d})$, which makes this particle slightly less likely to be visited in subsequent iterations. Similarly, nonvisited particles will see their weights decreased as in $c_{n+1}(i) = c_n(i) - \gamma/d$, making them more likely to be visited in subsequent iterations. Therefore, if the weight $c_n$ converges, it will be towards a limit $c_\star$ for which all particles are visited equally well, which can only happen if $c_\star(i) = a + z(\theta^{(i)})$, $1 \le i \le d$, for some finite constant $a$.

To fully describe the algorithm, we assume that, for each $i \in \{1, \dots, d\}$, a transition kernel $T_i^{(\mathcal{X})}$ on $\mathcal{X}$ with invariant distribution $e^{E(x, \theta^{(i)})}/Z(\theta^{(i)})$ is available. We use the subscript $\mathcal{X}$ to emphasize that $T_i^{(\mathcal{X})}$ is a kernel on $\mathcal{X}$. We will use a slowly decreasing sequence of (possibly random) positive numbers $\{\gamma_n\}$ that satisfies

$$\sum_n \gamma_n = +\infty \quad \text{and} \quad \sum_n \gamma_n^2 < +\infty.$$

**Algorithm 2.2.** *Set* $(X_0, I_0, c_0, \theta_0) \in \mathcal{X} \times \{1, \dots, d\} \times \mathbb{R}^d \times \Theta$ *as the initial state of the algorithm, with* $\zeta_0$ *as the initial random field estimate of* $\log Z$ (*for example* $\zeta_0 \equiv 0$).

*At time* $n$, *given* $(X_n, I_n, c_n, \theta_n)$ *and* $\zeta_n$:

1. *Generate* $X_{n+1}$ *from* $T_{I_n}^{(\mathcal{X})}(X_n, \cdot)$.
2. *Generate* $I_{n+1} \in \{1, \dots, d\}$ *from* $\Pr(I_{n+1} = i) \propto e^{E(X_{n+1}, \theta^{(i)}) - c_n(i)}$.
3. *Generate* $\theta_{n+1}$ *from* $P_{\zeta_n}(\theta_n, \cdot)$.
4. *Update* $c_n$ *to* $c_{n+1}$ *as*

$$c_{n+1}(i) = c_n(i) + \gamma_n\left(\mathbf{1}_{\{i\}}(I_n) - \frac{1}{d}\right), \qquad i = 1, \dots, d; \tag{2.4}$$

*and update* $\zeta_n$ *to* $\zeta_{n+1}$ *as*

$$\zeta_{n+1}(\theta) = \log\left(\sum_{i=1}^{d} \kappa(\theta, \theta^{(i)}) e^{c_{n+1}(i)} \left[\frac{\sum_{k=1}^{n+1} e^{E(X_k, \theta) - E(X_k, \theta^{(i)})} \mathbf{1}_i(I_k)}{\sum_{k=1}^{n+1} \mathbf{1}_i(I_k)}\right]\right). \tag{2.5}$$

Before proceeding to the analysis of the convergence properties of this algorithm and discussing its calibration, a few remarks are in order:

**Remark 2.2.** 1. The last term of (2.5) is not well-defined when all the $I_k$'s are different from $i$. In this case (which necessarily occurs in the first steps of the algorithm), it can either be replaced with one or with an approximative Rao–Blackwellized version, namely, replacing both $\mathbf{1}_i(I_k)$ terms with the conditional probability

$$\Pr(I_k = i \mid \mathcal{F}_{k-1}) = \frac{e^{E(X_{k-1}, \theta^{(i)}) - c_{k-1}(i)}}{\sum_{j=1}^{d} e^{E(X_{k-1}, \theta^{(j)}) - c_{k-1}(j)}}.$$

2. It is intuitively clear that the pair $\{(\zeta_n, \theta_n)\}$ falls within the framework of Section 2.1. In particular, we will show below that assumptions (A1)–(A3) are satisfied and that Theorem 2.1 applies, thus establishing the consistency of Algorithm 2.2.

3. We introduce $\kappa$ to serve as a smoothing factor so that the particles $\theta^{(i)}$ close to $\theta$ contribute more to the estimation of $Z(\theta)$. We expect this smoothing step to reduce the variance of the overall estimate of $Z(\theta)$. In the simulations we choose

$$\kappa(\theta, \theta^{(i)}) = \frac{e^{-1/(2h^2)\|\theta - \theta^{(i)}\|^2}}{\sum_{j=1}^{d} e^{-1/(2h^2)\|\theta - \theta^{(j)}\|^2}}.$$

The value of the smoothing parameter $h$ is set by trial and error for each example. More investigation is needed to better understand how to choose $h$ efficiently but, as a default, it can be chosen as a bandwidth for the Gaussian nonparametric kernel associated with the sample of $\theta_n$'s.

4. The implementation of the algorithm requires keeping track of all the samples $X_k$ that have been generated, due to the update in (2.5). Since $\mathcal{X}$ can be a very high-dimensional space, it is clear that, in practice, this bookkeeping can significantly slow down the algorithm. But, in many cases, the function $E$ takes the form $E(x,\theta) = \sum_{l=1}^{K} S_l(x)\theta_l$ for some real-valued functions $S_l$. In such cases, we only need to keep track of the sufficient statistics $\{(S_1(X_n), \ldots, S_K(X_n)), n \geq 0\}$. All the examples discussed below fall within this latter category.

Notice also that as $n$ increases, the computational cost of computing $\zeta_n$ in (2.5) increases. If $K$ and $d$ are large, this can potentially slow down the algorithm as $n$ increases. For the examples (admittedly low-dimensional) considered below, the slowing down was minor.

5. The update of $(X_n, I_n, c_n)$ is essentially the Wang–Landau algorithm of Atchade and Liu (2010), and, as explained in that paper, Algorithm 2.2 is not Markovian. Nevertheless, the marginal distribution of $\theta_n$ will typically converge to $\pi$ as shown in Section 3.

### 2.3 Choosing $d$ and the particles $\{\theta^{(i)}\}$

The Algorithm 2.2 is sensitive to the choice of both $d$ and $\{\theta^{(i)}\}$ and we provide here some guidelines. The leading factor is that the particles $\{\theta^{(i)}\}$ need to cover reasonably well the important range of the density $\pi$ and be such that for any $\theta \in \Theta$, the density $e^{E(x,\theta)}/Z(\theta)$ in $\mathcal{X}$ can be well approximated by at least one of the densities $e^{E(x,\theta^{(i)})}/Z(\theta^{(i)})$ from an importance sampling point of view. [This is related to the Kullback divergence topology; see, e.g., Cappé et al. (2008)]. One approach to selecting the $\theta^{(i)}$ that works well in practice consists in performing few iterations of the stochastic approximation recursion related to the maximum likelihood estimation of the model [Younes (1988)]. Indeed, the normal equation of the maximum likelihood estimation of $\theta$ in the model $\exp(E(x,\theta))/Z(\theta)$ is

$$\nabla_\theta E(x_0,\theta) - \mathbb{E}_\theta[\nabla_\theta E(X,\theta)] = 0. \tag{2.6}$$

In the above, $\nabla_\theta$ denotes the partial derivative operator with respect to $\theta$, $x_0$ is the observed data set and the expectation $\mathbb{E}_\theta$ is with respect to $X \sim \exp(E(x,\theta))/Z(\theta)$. Younes (1988) has proposed a stochastic approximation to solve this equation which works as follows. Given $(X_n, \theta_n)$, generate $X_{n+1} \sim T_{\theta_n}(X_n, \cdot)$, where $T_\theta$ is a transition kernel on $\mathcal{X}$ with invariant distribution $\exp(E(x,\theta))/Z(\theta)$, and then set

$$\theta_{n+1} = \theta_n + \rho_n\bigl(\nabla_\theta E(x_0,\theta) - \nabla_\theta E(X_n,\theta)\bigr) \tag{2.7}$$

for a sequence of positive numbers $\rho_n$ such that $\sum \rho_n = \infty$.

We thus use this algorithm to set the particles. In choosing $\{\theta^{(i)}\}$, we start by generating independently $d$ points from the prior distribution $\mu$. Then we carry each particle $\theta^{(i)}$ towards the important regions of $\pi$ using the stochastic recursion given in (2.7). We typically use a constant $\rho_n \equiv \rho$, with $\rho$ taken around 0.1 depending on the size of $\Theta$. A value of $\rho$ too small will make all the particles too close together. We found a few thousand iterations of the stochastic approximation to be largely sufficient.

The value of $d$, the number of particles, should then depend on the size and on the dimension of the compact set $\Theta$. In particular, the distributions $e^{E(x,\theta^{(i)})}/Z(\theta^{(i)})$ (as densities in $\mathcal{X}$) should overlap to some extent. Otherwise, the importance sampling approximation to $e^{E(x,\theta)}/Z(\theta)$ may be poor. In the simulation examples below, we choose $d$ between 100 and 500. Note also that a restarting provision can be made in cases when the variance of the approximation (2.5) is found to be too large.

### 2.4 Choosing the step-size $\{\gamma_n\}$

It is shown in Atchade and Liu (2010) that the recursion (2.4) can be written as a stochastic approximation algorithm with step-size $\{\gamma_n\}$, so that in theory [see, e.g., Andrieu et al. (2005)], any positive sequence $\{\gamma_n\}$ such that $\sum \gamma_n = \infty$ and $\sum \gamma_n^2 < \infty$ can be used. But the convergence of $c_n$ to $\log Z$ is very sensitive to the choice of the sequence $\{\gamma_n\}$. For instance, if the $\gamma_n$'s are overly small, the recursive equation in (2.4) will make very small steps and thus converge very slowly or not at all. But if these numbers are overly large, the algorithm will have a large variance. In both cases, the convergence to the solution will be slow or even problematic. Overall, it is a well-known issue that choosing the right step-size for a stochastic approximation algorithm is a difficult problem. Here we follow Atchade and Liu (2010), which have elaborated on a heuristic approach to this problem originally proposed by Wang and Landau (2001).

The main idea of this approach is that, typically, the larger $\gamma_n$ is, the easier it is for the algorithm to move around the state space (as in tempering). Therefore, when starting the algorithm, $\gamma_0$ is set at a relatively large value. The sequence $\gamma_n$ is kept constant until $\{I_n\}$ has visited equally well all the values in $\{1, \dots, d\}$. Let $\tau_1$ be the first time where the occupation measure of $\{1, \dots, d\}$ by $\{I_n\}$ is approximately uniform. Then $\gamma_{\tau_1+1}$ is set to a smaller value (e.g., $\gamma_{\tau_1+1} = \gamma_{\tau_1}/2$) and the process is iterated until $\gamma_n$ become sufficiently small. At which point, we choose to switch to a deterministic sequence of the form $\gamma_n = n^{-1/2-\varepsilon}$. Combining this idea with Algorithm 2.2, we get the following straightforward implementation, where $\gamma > \varepsilon_1 > 0$, $\varepsilon_2 > 0$ are constants to be calibrated:

**Algorithm 2.3.** *At time* 0, *set* $(X_0, I_0, c_0, \theta_0)$ *as the arbitrary initial state of the algorithm. Set* $v = 0 \in \mathbb{R}^d$.

*While* $\gamma > \varepsilon_1$,

1. *Generate* $(X_{n+1}, I_{n+1}, c_{n+1}, \theta_{n+1})$ *as in Algorithm* 2.2.
2. *For* $i = 1, \ldots, d$: *set* $v(i) = v(i) + \mathbf{1}_i(I_{n+1})$.
3. *If* $\max_i |v(i) - \frac{1}{d}| \le \frac{\varepsilon_2}{d}$, *then set* $\gamma = \gamma/2$ *and* $v = 0 \in \mathbb{R}^d$.

**Remark 2.3.** In the application section below, we use this algorithm with the following specifications: we set the initial $\gamma$ to 1, $\varepsilon_1 = 0.001$, $\varepsilon_2 = 0.2$ and the final deterministic sequence is $\gamma_n = \varepsilon_1/n^{0.7}$.

## 3 Convergence analysis

In this section, we characterize the asymptotics of Algorithm 2.2. We use the filtration $\mathcal{F}_n = \sigma\{(X_k, I_k, c_k, \theta_k), k \le n\}$ and recall that $\Theta$ is a compact space equipped with its Borel $\sigma$-algebra $\mathcal{B}(\Theta)$. We also assume the following constraints:

(B1) The prior density $\mu$ is positive and continuous and there exist $m, M \in \mathbb{R}$ such that

$$m \le E(x, \theta) \le M, \qquad x \in \mathcal{X}, \theta \in \Theta. \tag{3.1}$$

(B2) The sequence $\{\gamma_n\}$ is a random sequence adapted to $\{\mathcal{F}_n\}$ that satisfies $\gamma_n > 0$, $\sum \gamma_n = \infty$ and $\sum \gamma_n^2 < \infty$ Pr-a.s.

(B3) There exist $\varepsilon > 0$, an integer $n_0$, and a probability measure $\nu$ on $\mathcal{X}$ such that for any $i \in \{1, \ldots, d\}$, $x \in \mathcal{X}$, $A \in \mathcal{B}(\mathcal{X})$, $[T_i^{(\mathcal{X})}]^{n_0}(x, A) \ge \varepsilon \nu(A)$.

**Remark 3.1.** In many applications, and this is the case for the examples discussed below, $\mathcal{X}$ is a finite set (that is typically very large) and $\Theta$ is a compact set. In these cases, (B1) and (B3) are easily checked. [Note that the minorization condition (B3) amounts to uniform ergodicity of the kernel $T_i^{(\mathcal{X})}$.] These assumptions can be further relaxed, but the analysis of the algorithm would then require more elaborate techniques that are beyond the scope of this paper.

**Proposition 3.1.** *Assume* (B1)–(B3). *Then for any* $i \in \{1, \ldots, d\}$,

$$e^{c_n(i)} \left( \sum_{k=1}^{d} e^{c_n(k)} \right)^{-1} \xrightarrow{a.s.} CZ(\theta^{(i)})$$

*as* $n \to \infty$ *for some finite constant* $C$ *that does not depend on* $i$. *Moreover, for any bounded measurable function* $f : \mathcal{X} \times \{1, \ldots, d\} \to \mathbb{R}$,

$$n^{-1} \sum_{k=1}^{n} f(X_k, I_k) \xrightarrow{a.s.} \Lambda^\star(f)$$

*as* $n \to \infty$.

Under a few additional assumptions on the kernel $P_\zeta$, the conditions of Theorem 2.1 are met.

**Theorem 3.1.** *Consider Algorithm 2.2 and assume* (B1)–(B3) *hold. Take $P_\zeta$ as a random walk Metropolis kernel with invariant distribution $\pi_\zeta$ and a symmetric proposal kernel $q$ such that there exist $\varepsilon' > 0$ and an integer $n'_0 \geq 1$ such that $q^{n'_0}(\theta, \theta') \geq \varepsilon'$ uniformly over $\Theta$. Let $h : \Theta \to \mathbb{R}$ be a measurable bounded function. Then*

$$n^{-1} \sum_{k=1}^{n} h(\theta_k) \xrightarrow{a.s.} \pi_z(h)$$

*as $n \to \infty$.*

**Proof.** See Section 6.2. □

**Remark 3.2.** The uniform minorization assumption on $q$ is only imposed here as a practical way of checking (A3). It holds on all the examples considered below due to the compactness of $\Theta$. If $P_\zeta$ is not a random walk Metropolis kernel, that assumption should be adapted accordingly to obtain (A3).

## 4 Examples

### 4.1 Ising model

We first test our algorithm on the Ising model on a rectangular lattice. The energy function $E$ is

$$E(x) = \sum_{i=1}^{m} \sum_{j=1}^{n-1} x_{ij} x_{i,j+1} + \sum_{i=1}^{m-1} \sum_{j=1}^{n} x_{ij} x_{i+1,j}, \tag{4.1}$$

and $x_{i,j} \in \{1, -1\}$. In our implementation, $m = n = 64$ and we can generate the data $x_0$ from $e^{\theta E(x)}/Z(\theta)$ with $\theta = 0.4$ by perfect sampling through the Propp–Wilson algorithm [see, e.g., Møller et al. (2006)]. The prior used in this example is $\mu(\theta) = \mathbf{1}_{(0,3)}(\theta)$ and $d = 100$ points $\{\theta^{(i)}\}$ are generated using the stochastic approximation described in Section 2.3. As described in Section 2.4, we use the flat histogram approach in selecting $\{\gamma_n\}$ with an initial value $\gamma_0 = 1$, until $\gamma_n$ becomes smaller than 0.001. Then we start feeding the adaptive chain $\{\theta_n\}$ which is run for 10,000 iterations. In updating $\theta_n$, we use a random Walk Metropolis sampler with proposal distribution $\mathcal{U}(\theta_n - b, \theta_n + b)$ (with reflexion at the boundaries) for some $b > 0$. We adaptively update $b$ so as to reach an acceptance rate of 30% [see, e.g., Atchade (2006)]. We also discard the first 1999 simulations as a burn-in period. The results are plotted on Figure 1a. As far as we can judge from plots (b) and (c), the sampler appears to have converged to the posterior distribution $\pi$ in that the sequence is quite stable. The mixing rate of the algorithm as inferred from the autocorrelation graph in (d) seems fairly good. In addition, the algorithm yields an estimate of the partition function $\log Z(\theta)$ shown in (a) that can be reused in other sampling problems.

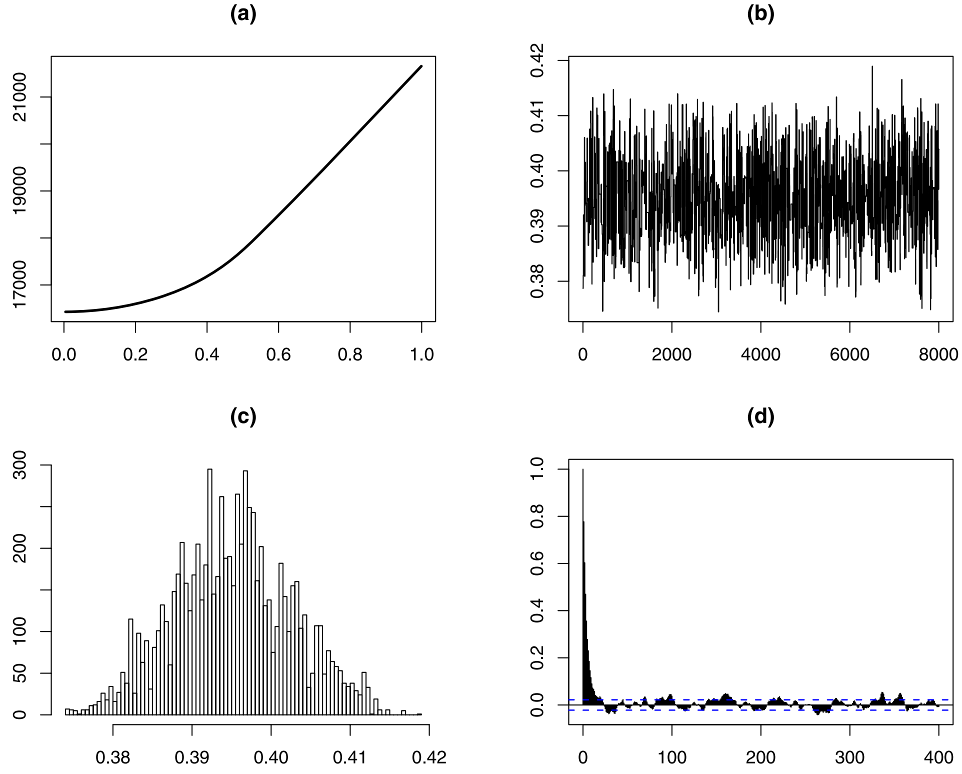

**Figure 1a** *Output for the Ising model* $\theta = 0.40$, $m = n = 64$. (a) *estimation of* $\log Z(\theta)$ *up to an additive constant*; (b)–(d) *trace plot*, *histogram and autocorrelation function of the adaptive sampler* $\{\theta_n\}$.

## 4.2 Comparison with the auxiliary variable method

We use the Ising model described above to compare the adaptive strategy of this paper with the auxiliary variable method (AVM) of Møller et al. (2006). We follow the description of the AVM method given in Murray et al. (2006). For the comparison we use $m = n = 20$. For the adaptive strategy, we use exactly the same sampler described above. For the AVM, the proposal kernel is a Random Walk proposal from $\mathcal{U}(x - b, x + b)$ with $b = 0.05$. We have run both samplers for 10,000 iterations and reached the following conclusions:

1. One limitation of the AVM that we have found is that the running time of the algorithm depends heavily on $b$ and the true value of the parameter. For larger values of $b$, large values of $\theta$ are more likely to be proposed and for those values, the time to coalescence in the Propp–Wilson perfect simulation algorithm can be significantly large.
2. Both samplers generate Markov chains with similar characteristics as assessed through a trace plot and an autocorrelation function. See Figure 1b.

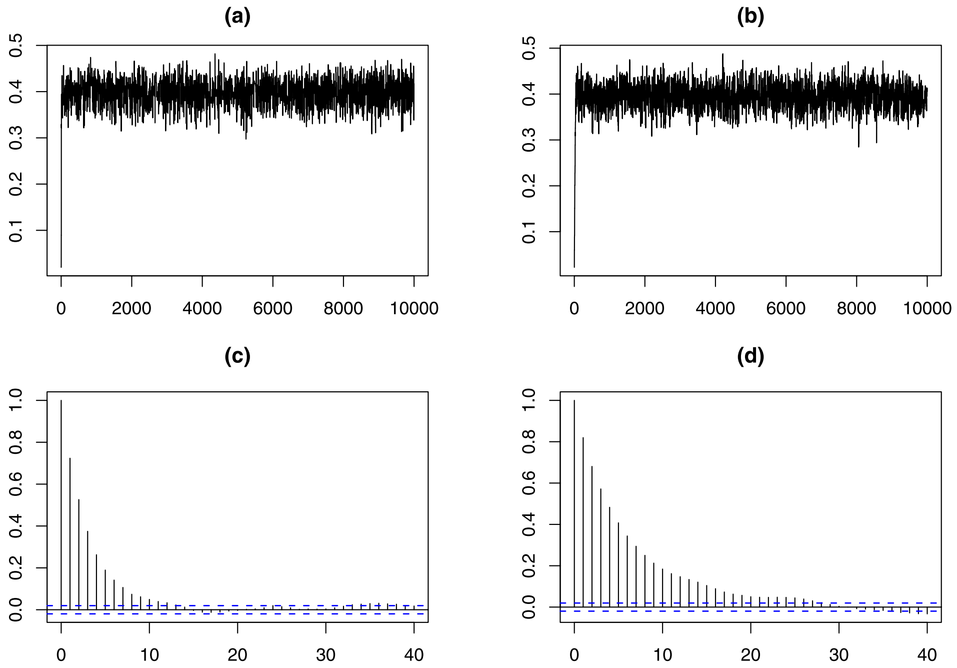

**Figure 1b** *Comparison with the AVM. Trace plot and autocorrelation function.* (a)–(c) *the adaptive sampler*; (b)–(d) *the AVM algorithm. Based on* 10,000 *iterations.*

3. The biggest difference between the two samplers is in terms of computing time. For this (relatively small) example the AVM took about 16 hours to run, whereas our adaptive MCMC took about 12 minutes, both on an IBM T60 Laptop.

### 4.3 An application to image segmentation

We use the above Ising model to illustrate an application of the methodology in image analysis [Ibanez and Simo (2003); Hurn et al. (2003)]. We represent the image by a vector $x = \{x_i, i \in \mathcal{S}\}$, where $\mathcal{S}$ is a $m \times n$ lattice and $x_i \in \{1, \dots, K\}$. Each $i \in \mathcal{S}$ represents a pixel, and $x_i$ is thus the color of the pixel $i$, with $K$ the number of colors. Here we assume that $K = 2$ and $x_i \in \{-1, 1\}$ is either black or white. In addition, we do not observe $x$ directly but through a noisy approximation $y$. We assume here that

$$y_i | x, \sigma^2 \overset{\text{ind}}{\sim} \mathcal{N}(x_i, \sigma^2), \tag{4.2}$$

for some unknown parameter $\sigma^2$. Even though (4.2) is a continuous model, it has been shown to provide a relatively good framework for image segmentation problems with multiple additive sources of noise [Ibanez and Simo (2003)].

As a standard assumption [see, e.g., Marin and Robert (2007)], we impose that the true image $x$ is generated from an Ising model with interaction parameter $\theta$. As in the previous section, $\theta$ follows a uniform prior distribution on $(0,3)$ and we assume in addition that $\sigma^2$ has an improper prior distribution that is proportional to $1/\sigma^2\mathbf{1}_{(0,\infty)}(\sigma^2)$. The posterior distribution $(\theta,\sigma^2,x)$ is then given by

$$\pi(\theta,\sigma^2,x|y)\propto\left(\frac{1}{\sigma^2}\right)^{|\mathcal{S}|/2+1}\frac{e^{\theta E(x)}}{Z(\theta)}e^{-1/(2\sigma^2)\sum_{s\in\mathcal{S}}(y(s)-x(s))^2}\mathbf{1}_{(0,3)}(\theta)\mathbf{1}_{(0,\infty)}(\sigma^2),$$

where $E$ is defined in (4.1).

We sample from this posterior distribution using the adaptive chain $\{(y_n,i_n,c_n,\theta_n,\sigma_n^2,x_n)\}$. The chain $\{(y_n,i_n,c_n)\}$ is updated following Steps 1–3 of Algorithm 2.2 and it provides the adaptive estimate of $Z(\theta)$ given by (2.5) (with $\{y_n,i_n\}$ replacing $\{X_n,I_n\}$). This sequence of estimates leads in its turn to update $(\theta_n,\sigma_n^2,x_n)$ using a Metropolis-within-Gibbs scheme. More specifically, given $(\sigma_n^2,x_n)$, we generate $\theta_{n+1}$ based on a random walk Metropolis step with proposal $\mathcal{U}(\theta_n-b,\theta_n+b)$ (with reflexion at the boundaries) and target proportional to $e^{\theta E(x_n)-\zeta_n(\theta)}$. Given $\theta_{n+1},x_n$, we generate $\sigma_{n+1}^2$ by sampling from the inverse Gamma distribution with parameters $(|\mathcal{S}|/2,\sum_{s\in\mathcal{S}}(y(s)-x(s))^2)/2$. At last, given $(\theta_{n+1},\sigma_{n+1})$, we sample each $x_{n+1}(s)$ from its full conditional distribution given $\{x(u),u\neq s\}$ and $y(s)$. This conditional distribution is given by

$$p\bigl(x(s)=a|x(u),u\neq s\bigr)\propto\exp\biggl(\theta a\sum_{u\sim s}x(u)-\frac{1}{2\sigma^2}(y(s)-a)^2\biggr),\qquad a\in\{-1,1\},$$

where $u\sim s$ in the summation means that the pixels $u$ and $s$ are neighbors.

To test our algorithm on this model, we have generated a simulated data set $y$ with $x$ generated from $e^{\theta E(x)}/Z(\theta)$ by perfect sampling. We use $m=n=64$, $\theta=0.40$ and $\sigma=0.5$. For the implementation details of the algorithm, we have made exactly the same calibration choices as in Example 4.1 above. In particular, we choose $d=100$ and generate $\{\theta^{(i)}\}$ using the stochastic approximation described in Section 2.3. The results are given in Figure 2. Once again, the sample path obtained for $\{\theta_n\}$ clearly suggests that the distribution of $\theta_n$ has converged to $\pi$ with a good mixing rate, as also inferred from the autocorrelation plots.

## 4.4 Social network modeling

We now give an application of the method to a Bayesian analysis of social networks. Statistical modeling of social networks is a growing subject in the social sciences [see, e.g., Robins et al. (2007) and the references therein for more details]. The setup is the following: given $n$ actors $I=\{1,\dots,n\}$, for each pair $(i,j)\in I\times I$, we define $y_{ij}=1$ if actor $i$ has ties with actor $j$ and $y_{ij}=0$ otherwise. In the example below, we only consider the case of a symmetric relationship where $y_{ij}=y_{ji}$ for all $i,j$. One of the most popular models for social networks

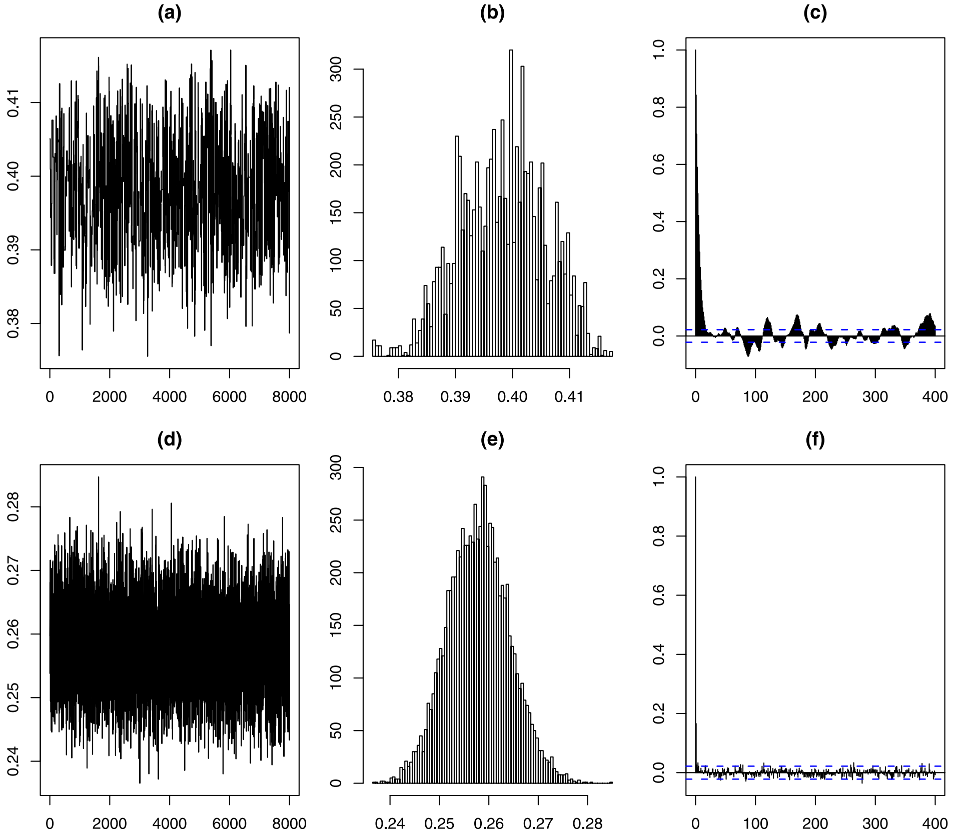

**Figure 2** *Output for the image segmentation model.* (a)–(c) *plots for* $\{\theta_n\}$; (d)–(f) *plots for* $\{\sigma_n^2\}$.

is the class of exponential random graph models. In these models, we assume that $\{y_{ij}\}$ is a sample generated from the parameterized distribution

$$p(y|\theta_1, \dots, \theta_K) \propto \exp\left(\sum_{i=1}^{K} \theta_i S_i(y)\right),$$

where $S_i(y)$ is a statistic used to capture some aspects of the network. For this example, and following Robins et al. (2007), we consider a 4-dimensional model with statistics

$$S_1(y) = \sum_{i<j} y_{ij}, \qquad \text{the total number of ties,}$$

$$S_2(y) = \sum_{i<j<k} y_{ik} y_{jk}, \qquad \text{the number of two-stars,}$$

$$S_3(y) = \sum_{i<j<k<l} y_{il} y_{jl} y_{kl}, \qquad \text{the number of three-stars,}$$

$$S_4(y) = \sum_{i<j<k} y_{ik} y_{jk} y_{ij}, \qquad \text{the number of transitive ties.}$$

We assume a uniform prior distribution on $D = (-50, 50)^4$ for $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$ and the corresponding posterior distribution is

$$\pi(\theta|y) \propto \frac{1}{Z(\theta)} \exp\left(\sum_{k=1}^{4} \theta_k S_k(y)\right) \mathbf{1}_D(\theta). \tag{4.3}$$

In this specific example, we study the Medici business network data set taken from Robins et al. (2007) which describes the business ties between 16 Florentine families. Numbering arbitrarily those families from 1 to 16, we plot the observed social network in Figure 3. The data set contains relatively few ties between families and even fewer transitive ties.

We use Algorithm 2.2 to sample from (4.3). For this example, we generate 400 particles $\{\theta^{(l)}\}$ using the stochastic approximation described in Section 2.3. We use the same parameterization as in the previous examples to update $(X_n, I_n, c_n)$. For the adaptive chain $\{\theta_n\}$ we use a slightly different strategy, though. It turns out that some of the components of the target distribution $\pi$ are strongly correlated. Therefore, we sample from $\pi$ in one block, using a random walk Metropolis algorithm with a Gaussian kernel $N(0, \sigma^2\Sigma)$ (restricted to $D$) for $\sigma > 0$ and a positive definite matrix $\Sigma$. We adaptively set $\sigma$ so as to reach the optimal acceptance rate of 30%. Ideally, we would like to choose $\Sigma$ equal to $\Sigma_\pi$ the variance–covariance matrix of $\pi$ which, of course, is not available. Instead, we adaptively estimate $\Sigma_\pi$ during the simulation as in Atchade (2006). As before, we run $(X_n, I_n, c_n)$ until $\gamma_n < 0.001$. Then we start $\{\theta_n\}$ and run the full chain $(X_n, I_n, c_n, \theta_n)$ for a total of 25,000 iterations. Figure 4 presents the output for parameter $\theta_4$. In Table 1, we summarize those graphs by providing the sample posterior mean together with the 2.5% and 97.5% quantiles of the marginal posterior distributions. Overall, these results are consistent with the maximum likelihood estimates obtained by Robins

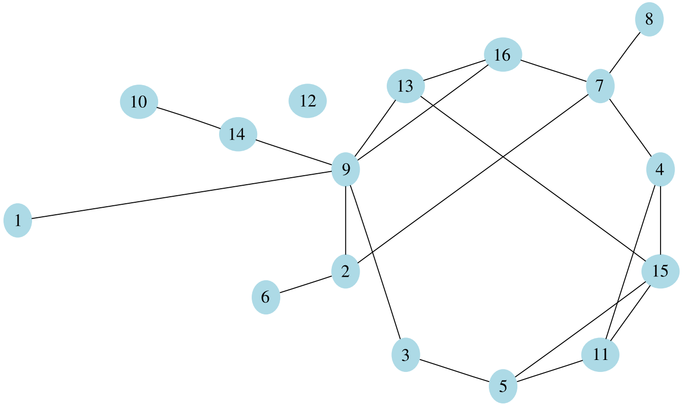

**Figure 3** *Business Relationships between* 16 *Florentine families.*

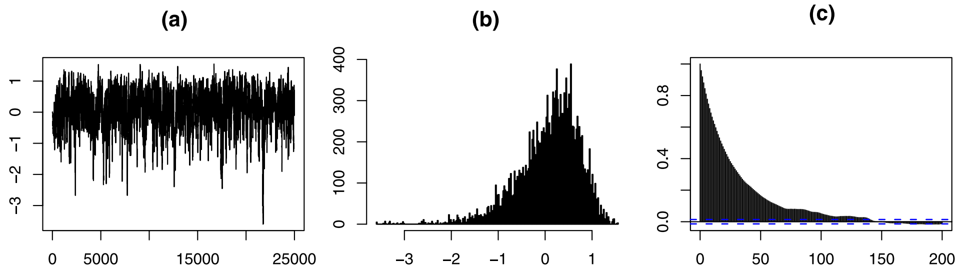

**Figure 4** *The adaptive MCMC output from* (4.3) *for* $\{\theta_4\}$. *Based on* 25,000 *iterations.*

**Table 1** *Summary of the posterior distribution of the parameters. Posterior means*, 2.5% *and* 97.5% *posterior quantiles*

| Parameters | Post. mean | Post. quantiles |
|---|---|---|
| $\theta_1$ | −2.05 | (−3.27, −0.77) |
| $\theta_2$ | 0.96 | (−0.40, 2.45) |
| $\theta_3$ | −1.10 | (−2.78, 0.08) |
| $\theta_4$ | 0.11 | (−1.41, 1.12) |

et al. (2007) using MCMC-MLE. The main difference appears in $\theta_4$ which we find here to be not significant. As a by-product, the sampler gives an estimate of the covariance matrix of the posterior distribution $\pi$:

$$\Sigma_\pi = \begin{bmatrix} 1.62 & -0.38 & 0.31 & -0.06 \\ -0.38 & 1.90 & -0.51 & -0.01 \\ 0.31 & -0.51 & 1.83 & -0.05 \\ -0.06 & -0.01 & -0.05 & 1.55 \end{bmatrix}.$$

## 5 Conclusion

Sampling from posterior distributions with intractable normalizing constants is a difficult computational problem. In this work, we attempt to address the problem in a certain generality, using adaptive Monte Carlo methods. The main advantage of the proposed method is that it does not require perfect sampling as in Møller et al. (2006). Even in cases where perfect sampling is possible, the proposed method can lead to computational gains over the exact sampling-based approach of Møller et al. (2006), as shown in Example 4.2.

An important question is knowing how well the proposed algorithm scales with dimensionality, more specifically, how $d$, the number of particles scales with the dimensionality of the parameter space $\Theta$ and the size of $\mathcal{X}$. Although we did not investigate this issue, it is likely that the method will suffer from the curse of dimensionality. The main limiting factor is the idea of estimating ratio of normalizing constants by importance sampling. Of course, how far the method can be pushed depends on the specific class of distributions under consideration. However, adaptations and improvements on the method can be imagined that can be automatically inserted in the proposed framework, for instance, using annealing/tempering schemes, or replacing importance sampling by alternative techniques such as path sampling.

## 6 Proof

### 6.1 Proof of Theorem 2.1

**Proof.** Throughout the proof, $C$ will denote a finite constant but whose actual value can change from one equation to the next. Also, and without any loss of generality, we will assume that $\theta_k$ is $\mathcal{F}_k$-measurable. For $\zeta \in \mathcal{C}_+(\Theta)$, define $\bar{P}_\zeta = P_\zeta - \pi_\zeta$. Fix $h:\Theta \to \mathbb{R}$ a bounded measurable function. For $\theta \in \Theta$ and $k \geq 0$, define $g_{\zeta_k}(\theta) = \sum_{j=0}^{\infty} \bar{P}_{\zeta_k} h(\theta)$. By (A3), $g_{\zeta_k}(\theta)$ is well defined and $|g_{\zeta_k}(\theta)| \leq (1-\rho)^{-1}|h|_\infty$, Pr-a.s. Moreover, $g_{\zeta_k}$ satisfies the equation

$$h(\theta) - \pi_{\zeta_k}(h) = g_{\zeta_k}(\theta) - P_{\zeta_k} g_{\zeta_k}(\theta). \tag{6.1}$$

For any $n, k \geq 0$, we have $\bar{P}^n_{\zeta_k} - \bar{P}^n_{\zeta_{k-1}} = \sum_{j=1}^{n} \bar{P}^{n-j}_{\zeta_k}(\bar{P}_{\zeta_k} - \bar{P}_{\zeta_{k-1}})\bar{P}^{j-1}_{\zeta_{k-1}}$, Pr-a.s. We deduce from this and (A2) that $|\bar{P}^n_{\zeta_k} - \bar{P}^n_{\zeta_{k-1}}| \leq |h|_\infty \sup_{\theta\in\Theta} \|\bar{P}_{\zeta_k}(\theta,\cdot) - \bar{P}_{\zeta_{k-1}}(\theta,\cdot)\|_{\mathrm{TV}} n\rho^{n-1}$ which implies that

$$|g_{\zeta_k}(\theta) - g_{\zeta_{k-1}}(\theta)| \leq \frac{|h|_\infty}{(1-\rho)^2} \sup_{\theta\in\Theta} \|\bar{P}_{\zeta_k}(\theta,\cdot) - \bar{P}_{\zeta_{k-1}}(\theta,\cdot)\|_{\mathrm{TV}}, \qquad \text{Pr-a.s.} \tag{6.2}$$

By the triangular inequality,

$$\begin{aligned} \|\bar{P}_{\zeta_k}(\theta,\cdot) - \bar{P}_{\zeta_{k-1}}(\theta,\cdot)\|_{\mathrm{TV}} &\leq \|P_{\zeta_k}(\theta,\cdot) - P_{\zeta_{k-1}}(\theta,\cdot)\|_{\mathrm{TV}} \\ &\quad + \|\pi_{\zeta_k}(\cdot) - \pi_{\zeta_{k-1}}(\cdot)\|_{\mathrm{TV}}, \qquad \text{Pr-a.s.} \end{aligned} \tag{6.3}$$

Now, for any measurable function $f:\Theta \to \mathbb{R}$ such that $|f| \leq 1$ and for any $n \geq 0$,

$$\begin{aligned} &\pi_{\zeta_k}(f) - \pi_{\zeta_{k-1}}(f) \\ &\quad = \pi_{\zeta_k}[P^n_{\zeta_k}(f - \pi_{\zeta_{k-1}}(f))] \\ &\quad = \pi_{\zeta_k}[P^n_{\zeta_{k-1}}(f - \pi_{\zeta_{k-1}}(f)) + P^n_{\zeta_k}(f - \pi_{\zeta_{k-1}}(f)) - P^n_{\zeta_{k-1}}(f - \pi_{\zeta_{k-1}}(f))] \\ &\quad = \pi_{\zeta_k}\left[P^n_{\zeta_{k-1}}(f - \pi_{\zeta_{k-1}}(f)) + \sum_{j=1}^{n} P^{n-j}_{\zeta_k}(P_{\zeta_k} - P_{\zeta_{k-1}})P^{j-1}_{\zeta_{k-1}}(f - \pi_{\zeta_{k-1}}(f))\right]. \end{aligned}$$

Using (A2)–(A3) and letting $n \to \infty$, it follows that

$$\begin{aligned}\|\pi_{\zeta_k} - \pi_{\zeta_{k-1}}\|_{\mathrm{TV}} &\le 2(1-\rho)^{-1} \sup_{\theta\in\Theta} \|P_{\zeta_k}(\theta,\cdot) - P_{\zeta_{k-1}}(\theta,\cdot)\|_{\mathrm{TV}} \\ &\to 0, \qquad \mathrm{Pr}\text{-a.s. as } k \to \infty.\end{aligned} \tag{6.4}$$

Using this and (6.3) in (6.2), we can therefore conclude that

$$\sup_{\theta\in\Theta} |g_{\zeta_k}(\theta) - g_{\theta_{k-1}}(\theta)| \to 0, \qquad \mathrm{Pr}\text{-a.s. as } k \to \infty. \tag{6.5}$$

Define $S_n(h) = \sum_{k=1}^n h(\theta_k) - \pi_{\zeta_{k-1}}(h)$. Using (6.1), we can rewrite $S_n(h)$ as

$$\begin{aligned}S_n(h) &= \sum_{k=1}^n h(\theta_k) - \pi_{\zeta_{k-1}}(h) = \sum_{k=1}^n g_{\zeta_{k-1}}(\theta_k) - P_{\zeta_{k-1}} g_{\zeta_{k-1}}(\theta_k) \\ &= \sum_{k=1}^n g_{\zeta_{k-1}}(\theta_k) - P_{\zeta_{k-1}} g_{\zeta_{k-1}}(\theta_{k-1}) + \bigl(P_{\zeta_0} g_{\zeta_0}(\theta_0) - P_{\zeta_n} g_{\zeta_n}(\theta_n)\bigr) \\ &\quad + \left(\sum_{k=1}^n \bigl(g_{\zeta_k}(\theta_k) - g_{\zeta_{k-1}}(\theta_k)\bigr)\right) + \left(\sum_{k=1}^n \bigl(\pi_{\zeta_k}(h) - \pi_{\zeta_{k-1}}(h)\bigr)\right).\end{aligned}$$

The term $\sum_{k=1}^n g_{\zeta_{k-1}}(\theta_k) - P_{\zeta_{k-1}} g_{\zeta_{k-1}}(\theta_{k-1})$ is a $\{\mathcal{F}_n\}$-martingale with bounded increment. From martingale limit theory, we conclude that $n^{-1}\sum_{k=1}^n g_{\zeta_{k-1}}(\theta_k) - P_{\zeta_{k-1}} g_{\zeta_{k-1}}(\theta_{k-1})$ converges to zero, Pr-a.s. as $n \to \infty$. Again, since $g_{\zeta_k}$ are uniformly bounded, $n^{-1}(P_{\zeta_0} g_{\zeta_0}(\theta_0) - P_{\zeta_n} g_{\zeta_n}(\theta_n))$ converges to zero as $n$ goes to infinity. We conclude from (6.4) and (6.5) that the last two terms also converge to zero. In conclusion, $n^{-1}S_n(h)$ converges a.s. to zero.

Now, since $n^{-1}\sum_{k=1}^n h(\theta_k) = n^{-1}S_n(h) + n^{-1}\sum_{k=1}^n \pi_{\zeta_k}(h)$, and since by (A1), $n^{-1}\sum_{k=1}^n \pi_{\zeta_k}(h) \to \pi_Z(h)$ Pr-almost surely, we conclude that the limit of $n^{-1}\sum_{k=1}^n h(\theta_k)$ as $n$ goes to infinity is also $\pi_Z(h)$, Pr-a.s, which concludes the proof. □

## 6.2 Proof of Theorem 3.1

**Proof.** We will show that (A1)–(A3) hold and then apply Proposition 2.1. Define $\tilde{\zeta}_n(\theta) = \zeta_n(\theta)(\sum_{i=1}^d e^{c_n(i)})^{-1}$. Using Proposition 3.1, we see that

$$\begin{aligned}\tilde{\zeta}_n(\theta) &= \sum_{i=1}^d \kappa(\theta,\theta^{(i)}) \frac{e^{c_n(i)}}{\sum_{l=1}^d e^{c_n(l)}} \frac{\sum_{k=1}^n e^{E(X_k,\theta)-E(X_k,\theta^{(i)})}\mathbf{1}_i(I_k)}{\sum_{k=1}^n \mathbf{1}_i(I_k)} \\ &\to \sum_{i=1}^n \kappa(\theta,\theta^{(i)}) C Z(\theta^{(i)}) \frac{Z(\theta)Z(\theta^{(i)})^{-1}}{d^{-1}} \\ &= \frac{C}{d} Z(\theta),\end{aligned}$$

with probability one as $n \to \infty$. Furthermore, using (B1), we deduce easily that

$$\inf_{\theta,\theta'\in\Theta} \kappa(\theta,\theta')d^{-1}e^{m-M} \le \sup_{\theta\in\Theta} \tilde{\zeta}_n(\theta) \le \sup_{\theta,\theta'\in\Theta} \kappa(\theta,\theta')e^{M-m}. \tag{6.6}$$

It follows that for any bounded measurable function $h:\Theta \to \mathbb{R}$,

$$\pi_{\zeta_n}(h) = \pi_{\tilde{\zeta}_n}(h) = \frac{\int_\Theta e^{E(x_0,\theta)}\tilde{\zeta}_n^{-1}(\theta)h(\theta)\,d\theta}{\int_\Theta e^{E(x_0,\theta)}\tilde{\zeta}_n^{-1}(\theta)\,d\theta} \to \pi_Z(h),$$

as $n \to \infty$, by the Lebesgue dominated convergence theorem. (A1) is proved.

Let $f:\Theta \to \mathbb{R}$ be a measurable function such that $|f| \le 1$. For any $n \ge 1$, $\theta \in \Theta$, we have

$$\begin{aligned}
&|P_{\tilde{\zeta}_n} f(\theta) - P_{\tilde{\zeta}_{n-1}} f(\theta)| \\
&\quad = \biggl| \int \biggl[ \min\biggl(1, \frac{\tilde{\zeta}_n(\theta)e^{E(x_0,\theta')}\mu(\theta')}{\tilde{\zeta}_n(\theta')e^{E(x_0,\theta)}\mu(\theta)}\biggr) - \min\biggl(1, \frac{\tilde{\zeta}_{n-1}(\theta)e^{E(x_0,\theta')}\mu(\theta')}{\tilde{\zeta}_{n-1}(\theta')e^{E(x_0,\theta)}\mu(\theta)}\biggr)\biggr] \\
&\qquad\quad \times f(\theta')q(\theta,\theta')\,d\theta' \biggr| \\
&\quad \le \int \biggl| \frac{\tilde{\zeta}_n(\theta)}{\tilde{\zeta}_n(\theta')} - \frac{\tilde{\zeta}_{n-1}(\theta)}{\tilde{\zeta}_{n-1}(\theta')} \biggr| \frac{e^{E(x_0,\theta')}\mu(\theta')}{e^{E(x_0,\theta)}\mu(\theta)} q(\theta,\theta')\,d\theta' \\
&\quad \le C \sup_{\theta,\theta'\in\Theta} |\tilde{\zeta}_n(\theta) - \tilde{\zeta}_{n-1}(\theta)|
\end{aligned}$$

for some finite constant $C$. For the first inequality we use $|f| \le 1$ and the fact that for any $a, x, y \ge 0$, $|\min(1, ax) - \min(1, ay)| \le a|x - y|$, whereas in the last inequality, we use (B1) and (6.6).

We now bound the term $\sup_{\theta,\theta'\in\Theta} |\tilde{\zeta}_n(\theta) - \tilde{\zeta}_{n-1}(\theta)|$. Let $w_n(i) = e^{c_n(i)} \times (\sum_{l=1}^d e^{c_n(l)})^{-1}$, $v_n(i) = \sum_{k=1}^n e^{E(X_k,\theta)-E(X_k,\theta^{(i)})}\mathbf{1}_i(I_k)(\sum_{k=1}^n \mathbf{1}_i(I_k))^{-1}$, so that $\tilde{\zeta}_n(\theta) = \sum_{i=1}^d \kappa(\theta,\theta^{(i)})w_n(i)v_n(i)$. We have

$$\begin{aligned}
|\tilde{\zeta}_n(\theta) - \tilde{\zeta}_{n-1}(\theta)| &\le \biggl| \sum_{i=1}^d \kappa(\theta,\theta^{(i)})\bigl(w_n(i) - w_{n-1}(i)\bigr)v_n(i) \biggr| \\
&\quad + \biggl| \sum_{i=1}^d \kappa(\theta,\theta^{(i)})w_{n-1}(i)\bigl(v_n(i) - v_{n-1}(i)\bigr) \biggr|.
\end{aligned}$$

The term $w_n(i)$ satisfies $|w_n(i) - w_{n-1}(i)| \le 2e^{\gamma_0}\gamma_n$, whereas the term $v_n(i)$ satisfies $|v_n(i) - v_{n-1}(i)| \le e^{M-m}(\sum_{l=1}^n \mathbf{1}_i(I_l))^{-1}$. It follows that

$$\sup_{\theta,\theta'\in\Theta} |\tilde{\zeta}_n(\theta) - \tilde{\zeta}_{n-1}(\theta)| \le C\biggl(\gamma_n + \frac{1}{\sum_{l=1}^n \mathbf{1}_i(I_l)}\biggr).$$

(A2) follows.

For $n \geq 1$, $\theta \in \Theta$ and $A \in \mathcal{B}(\Theta)$,

$$P_{\zeta_n}(\theta, A) \geq \int_A \min\left(1, \frac{\zeta_n(\theta) e^{E(x_0,\theta')} \mu(\theta')}{\zeta_n(\theta') e^{E(x_0,\theta)} \mu(\theta)}\right) q(\theta, \theta')\, d\theta'$$

$$\geq \left[\inf_{\theta,\theta' \in \Theta} \frac{\zeta_n(\theta) e^{E(x_0,\theta')} \mu(\theta')}{\zeta_n(\theta') e^{E(x_0,\theta)} \mu(\theta)}\right] q(\theta, A) \geq \delta q(\theta, A)$$

for some $\delta > 0$, using (B1) and (6.6). The constant $\delta$ does not depend on $n$ or $\theta$. Therefore, if $q^{n_0}(\theta, \theta') \geq \varepsilon$, then for any $n \geq 1$, $\theta \in \Theta$, $\|P^j_{\tilde{\zeta}_n}(\theta, \cdot) - \pi_{\tilde{\zeta}_n}(\cdot)\|_{\mathrm{TV}} \leq (1-\delta)^{j/n_0}$, which implies (A3). □

## Acknowledgments

We are grateful to the referees and to Mahlet Tadesse for very helpful comments that have improved the quality of this work.

## References

Andrieu, C., Moulines, É. and Priouret, P. (2005). Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim.* **44**, 283–312 (electronic). MR2177157

Atchade, Y. F. (2006). An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift. *Methodol. Comput. Appl. Probab.* **8**, 235–254. MR2324873

Atchade, Y. F. and Liu, J. S. (2010). The Wang–Landau algorithm for Monte Carlo computation in general state spaces. *Statistica Sinica* **20**, 209–233. MR2640691

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36**, 192–236. With discussion by D. R. Cox, A. G. Hawkes, P. Clifford, P. Whittle, K. Ord, R. Mead, J. M. Hammersley and M. S. Bartlett and with a reply by the author. MR0373208

Cappé, O., Douc, R., Guillin, A., Marin, J.-M. and Robert, C. (2008). Adaptive importance sampling in general mixture classes. *Statist. Comput.* **18**, 447–459. MR2461888

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statist. Sci.* **13**, 163–185. MR1647507

Geyer, C. J. (1994). On the convergence of Monte Carlo maximum likelihood calculations. *J. Roy. Statist. Soc. Ser. B* **56**, 261–274. MR1257812

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *J. Roy. Statist. Soc. Ser. B* **54**, 657–699. With discussion and a reply by the authors. MR1185217

Hurn, M., Husby, O. and Rue, H. (2003). A tutorial on image analysis. In *Spatial Statistics and Computational Methods (Aalborg, 2001). Lecture Notes in Statistics* **173**, 87–141. New York: Springer. MR2001386

Ibanez, M. V. and Simo, A. (2003). Parameter estimation in markov random field image modeling with imperfect observations. a comparative study. *Pattern Recognition Letters* **24**, 2377–2389.

Kleinman, A., Rodrigue, N., Bonnard, C. and Philippe, H. (2006). A maximum likelihood framework for protein design. *BMC Bioinformatics* **7**, 20–39.

Marin, J.-M. and Robert, C. (2007). *Bayesian Core*. New York: Springer Verlag. MR2289769

Marin, J.-M., Robert, C. and Titterington, D. (2009). A Bayesian reassessment of nearest-neighbour classification. *J. Amer. Statist. Assoc.* **104**, 263–273. MR2663042

Møller, J., Pettitt, A. N., Reeves, R. and Berthelsen, K. K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika* **93**, 451–458. MR2278096

Møller, J. and Waagepetersen, R. P. (2004). *Statistical Inference and Simulation for Spatial Point Processes*. *Monographs on Statistics and Applied Probability* **100**. Boca Raton, FL: Chapman & Hall/CRC. MR2004226

Murray, I., Ghahramani, Z. and MacKay, D. (2006). MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. Corvallis, OR: AUAI Press.

Robins, G., P., P., Kalish, Y. and Lusher, D. (2007). An introduction to exponential random graph models for social networks. *Social Networks* **29**, 173–191.

Wang, F. and Landau, D. P. (2001). Efficient, multiple-range random walk algorithm to calculate the density of states. *Physical Review Letters* **86**, 2050–2053.

Younes, L. (1988). Estimation and annealing for gibbsian fields. *Annales de l'Institut Henri Poincaré. Probabilité et Statistiques* **24**, 269–294. MR0953120

Y. F. Atchadé
Department of Statistics
University of Michigan
Ann Arbor, Michigan 48109-1107
USA
E-mail: yvesa@umich.edu

N. Lartillot
Dept. de Biologie
Université de Montreal
Montreal, Quebec
Canada
E-mail: nicolas.lartillot@umontreal.ca

C. Robert
Université Paris-Dauphine and CREST, INSEE
Paris
France
E-mail: xian@ceremade.dauphine.fr