Optimal scaling of the independence sampler: Theory and Practice

Clement Lee (Lancaster University – currently at Newcastle University)
and Peter Neal* (Lancaster University)

October 3, 2016

Correspondence address: Department of Mathematics and Statistics, Fylde College, Lancaster University, Lancaster, LA1 4YF, UK

Running title: Optimal scaling of the independence sampler

Abstract

The independence sampler is one of the most commonly used MCMC algorithms usually as a component of a Metropolis-within-Gibbs algorithm. The common focus for the independence sampler is on the choice of proposal distribution to obtain an as high as possible acceptance rate. In this paper we have a somewhat different focus concentrating on the use of the independence sampler for updating augmented data in a Bayesian framework where a natural proposal distribution for the independence sampler exists. Thus we concentrate on the proportion of the augmented data to update to optimise the independence sampler. Generic guidelines for optimising the independence sampler are obtained for independent and identically distributed product densities mirroring findings for the random walk Metropolis algorithm. The generic guidelines are shown to be informative beyond the narrow confines of idealised product densities in two epidemic examples.

Keywords: Augmented data; Birth-Death-Mutation model; Markov jump process; MCMC; SIR epidemic model.

1 Introduction

The independence sampler is the incorporation of rejection sampling within an MCMC framework. The rejection sampler obtains samples from a ran-

dom variable, X, with probability density function $f(\cdot)$ by first proposing a candidate value y from a random variable, Y, with probability density function $q(\cdot)$, and secondly accepting y as a sample from X with probability $f(y)/\{Kq(y)\}$, where $K = \sup_x f(x)/q(x)$. Otherwise y is rejected, see Ripley (1987), page 60. The success of the rejection sampler depends upon making a good choice of $q(\cdot)$ such that $K(\geq 1)$ is small and that $q(\cdot)$ is straightforward to sample from. The MCMC independence sampler is the modification of the above where a Markov chain $X_0, X_1, \ldots$ is constructed with at iteration t, a candidate y proposed from Y and if accepted X_t is set equal to y. Otherwise $X_t = X_{t-1}$. The rejection sampler, and consequently, the independence sampler can usually be implemented in a straightforward and efficient manner for low dimensional (target) distributions but as the dimension of X increases it becomes increasingly more challenging to obtain a good choice of $q(\cdot)$. Therefore the independence sampler is rarely used as an MCMC algorithm in its own right but instead independence sampler moves are often incorporated within Metropolis-within-Gibbs to effectively update low dimensional subsets of X, see Dellaportas and Roberts (2013), page 15.

The main focus for independence samplers has been to choose the proposal density $q(\cdot)$ so as to have an acceptance probability as close to 1 as possible. Whilst this makes intuitive sense, the aim of the current paper is to challenge the idea of aiming for an acceptance probability as close to 1 as possible within the context of using independence samplers for updating augmented data in MCMC algorithms. Specifically, we are interested in the Bayesian statistical problem of obtaining samples from the posterior distribution of the parameters $\boldsymbol{\theta}$ of a model given data $\mathbf{x}$, $\pi(\boldsymbol{\theta}|\mathbf{x})$ in the case where the likelihood, $\pi(\mathbf{x}|\boldsymbol{\theta})$ is intractable. We assume that given augmented data $\mathbf{y}$, $\pi(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta})$ is tractable and an MCMC algorithm can be constructed to obtain samples from the joint posterior of $\boldsymbol{\theta}$ and $\mathbf{y}$, $\pi(\boldsymbol{\theta}, \mathbf{y}|\mathbf{x})$. Then it is natural to construct an MCMC algorithm which alternates between updating the parameters and the augmented data as follows:

1. Update $\boldsymbol{\theta}$ given $\mathbf{x}$ and $\mathbf{y}$. *i.e.* Use $\pi(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y})$.
2. Update $\mathbf{y}$ given $\mathbf{x}$ and $\boldsymbol{\theta}$. *i.e.* Use $\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$.

Our focus is the use of independence samplers to update $\mathbf{y}$ given $\mathbf{x}$ and $\boldsymbol{\theta}$. For updating augmented data a natural independence sampler often presents itself. For example, in an epidemic modelling context where $\mathbf{x}$ denotes the removal times of infected individuals, $\boldsymbol{\theta}$ denotes the infection and infectious period parameters and $\mathbf{y}$ denotes the infection times of individuals, a natural candidate for the infection time of individual i who is removed at time x_i is $y_i = x_i - D$, where D denotes the infectious period distribution, see Neal and Roberts (2005), Xiang and Neal (2014) and Section 3.2. For non-centered

parameterisations, Papaspoliopoulos *et al.* (2003), we can often denote $\mathbf{Y}$ as a deterministic function $h(\boldsymbol{\theta}, \mathbf{U})$ with $\pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ easy to compute, where $\mathbf{U}$ is a vector of independent and identically distributed uniform random variables, see Neal and Huang (2015) and Section 3.3. Then to update U_i we can propose a new value from $U(0, 1)$. The dimension of the augmented data, $\mathbf{y}$, can be orders of magnitude higher than $\boldsymbol{\theta}$ and $\mathbf{x}$, so updating one component of $\mathbf{y}$ at a time can be prohibitive. Therefore we seek generic guidelines for updating multiple components of $\mathbf{y}$ at a time and optimising the performance of the resulting independent sampler. Specifically, this work formalises findings in Xiang and Neal (2014) and Neal and Huang (2015) in using the independence sampler for data augmentation giving simple guidelines for producing close to optimal independence samplers. The guidelines obtained are similar to those given in Roberts *et al.* (1997) for the random walk Metropolis algorithm and comparisons with the random walk Metropolis algorithm are made.

The paper is structured as follows. In Section 2, we study the properties of the independence sampler for independent and identically distributed product densities $\pi(\mathbf{x}) = \prod_{i=1}^{n} f(x_i)$. This idealised scenario mimics the set up in Roberts *et al.* (1997) where optimal scaling of the random walk Metropolis algorithm was first explored and as in Roberts *et al.* (1997) allows us to get a handle on understanding the key factors in optimising the independence sampler. In particular, we show that the optimal number of components, k, of $\mathbf{x}$ to update, is the k which maximises the mean number of components per move. In the case where this optimal k is large this corresponds to a mean acceptance rate of approximately 23.4%. Thus there is a somewhat surprising link with the optimal scaling of the random walk Metropolis algorithm, Roberts *et al.* (1997) with which we make comparison and highlight the benefits of the independence sampler. In Section 3, we explore the optimal performance of the independence sampler for increasingly complex problems. In Section 3.1, we study product Gaussian target densities with Gaussian and t-distribution proposals demonstrating the optimal scaling results obtained in Section 2. In Sections 3.2 and 3.3 we apply the independence sampler to two epidemic models, the classic homogeneously mixing SIR epidemic model, Bailey (1975) and O'Neill and Roberts (1999) and a birth-death-mutation (BDM) model for an emerging, evolving disease, Tanaka *et al.* (2006) and Fearnhead and Prangle (2012). In Section 3.2, we show that for the homogeneously mixing SIR epidemic model updating a proportion of the infection times so as to obtain a mean acceptance rate of approximately 23.4% is optimal. This demonstrates that as observed with the random walk Metropolis algorithm the findings of Section 2 are informative in designing independence samplers beyond the limited confines of product densities. For the BDM model in Section 3.3 the findings are somewhat different with a lower optimal mean acceptance rate correspond-

ing to large scale data augmentation. Finally, in Section 4, we make some concluding remarks highlighting the possible benefits of the independence sampler over random walk Metropolis for large scale data augmentation and the differences seen between the two epidemic models in Sections 3.2 and 3.3.

2 Theoretical properties of the independent sampler

In this Section we consider the theoretical properties of the independence sampler for the special case where $\pi_n(\mathbf{x}^n) = \prod_{i=1}^n f(x_i)$, a product of independent and identically distributed univariate densities, $f(x)$. The main focus is on the asymptotic behaviour as the number of components, $n \to \infty$ mirroring analysis performed in Roberts *et al.* (1997) for the random walk Metropolis algorithm. The aim is to characterise the optimal performance of the independence sampler in terms of the number of components to update and to draw interesting comparisons of similarities and differences with the random walk Metropolis algorithm.

For the independence sampler we propose to select uniformly at random k components $\{I_1, I_2, \ldots, I_k\}$ from $\{1, 2, \ldots, n\}$ to update. For $j \in \{I_1, I_2, \ldots, I_k\}$, y_j is drawn from Y with probability density function $q(y)$, whilst for $l \notin \{I_1, I_2, \ldots, I_k\}$, $y_l = x_l$. Therefore the acceptance probability for the proposed move from $\mathbf{x}^n$ to $\mathbf{y}^n$ is

$$\min\left\{1, \frac{\pi_n(\mathbf{y}^n)}{\pi_n(\mathbf{x}^n)} \times \frac{q(\mathbf{y}^n \to \mathbf{x}^n)}{q(\mathbf{x}^n \to \mathbf{y}^n)}\right\} = \min\left\{1, \prod_{j=1}^{k} \frac{f(y_{I_j})/q(y_{I_j})}{f(x_{I_j})/q(x_{I_j})}\right\}. \quad (2.1)$$

For $n = 1, 2, \ldots$ and $t = 0, 1, \ldots$, let $\mathbf{X}_t^n = (X_{t,1}^n, X_{t,2}^n, \ldots, X_{t,n}^n)$ denote the position of the Markov chain after t iterations. As in Roberts *et al.* (1997), we assume that the Markov chain is initiated with $\mathbf{X}_0^n$ drawn from $\pi_n(\cdot)$ and thus for all $t \geq 0$, $\mathbf{X}_t^n \sim \pi_n(\cdot)$. The independent and identically distributed nature of the stationary and proposal distributions means that as in Roberts *et al.* (1997) it suffices to focus on the behaviour and performance of the independence sampler on the first component only. Specifically, for $t \geq 0$, letting $\mathbf{Z}_t^n = \mathbf{X}_{[nt]}^n$ we show that for fixed k, as $n \to \infty$, the movement in the first component of $\mathbf{Z}_t^n$ converges to a Markov jump process with jumps governed by $f(\cdot)$ and $q(\cdot)$.

Let $\omega(x) = f(x)/q(x)$, then for the independence sampler to be well-behaved we require that $\sup_x \omega(x) < \infty$, see Tiernay (1994) and we make this assumption throughout. For a move to occur in the first component we must

propose to move the first component and $k-1$ other components from $\{2,3,\ldots,n\}$. Let $\{J_1,J_2,\ldots,J_{k-1}\}$ be a random sample from $\{2,3,\ldots,n\}$ with $W_{k-1}(\mathbf{x}^{n-}) = \prod_{i=1}^{k-1}\omega(Y_{J_i})/\omega(x_{J_i})$, where $\mathbf{x}^{n-} = (x_2,x_3,\ldots x_n)$. Define $\mathbf{Y}^{n-}$, $\mathbf{X}^{n-}$ and $\mathbf{y}^{n-}$ in the obvious fashion. Then we define

$$\begin{aligned} H(y,\mathbf{x}^n) &= H(y,x_1,\mathbf{x}^{n-}) \\ &= \mathbb{E}_{\mathbf{Y}^{n-},\mathbf{J}_{k-1}}\left[1\wedge\frac{\omega(y)}{\omega(x_1)}W_{k-1}(\mathbf{x}^{n-})\right] \\ &= \mathbb{E}_{\mathbf{Y}^{n-},\mathbf{J}_{k-1}}\left[1\wedge\frac{\omega(y)}{\omega(x_1)}\prod_{i=1}^{k-1}\frac{\omega(Y_{J_i})}{\omega(x_{J_i})}\right], \end{aligned} \tag{2.2}$$

where $\mathbf{J}_{k-1} = (J_1,J_2,\ldots,J_{k-1})$. A useful observation is that the proposed values $(Y_1,Y_{J_1},\ldots,Y_{J_{k-1}})$ are independent of $\mathbf{x}^n$. Let $H^*(y,x_1) = \mathbb{E}_{\mathbf{X}^{n-}}[H(y,x_1,\mathbf{X}^{n-})]$ and let

$$\mathcal{A}_n = \left\{\mathbf{x}^n; \int |H(y,\mathbf{x}^n) - H^*(y,x_1)|q(y)\,dy \le n^{-1/8}\right\} \tag{2.3}$$

We have the following Lemma which mirrors Roberts *et al.* (1997), Lemma 2.1, which states that with sufficiently high probability we can focus upon $\mathbf{X}^n_{[nt]}$ ($\mathbf{Z}^n_t$) contained in $\mathcal{A}_n$. The proof of Lemma 2.1 is given in appendix A.

Lemma 2.1 *For $t>0$,*

$$\mathbb{P}(\mathbf{Z}^n_s \in \mathcal{A}_n, 0\le s\le t) \to 1 \quad \text{as } n\to\infty. \tag{2.4}$$

We are now in position to state and prove the main result of this Section, Theorem 2.2.

Theorem 2.2 *For $k\in\mathbb{N}$, let $\mathbf{X}^n_0 \sim \pi_n$, then*

$$Z^n_{\cdot,1} \Rightarrow Z_\cdot \quad \text{as } n\to\infty, \tag{2.5}$$

where $Z_\cdot$ is a Markov jump process with infinitesimal generator

$$Gh(x) = k\int\{h(y)-h(x)\}H^*(y,x)q(y)\,dy, \tag{2.6}$$

for any C_c^∞ function h.

Proof. We begin by defining the (discrete time) generator of $\mathbf{X}^n$,

$$G_nh(\mathbf{x}^n) = n\mathbb{E}\left[\{h(\mathbf{Y}^n)-h(\mathbf{x}^n)\}\left\{1\wedge\frac{\pi_n(\mathbf{Y}^n)}{\pi_n(\mathbf{x}^n)}\right\}\right], \tag{2.7}$$

where h is any C_c^∞ function of the first component. Note that if there is no proposed update in the first component then $Y_1^n = x_1$. Therefore letting $\chi^n = 1$ if there is a proposed update of the first component and 0 otherwise, we have that

$$\begin{aligned} G_n h(\mathbf{x}^n) &= \sum_{i=0}^{1} n\mathbb{P}(\chi^n = i)\mathbb{E}\left[\{h(\mathbf{Y}^n) - h(\mathbf{x}^n)\}\left\{1 \wedge \frac{\pi_n(\mathbf{Y}^n)}{\pi_n(\mathbf{x}^n)}\right\}\middle| \chi^n = i\right] \\ &= n \times \frac{k}{n} \times \mathbb{E}\left[\{h(\mathbf{Y}^n) - h(\mathbf{x}^n)\}\left\{1 \wedge \frac{\pi_n(\mathbf{Y}^n)}{\pi_n(\mathbf{x}^n)}\right\}\middle| \chi^n = 1\right] \\ &= k\mathbb{E}_{Y_1}\left[(h(Y_1) - h(x_1))\mathbb{E}_{\mathbf{Y}^{n-},\mathbf{J}_{k-1}}\left[1 \wedge \frac{\omega(Y_1)}{\omega(x_1)}\prod_{j=1}^{k-1}\frac{\omega(Y_{J_j})}{\omega(x_{J_j})}\right]\right]. \end{aligned} \tag{2.8}$$

We compare $G_n h(\mathbf{x}^n)$ with the generator $Gh(x)$ defined in (2.6) for the limiting jump process. Now by (2.3), for all $\mathbf{x}^n \in \mathcal{A}_n$ and $h \in C_c^\infty$,

$$\begin{aligned} & |G_n h(\mathbf{x}^n) - Gh(x_1)| \\ = & \left|\int \{h(y) - h(x_1)\}q(y)\left(\mathbb{E}\left[1 \wedge \frac{\omega(y)}{\omega(x_1)}\prod_{j=1}^{k-1}\frac{\omega(Y_{J_j})}{\omega(x_{J_j})}\right] - H^*(y,x)\right)\,dy\right| \\ = & \left|\int \{h(y) - h(x_1)\}q(y)\left(H(y,x_1,\mathbf{x}^{n-}) - H^*(y,x)\right)\,dy\right| \\ \leq & \; 2\sup_z |h(z)| \int q(y)\left(H(y,\mathbf{x}^n) - H^*(y,x)\right)\,dy \\ \leq & \; 2\sup_z |h(z)| n^{-\frac{1}{8}} \to 0 \quad \text{as } n \to \infty. \end{aligned} \tag{2.9}$$

Hence,

$$\sup_{\mathbf{x}^n \in \mathcal{A}_n} |G_n h(\mathbf{x}^n) - Gh(x_1)| \to 0 \quad \text{as } n \to \infty. \tag{2.10}$$

The Theorem follows along identical lines to Roberts *et al.* (1997), Theorem 1.1. Since C_c^∞ separates points (see, Ethier and Kurtz (1986), page 113), the Theorem follows from (2.10) and Lemma 2.1 by Corollary 8.7 (f) of Chapter 4 of Ethier and Kurtz (1986). □

We proceed by discussing properties of the limiting jump process. Let

$$W_k^* \stackrel{D}{=} \prod_{i=1}^{k}\frac{\omega(Y_i)}{\omega(X_i)}, \tag{2.11}$$

where $Y_i \sim q(\cdot)$ and $X_i \sim f(\cdot)$. Then $\mathbb{E}[1 \wedge W_k^*]$ denotes the mean acceptance probability, in stationarity, of a proposed move and $k\mathbb{E}[1 \wedge W_k^*]$ denotes the

corresponding mean number of components updated. Moreover, $k\mathbb{E}[1 \wedge W_k^*]$ denotes the mean number of jumps, per unit time, of the limiting jump process, and hence, we seek k which maximises $k\mathbb{E}[1 \wedge W_k^*]$.

The distribution of W_k^* depends largely on the *closeness* of the target $(f(\cdot))$ and proposal $(q(\cdot))$ distributions with $W_k^* \equiv 1$ if for all x, $f(x) \equiv q(x)$. Let $g(x) = \log \omega(x) = \log f(x) - \log q(x)$, then

$$\log W_k^* \stackrel{D}{=} \sum_{i=1}^{k} \{g(Y_i) - g(X_i)\}, \tag{2.12}$$

where the $\{g(Y_i) - g(X_i)\}$ are independent and identically distributed. Note that $\mathbb{E}[g(Y_1)] = -D(q\|f)$ and $\mathbb{E}[g(X_1)] = D(f\|q)$, where for two probability density functions u and v,

$$D(u\|v) = \int u(x) \log\{u(x)/v(x)\}\, dx \tag{2.13}$$

is the Kullback-Leibler divergence. Hence,

$$\mathbb{E}[g(Y_1) - g(X_1)] = -\{D(q\|f) + D(f\|q)\} = -I, \text{ say,} \tag{2.14}$$

which makes explicit the role played by the *closeness* of the two densities. It should be noted that $I = \infty$ if there exists x such that $q(x) > 0$ and $f(x) = 0$, in such cases efficient independence sampling may still exist, for example, $X \sim U(0,1)$ and $Y \sim U(0, 1+\epsilon)$ for small, positive ϵ.

For finite I, it follows from (2.12) by the Central limit Theorem that for large k, $\log W_k^*$ is approximately Gaussian with mean $k\mathbb{E}[g(Y_1) - g(X_1)]$ and variance $kvar(g(Y_1) - g(X_1)) = kJ$, say. Now if I is small, which will be the case where the Central limit theorem is relevant, then $q(x) \approx f(x)$. Moreover, if $f(x) = q(x)\{1 + \epsilon(x)\}$ where $\epsilon(x)$ is small, then it is straightforward to show that $I = \int q(x)\{\epsilon(x)^2 + O(\epsilon(x)^3\}\, dx$ and that $J = 2\int q(x)\{\epsilon(x)^2 + O(\epsilon(x)^3\}\, dx \approx 2I$. Thus for k large, with $\log W_k^* \approx V_k^* \equiv N(-kI, kJ)$, we have by Roberts *et al.* (1997), Proposition 2.4, that

$$\begin{aligned} k\mathbb{E}[1 \wedge \exp(\log W_k^*)] &\approx k\mathbb{E}[1 \wedge \exp(V_k^*)] \\ &= k \times \left\{ \Phi\left(-\frac{kI}{\sqrt{kJ}}\right) + \exp\left(-kI + \frac{kJ}{2}\right) \Phi\left(-\sqrt{kJ} + \frac{kI}{\sqrt{kJ}}\right) \right\} \\ &\approx k \times 2\Phi\left(-\sqrt{\frac{kI}{2}}\right), \end{aligned} \tag{2.15}$$

where the latter approximation follows from setting $J = 2I$. Replacing k by z^2 and I by $\tilde{I} = \sqrt{2}I$ in the right hand side of (2.15), we obtain $j(z) = 2z^2\Phi(-z\sqrt{\tilde{I}}/2)$, which is the function maximized in Roberts *et al.*

(1997), Corollary 1.2 to maximise the optimal scaling of the random walk Metropolis algorithm. The only difference is the form of I which here depends upon the Kullback-Leibler divergence between the target and proposal distribution, whereas in Roberts *et al.* (1997) $I \equiv \mathbb{E}_f[(f'(X)/f(X))^2]$ and depends upon the *smoothness* of $f(\cdot)$. Most importantly, $z^2 I = 2.835$ maximises $j(z)$ and therefore k should be chosen approximately equal to $2.835/I$. Thus if I is small (there is close agreement between $f(\cdot)$ and $q(\cdot)$) k will be large. Moreover, mirroring Roberts *et al.* (1997), Corollary 1.2, such a k corresponds to a mean acceptance probability of (approximately) 0.234. Thus it is not necessary to compute I but instead suffices to monitor the mean acceptance probability. This will be shown to be a useful guiding principle in the examples below. However, it should be noted that scenarios exist, see Section 3.2 below, where the acceptance rate is above (below) 0.234 for all k, in such cases it is optimal to choose $k = n$ ($k = 1$).

Returning to optimising the independence sampler in the case $X \sim U(0,1)$ and $Y \sim U(0, 1+\epsilon)$, it is straightforward to show that the probability a proposed move is accepted is $(1+\epsilon)^{-k}$. Optimising the function $k(1+\epsilon)^{-k}$ gives $k = 1/\log(1+\epsilon)$, and hence for small ϵ, $k \approx 1/\epsilon$. Thus as $\epsilon \downarrow 0$, the optimal acceptance probability ($(1+\epsilon)^{-1/\log(1+\epsilon)} \approx (1-\epsilon)^{1/\epsilon}$) converges to $\exp(-1) = 0.368$. Therefore non-trivial asymptotic acceptance probabilities can exist in the case $I = \infty$ and typically these will be different from 0.234.

A key question is how does the independence sampler compare to the random walk Metropolis algorithm. Provided $\sup_x \omega(x) < \infty$, Theorem 2.2 holds and we have that the mixing of the independence sampler algorithm is $O(n)$, the same order of mixing as for the random walk Metropolis algorithm for continuous (and sufficiently differentiable) densities. The mixing of the random walk Metropolis algorithm for discontinuous densities is $O(n^2)$, Neal *et al.* (2012) whilst modifications such as Metropolis adjusted Langevin algorithms (MALA) and hybrid Monte Carlo (HMC) algorithms mix in $O(n^{\frac{1}{3}})$ and $O(n^{\frac{1}{4}})$ iterations, see Roberts and Rosenthal (1998) and Beskos *et al.* (2013), respectively, for sufficiently well behaved (continuous) target densities. Thus the independence sampler is competitive with the random walk Metropolis algorithm and Theorem 2.2 holds under very weak conditions compared with those imposed for corresponding random walk Metropolis algorithms. The similarity of the right hand side of (2.15) to $j(z)$ might suggest that computing I for the two algorithms would assist in comparing there performances with smaller I the better. However, the different nature of the moves, global in the independence sampler and local in the random walk Metropolis, means that this is not the case. In simulation studies with $X \sim N(0,1)$, $Y \sim N(0, \phi^2)$ and a range of $n \geq 50$, the independence sampler, with appropriately chosen k was found to outperform the optimal random walk Metropolis algorithm ($\sigma = 2.4/\sqrt{n}$) for $1 \leq \phi \leq 2.4$.

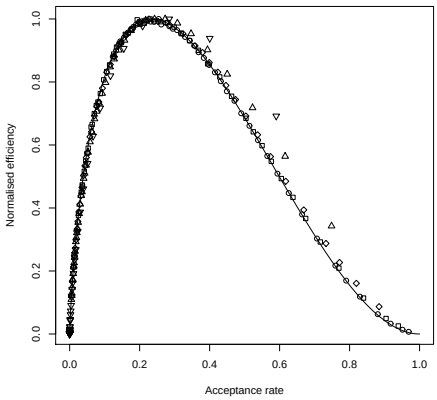
Thus the independence sampler is competitive with, and often superior to, random walk Metropolis, for continuous target densities so long as a reasonable choice of $q(\cdot)$ is made, and is clearly preferable for discontinuous target densities which is often the case in real life Bayesian problems, see Section 3.

3 Examples

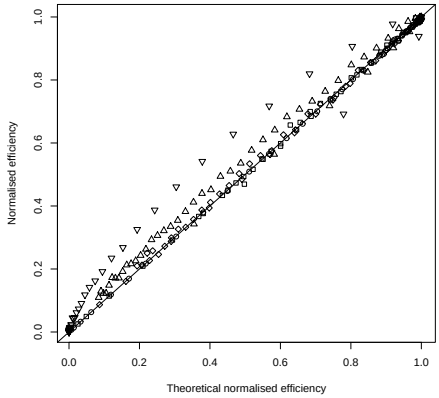
3.1 Introduction

In this Section we illustrate how large scale independence sampling can be exploited to construct effective MCMC algorithms. We start with an independent and identically distributed Gaussian product density as the target distribution and consider both Gaussian and t-distribution proposals. Specifically, we take $\pi(\mathbf{x}) = \prod_{i=1}^{n} f(x_i)$, where $f(x)$ is a standard Gaussian density. The proposal distributions are symmetric about 0 with Gaussian proposals $q_N(y) = (\sqrt{2\pi}\lambda)^{-1}\exp(-y^2/2\lambda^2)$, where $\lambda \geq 1$ and t-distribution proposals $q_t(y) = \Gamma((\nu+1)/2)/(\sqrt{\nu\pi}\Gamma(\nu/2))(1+x^2/\nu)^{-\frac{\nu+1}{2}}$ $(\nu \in \mathbb{N})$. We conducted a simulation study using 5 Gaussian and 5 t-distribution proposals with $n = 1000$ and 10^6 iterations of the MCMC algorithm starting from the stationary distribution. For each proposal distribution we considered 50 choices of k, the exact choices of which depended on I and were chosen to give acceptance rates on the full range 0 to 1.

For the Gaussian proposal it is straightforward to show that $I = 1/2(\lambda - 1/\lambda)^2$. We considered $\lambda = 1.05, 1.1, 1.2, 1.5, 2$ with corresponding $I = 0.0048, 0.0182, 0.0672, 0.347, 1.125$. A key quantity for comparing the independence sampler for different choices of λ, and hence I, is the normalised efficiency. We define the normalised efficiency for k as the mean number of components updated ($k\times$ acceptance rate) when proposing to update k components divided through by the maximum mean number of components updated for $j = 1, 2, \ldots, n$. Correspondingly the normalised theoretical efficiency is given by $j(z) = 2z^2\Phi(-z/2)/\sup_y\{2y^2\Phi(-y/2)\} = 2z^2\Phi(-z/2)/1.3257$ from applying the central limit theorem approximation obtained in Section 2. The plots in Figure 1 show that in all cases the optimal acceptance rate is close to 0.234 with very similar behaviour for the normalised efficiency varying with acceptance rate, even for $\lambda = 2$ with $I = 1.125$. Similar results are obtained in Section in Neal and Roberts (2006), Section 6 for the optimal performance of the random walk Metropolis algorithm. As $\lambda \downarrow 1$, $I \downarrow 0$ and the agreement between the observed normalised efficiency normalised theoretical efficiency becomes very close.



(a) Normalised efficiency against acceptance rate



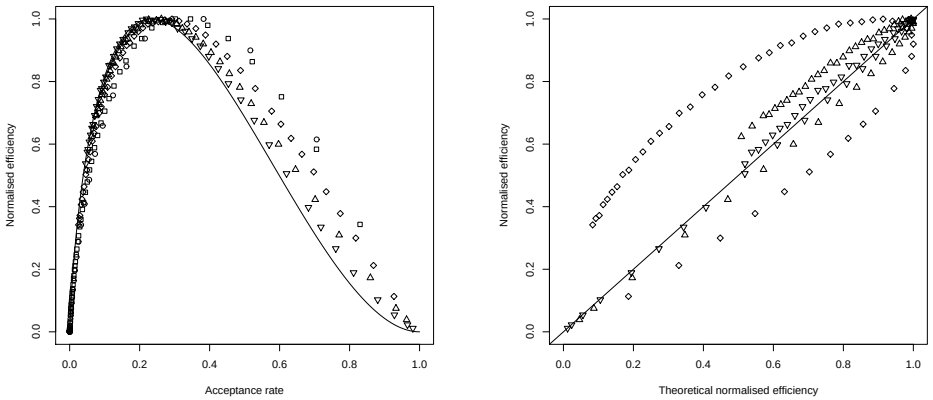
(b) Normalised efficiency against normalised theoretical efficiency

Figure 1: Gaussian proposal $\lambda = 1.05(\bigcirc), 1.1(\square), 1.2(\Diamond), 1.5(\triangle), 2.0(\bigtriangledown)$. (a) Solid line given by $j(z) = 2z^2\Phi(-z/2)/1.3257$ plotted against acceptance rate. (b) Solid line $x = y$.

For the t-distribution, $I = \infty$ for $\nu = 1, 2$, otherwise

$$I = \frac{1}{\nu - 2} + \frac{\nu + 1}{2} \left\{ \mathbb{E}[\log(1 + X^2/\nu)] - \mathbb{E}[\log(1 + Y_\nu^2/\nu)] \right\},$$

where $X \sim N(0, 1)$ and $Y \sim t_\nu$. It is not possible to obtain a closed form analytical expression for I but it is straightforward to estimate using Monte Carlo integration. We consider $\nu = 1, 2, 5, 10, 20$ with corresponding $I = \infty, \infty, 0.1582, 0.0338, 0.0083$. The plots in Figure 2 show that the optimal acceptance rate is higher than 0.234 for a t-distribution proposal with an optimal acceptance rate of 0.383 corresponding to $k = 3$ for a t_1 proposal. Note that this is close to $\exp(-1)$, the optimal acceptance rate of the uniform distributions example given in Section 2. It is worth noting that choosing k to obtain an acceptance rate of approximately 0.234 is in general a good approach as only a small loss in efficiency is observed. As ν increases the optimal acceptance rate converges towards 0.234 and the normalised efficiency tends towards the theoretical normalised efficiency given by the central limit theorem approximation. This is further demonstrated in Figure 2b by plotting normalised efficiency against normalised theoretical efficiency. Note that $\nu = 1$ and $\nu = 2$ do not feature on this plot as $I = \infty$.



(a) Normalised efficiency against acceptance rate

(b) Normalised efficiency against normalised theoretical efficiency

Figure 2: Gaussian proposal $t = 1(\bigcirc), 2(\square), 5(\diamondsuit), 10(\triangle), 20(\triangledown)$. (a) Solid line given by $j(z) = 2z^2\Phi(-z/2)/1.3257$ plotted against acceptance rate. (b) Solid line $x = y$.

3.2 Homogeneously mixing SIR epidemic

In this Section we show how the importance sampler can be applied to temporally observed, homogeneously mixing SIR epidemic models, Bailey (1975); O'Neill and Roberts (1999). We assume that there is a population of size N with the disease introduced into the population via a single introductory case. (The extension to multiple introductory cases is trivial.) We assume that the disease follows an SIR epidemic model, where initially all individuals, except the introductory case, are susceptible. On becoming infectious, an individual is infectious for a given period of time, distributed according to a Gamma random variable $Q \sim \text{Gamma}(\alpha, \delta)$. (Alternative infectious period distributions can easily be considered.) Whilst infectious, an individual i, say, makes infectious contacts at the points of a homogeneous Poisson point process with rate β with the individual contacted chosen uniformly at random from the entire population. Infectious contacts with susceptible individuals result in the immediate infection of the individual and the start of their infectious period. Infectious contacts with infectives have no effect on the recipient.

Suppose that m individuals are infected during the course of the epidemic and we are analysing the completed epidemic data. For each individual, i say, infected during the course of the epidemic there will be an infection time, I_i and a removal (recovery) time, R_i, which mark the start and end of

the infectious period, respectively. We follow O'Neill and Roberts (1999), Neal and Roberts (2005) and Xiang and Neal (2014) in assuming that the removal times, $\mathbf{R} = (R_1, \ldots, R_m)$ are observed, whilst the infection times $\mathbf{I} = (I_1, \ldots, I_m)$ are unobserved. Furthermore, we assume that the removal times are ordered such that $R_1 \leq R_2 \leq \ldots \leq R_m$. The key interest is in the posterior distribution of $\pi(\beta, \alpha, \delta|\mathbf{R})$ and to obtain samples from this distribution imputation of $\mathbf{I}$ is required.

We use the MCMC algorithm proposed in Xiang and Neal (2014), Section 3 with the modification that the number of components to be updated is fixed to $k \in \{1, 2, \ldots, m\}$. As with Xiang and Neal (2014), the MCMC algorithm is applied to the extensively studied Abakaliki smallpox outbreak, (Bailey, 1975, p.125), O'Neill and Roberts (1999); O'Neill and Becker (2001); McKinley *et al.* (2014), where $m = 30$ and $N = 120$. We considered various fixed values of $\alpha = 1, 3, 10$ with optimal $k = 9, 17$ and 30, respectively, based upon the maximised mean number of components updated over 100000 iterations, see Figure 3. For $\alpha = 1, 3, 10$, the corresponding values of k which had acceptance rates closest to 23.4% were $k = 10, 19$ and 29, respectively. Thus choosing k so that the acceptance rate is close to 23.4% is effective in obtaining a close to optimal algorithm. In Xiang and Neal (2014), the situation where α is assumed to be unknown is also considered with the posterior mean of α being 33.8. For unknown α, the acceptance rate is above 23.4% for all k and thus $k = m(= 30)$ performs optimally.

We can go further in illustrating the usefulness of the theoretical results derived in Section 2 for choosing k. In Figure 4, we plot the normalised efficiency for $\alpha = 1, 2, \ldots, 9$, since for $\alpha > 9$, the acceptance rate is always above 23.4%. Also on the plot (in red) is the normalised theoretical curve $j(z) = 2z^2\Phi(-z/2)/1.3257$ given by (2.15) against acceptance rate $2\Phi(-z)$. In a similar fashion to Section 3.1 this illustrates that the asymptotic results which are valid as the number of components updated tend to ∞ are applicable for small k.

A simulation study was conducted to study the general applicability of the results obtained above for the Abakaliki data. Data sets were simulated with $N = 200, 400, 600, 800, 1000, 1200$, $m = 0.25N, 0.5N, 0.75N$ and $\alpha = 1, 2, 3, 5, 10, 15, 20$ with $\delta = 0.1\alpha$ chosen to give a mean infectious period of 10 and β to give the mean size of a major epidemic outbreak to be 10. For each α, the optimal k increases with N and vice versa. Throughout choosing k with acceptance rate closest to 23.4% produced close to optimal performance. Plots of the normalised efficiency against the acceptance rate showed increasing agreement with the asymptotic theoretical curve as N increases.

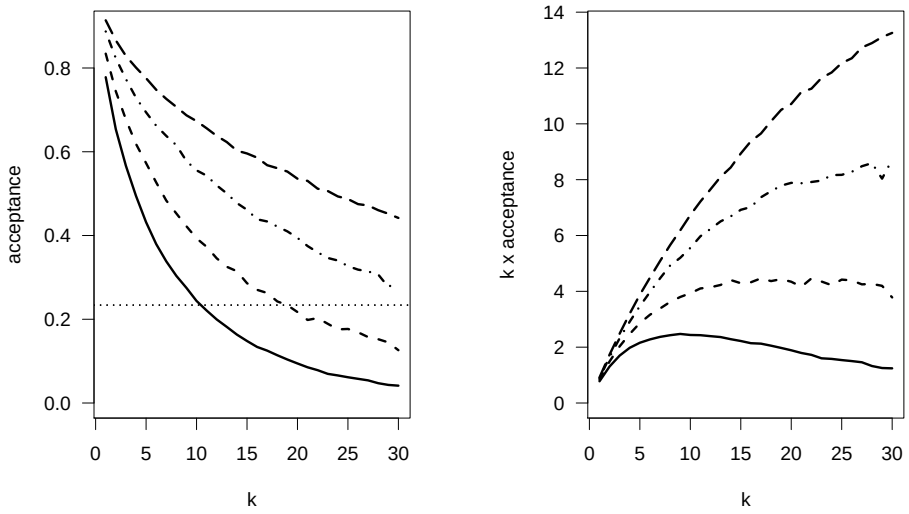


Figure 3: Acceptance rate (left) and mean number of components updated (right) against k for $\alpha = 1$ (solid), 3 (dashed), 10 (dot-dashed) and unknown (posterior mean 33.8) (long dashed).

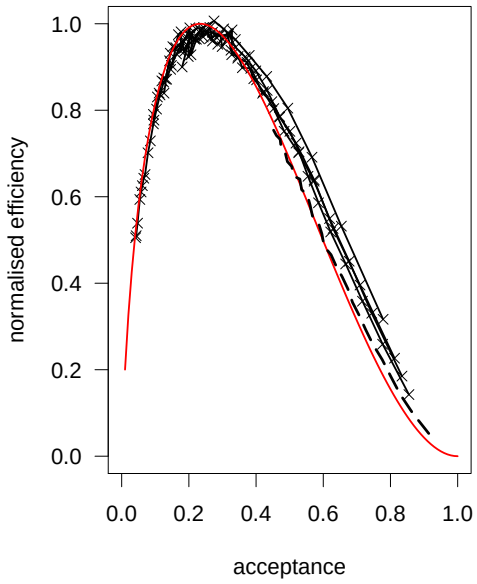


Figure 4: Normalised mean number of components updated against acceptance rate, overlaid by the theoretical normalised curve (red), given by $j(z) = 2z^2\Phi(-z/2)/1.3257$.

Table 1: Observed cluster size distribution of Tuberculosis bacteria genotype data, Small *et al.* (1994).

Cluster size	1	2	3	4	5	8	10	15	23	30
Number of clusters	282	20	13	4	2	1	1	1	1	1

3.3 Birth-Death-Mutation model

In this section we consider a birth-death-mutation (BDM) model which is applicable to the early stages of a mutating disease. The model has previously been used by Tanaka *et al.* (2006); Sisson *et al.* (2007); Fearnhead and Prangle (2012); Del Moral *et al.* (2012); Neal and Huang (2015) to analyse data from a tuberculosis outbreak in San Francisco in the early 1990s reported in Small *et al.* (1994). We explore and seek to optimise the performance of the forward simulation MCMC algorithm introduced by Neal and Huang (2015). Note that all the other analyses reported above used ABC algorithms.

The data consist of the genotypes of 473 bacteria samples sampled from individuals infected with tuberculosis in San Francisco during an observational period in 1991-92. The data are clustered by genotype and summarised in Table 1. Let N_t denote the total number of tuberculosis cases at time t. The data are assumed to be a random sample taken at time T, where $T = \min\{t; N_t = 10000\}$ evolving from $N_0 = 1$.

The BDM model is a Markov process defined as follows. Individuals are classified by (geno)type. Each individual born into the process has an exponentially distributed lifetime (infectious period) with mean $1/\delta$. Whilst alive individuals give birth (infects) and mutates at the points of independent homogeneous Poisson point processes with rates α and ϑ, respectively. Each individual born inherits the (geno)type of their parent and all mutations result in the creation of a new, previously unseen (geno)type (infinite allele model, Kimura and Crow (1964)). We reparameterise the model by setting $\phi = \alpha + \delta + \vartheta$, $a = \alpha/\phi$ and $d = \delta/\phi$, where ϕ is the rate at which events occur for an individual, a is the probability that the event is a birth (infection) and d is the probability that the event is a death (recovery). Since the stopping time T at which the population is observed only depends upon the number of individuals alive in the population, there is no information in the data about ϕ. Thus, without loss of generality, we assume $\phi = 1$ making inference about (a, d) given the genotype data $\mathbf{x}$. In order to construct a tractable likelihood it is necessary to generate the state of the population at

time T, $N_T = 10000$. This can be done using a non-centered parameterisation Papaspoliopoulos *et al.* (2003) where the augmented data $\mathbf{y} = (\mathbf{u}, \mathbf{w}, \mathbf{v})$ consist of realisations of $U(0,1)$ with $(\mathbf{u}, \mathbf{w})$ combine with (a, d) to generate the underlying state of the BDM model at time T and $\mathbf{v}$ is used to estimate the probability of observing $\mathbf{x}$. Details of the construction are given in Neal and Huang (2015), Section 4.

The time consuming step of the MCMC algorithm for the BDM model is the simulation of the state of the process using $(\mathbf{u}, \mathbf{w})$ and (a, d). In Neal and Huang (2015), (a, d) are updated using random walk Metropolis keeping $(\mathbf{u}, \mathbf{w})$ fixed and $(\mathbf{u}, \mathbf{w})$ are updated using an independence sampler, draws from $U(0,1)$, keeping (a, d) fixed. We thus focus on the independence sampler for updating $(\mathbf{u}, \mathbf{w})$. Note that $\mathbf{v}$ is updated by a separate independence sampler but this is very fast to implement (no need to simulate the BDM process), and so we don't comment on this step. The dimensions of $\mathbf{u}$ and $\mathbf{w}$ are the same but vary from iteration to iteration, typically being around 30000. To circumvent issues with this Neal and Huang (2015) used random vectors of a fixed length $n = 100000$ with only those elements needed to simulate the process used. In this paper we also used a fixed length vector updating k out of n components in $\mathbf{u}$ and $\mathbf{w}$ noting that in each simulation not all (updated) components will be used.

In Neal and Huang (2015), $\mathbf{u}$ and $\mathbf{w}$ are broken down into blocks of 50 components with 1 component in each block proposed to be updated. This amounts to proposing to update $n/50 = 2000$ values in each iteration of which typically around 600 are used in the simulation. In this paper we propose to update k components each of $\mathbf{u}$ and $\mathbf{w}$, $(u_{I_1^u}, u_{I_2^u}, \ldots, u_{I_k^u})$ and $(w_{I_1^w}, w_{I_2^w}, \ldots, w_{I_k^w})$, where $\{I_1^u, I_2^u, \ldots, I_k^u\}$ ($\{I_1^w, I_2^w, \ldots, I_k^w\}$) is a uniformly random sample without replacement from $\{1, 2, \ldots, n\}$, for the sake of consistency with the updating strategy throughout this paper. In addition to using different values for k, we also examine the performance of the algorithm using $n = 60000, 80000$ and the original 100000, which are all found to be empirically sufficient. We ran the MCMC algorithm for 1.1×10^6 iterations with the first 10^5 iterations discarded as burn-in. The acceptance rate is plotted against k for all three values of n on the left of Figure 5, which is analogous to Figure 3, with the mean number of components updated on the right. The results shown in Figures 5 demonstrate an interesting departure from those found earlier in the paper with an optimal acceptance rate of 23.4%. The mean number of components updated increases with k even as the acceptance rate drops below 5%. However, for both parameters a and d, the effective sample size levels off at around 3000 for all $k \geq 2000$, which suggests that seeking to optimise the mean number of components updated does not tell the full story in this case.

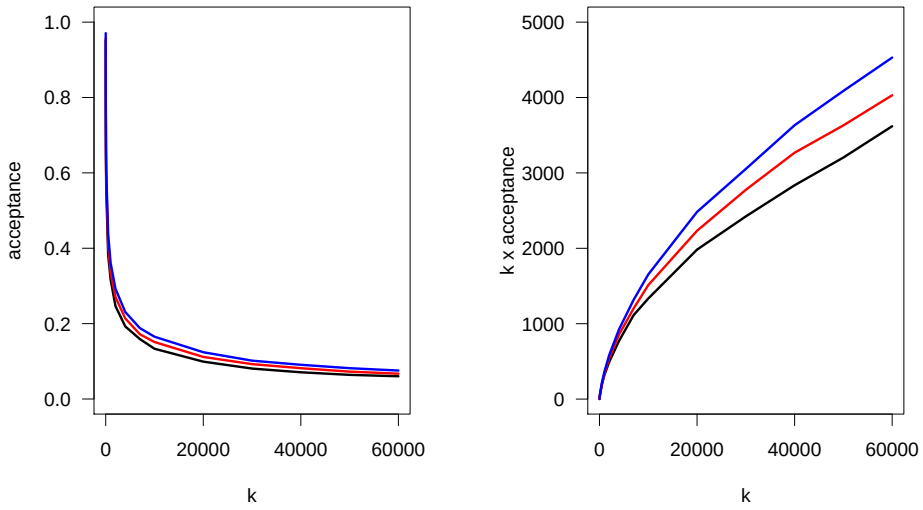


Figure 5: Acceptance rate (left) and mean number of components updated (right) against k for $n = 60000$ (black), 80000 (red) and 100000 (blue).

4 Conclusions

In this paper we have demonstrated the potential benefits, both theoretical and practical, of the independence sampler over the random walk Metropolis algorithm. In particular, we have shown that simple choices of proposal distributions can be used to construct effective independence samplers and that similar considerations to the tuning of the random walk Metropolis algorithm are required. There are a number of points to consider in the wider application of the results derived in Section 2 and applied in Section 3. Firstly, we have not considered the computational time required to update k components. In the homogeneously mixing epidemic model (Section 3.2), and in particular, the BDM model (Section 3.3) the time taken per iteration was essentially independent of k. However, it is possible for the homogeneously mixing epidemic model by careful updating of the calculation of the likelihood for the time taken per iteration to be smaller for smaller k. In such cases the optimal acceptance rate will be larger than 23.4% and if the time per iteration is proportional to k it will be optimal to update a single component at a time. Secondly, the theoretical results of Section 2 for independent and identically distributed product densities are shown to give clear guidance for optimising the independence sampler for the homogeneously mixing epidemic model but not for the BDM model. The reason for this difference is not immediately obvious but is likely to depend on the relationship of the observed data to the augmented data. For the

homogeneously mixing epidemics the local behaviour of $\mathbf{I}$ is important, for example ensuring $\mathbf{I}$ is consistent with an epidemic outbreak, whereas for the BDM model it is global properties of $(\mathbf{U}, \mathbf{W})$, the total numbers of births, deaths and mutations which are most important. For the random walk Metropolis algorithm optimal scaling results differ depending upon whether the acceptance probability depends on local behaviour (discontinuous product densities, Neal *et al.* (2012)) or global behaviour (continuous product densities, Roberts *et al.* (1997), elliptically symmetric densities Sherlock and Roberts (2009)) of the proposed moves.

Acknowledgements

This research was supported by the Engineering and Physical Sciences Research Council under grant EP/J008443/1. We would like to thank an anonymous referee for their careful reading of the paper and suggestions for improving presentation of the findings.

References

Bailey, N.T.J. (1975) *The Mathematical Theory of Infectious Diseases and its Applications. Second edition.* Griffin, London.

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J-M. and Stuart, A. (2013) Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli*, **19**, 1501–1534.

Dellaportas, P. and Roberts, G.O. (2013) An introduction to MCMC. *Spatial Statistics and Computational Methods* (J. Møller, eds.) Springer, New York, 1–43.

Del Moral, P., Doucet, A. and Jasra, A. (2012) An adaptive sequential Monte Carlo method for approximate Bayesian compuation. *Stat. Comput.* **22**, 1009–1020.

Ethier, S.N. and Kurtz, T.G. (1986) *Markov processes, characterization and convergence.* Wiley, New York.

Fearnhead, P. and Prangle, D. (2012) Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation (with discussion). *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74**, 419–474

Kimura, M. and Crow, J. (1964). The number of alleles that can be maintained in a finite population. *Genetics* **49**, 725-738.

McKinley, T.J., Ross, J.V., Deardon, R. and Cook, A.R. (2014) Simulation-based Bayesian inference for epidemic models. *Comput. Statist. Data Anal.* **71**, 434–447.

Neal, P. and Huang, C.L.T. (2015) Forward Simulation MCMC with applications to stochastic epidemic models. *Scand. J. Stat.* **42**, 378–396.

Neal, P.J. and Roberts, G.O. (2005) A case study in non-centering for data augmentation: Stochastic epidemics. *Stat. Comput.* **15**, 315–327.

Neal, P.J. and Roberts, G.O. (2006) Optimal scaling for partially updating MCMC algorithms. *Ann. Appl. Probab.* **16**, 475–515.

Neal, P.J., Roberts, G.O. and Yuen, W.K. (2012) Optimal Scaling of Random Walk Metropolis algorithms with discontinuous target densities. *Ann. Appl. Probab.* **22**, 1880–1927

O'Neill, P.D. and Becker, N.G. (2001). Inference for an epidemic when susceptibility varies. *Biostatistics* **2**, 99–108.

O'Neill, P.D. and Roberts, G.O. (1999). Bayesian inference for partially observed stochastic epidemics. *J. Roy. Statist. Soc. Ser. A* **162**, 121–129.

Papaspoliopoulos, O. , Roberts, G.O. and Sköld, M. (2003) Non-centered parameterisations for hierarchical models and data augmentation. *Bayesian Statistics 7* (J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith and M. West, eds.) Oxford University Press, 307–326.

Ripley, B. (1987) *Stochastic Simulation.* Wiley, New York.

Roberts, G.O., Gelman, A. and Gilks, W.R. (1997) Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, **7**, 110–120.

Roberts, G. O. and Rosenthal, J. S. (1998) Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 255–268.

Sherlock, C. and Roberts, G.O. (2009) Optimal scaling of the random walk Metropolis on elliptically symmetric unimodal targets. *Bernoulli* **15**, 774–798.

Sisson, S. A., Fan, Y. and Tanaka, M. M. (2007) Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA*, **104**, 1760–1765.

Small, P. M., Hopewell, P. C., Singh, S. P., Paz, A., Parsonnet, J., Ruston, D. C., Schecter, G. F., Daley, C. L., and Schoolnik, G. K. (1994) The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. *New England Journal of Medicine* **330** 1703-1709.

Tanaka, M. M., Francis, A. R., Luciani, F. and Sisson, S. A. (2006) Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. *Genetics* **173**, 1511–1520.

Tiernay, L. (1994) Markov chains for exploring posterior distributions. *Ann. Statist.* **22**, 1701–1728.

Xiang, F. and Neal, P. (2014) Efficient MCMC for temporal epidemics via parameter reduction. *Comput. Statist. Data Anal.* **80**, 240–250.

A Proof of Lemma 2.1

Since $\mathbf{Z}_0^n \sim \pi_n$, for all $0 \leq s \leq t$, $\mathbf{Z}_s^n \sim \pi_n$, since π_n is the stationary distribution of $\mathbf{Z}_t^n$. Therefore, we have that

$$\mathbb{P}(\mathbf{Z}_s^n \notin \mathcal{A}_n, \text{for some } 0 \leq s \leq t) \leq tn\mathbb{P}(\mathbf{X}_0^n \notin \mathcal{A}_n). \tag{A.1}$$

Now,

$$\begin{aligned}\mathbb{P}(\mathbf{X}_0^n \notin \mathcal{A}_n) &= \mathbb{P}\left(\int |H(y, \mathbf{X}_0^n) - H^*(y, X_{0,1})|q(y)\,dy > n^{-\frac{1}{8}}\right) \\ &= \int \mathbb{P}\left(\int |H(y, \mathbf{x}^n) - H^*(y, x_1)|q(y)\,dy > n^{-\frac{1}{8}}\right) \pi_n(\mathbf{x}^n)\,d\mathbf{x}^n.\end{aligned} \tag{A.2}$$

Applying Markov's inequality to the right hand side of (A.2), we have that

$$\mathbb{P}(\mathbf{X}_0^n \notin \mathcal{A}_n) \leq \int \sqrt{n}\left\{\int |H(y, \mathbf{x}^n) - H^*(y, x_1)|q(y)\,dy\right\}^4 \pi_n(\mathbf{x}^n)\,d\mathbf{x}^n. \tag{A.3}$$

It then follows by Jensen's inequality that

$$\begin{aligned}\mathbb{P}(\mathbf{X}_0^n \notin \mathcal{A}_n) &\leq \int \sqrt{n}\left\{\int (H(y, \mathbf{x}^n) - H^*(y, x_1))^4 q(y)\,dy\right\} \pi_n(\mathbf{x}^n)\,d\mathbf{x}^n \\ &= \sqrt{n}\int \left\{\int (H(y, \mathbf{x}^n) - H^*(y, x_1))^4 \pi_n(\mathbf{x}^n)\,d\mathbf{x}^n\right\} q(y)\,dy.\end{aligned} \tag{A.4}$$

We now focus on the inner integral on the right hand side of (A.4). Since $\mathbb{E}_{\mathbf{X}^{n-}}[H(y,x_1,\mathbf{X}^{n-})] = H^*(y,x_1)$, we have that

$$\begin{aligned} & \int (H(y,\mathbf{x}^n) - H^*(y,x_1))^4 \pi_n(\mathbf{x}^n)\, d\mathbf{x}^n \\ = & \int \mathbb{E}[(H(y,x_1,\mathbf{X}_0^{n-}) - \mathbb{E}_{\mathbf{X}_0^{n-}}[H(y,x_1,\mathbf{X}_0^{n-})])^4] f(x_1)\, dx_1. \quad \text{(A.5)} \end{aligned}$$

Let $\mathcal{I}_n = \{\mathbf{i} \in \{2,3,\ldots,n\}^{k-1}; i_1 < i_2 < \ldots < i_{k-1}\}$. Then letting

$$\hat{H}_{\mathbf{i}}(y,x_1,\mathbf{x}^{n-}) = \mathbb{E}_{\mathbf{Y}^n}\left[1 \wedge \frac{\omega(Y_1)}{\omega(x_1)} \prod_{l=1}^{k-1} \frac{\omega(Y_{i_l})}{\omega(x_{i_l})}\right], \quad \text{(A.6)}$$

we note that for all $\mathbf{i},\mathbf{j} \in \mathcal{I}_n$, $\hat{H}_{\mathbf{i}}(y,x_1,\mathbf{X}_0^{n-}) \stackrel{D}{=} \hat{H}_{\mathbf{i}}(y,x_1,\mathbf{X}_0^{n-})$, where $\stackrel{D}{=}$ denotes equality in distribution. Hence for all $\mathbf{i} \in \mathcal{I}_n$, $\mathbb{E}[\hat{H}_{\mathbf{i}}(y,x_1,\mathbf{X}_0^{n-})] = H^*(y,x_1)$. Therefore given that

$$H(y,x_1,\mathbf{X}_0^{n-}) = \binom{n-1}{k-1}^{-1} \sum_{\mathbf{i}} \hat{H}_{\mathbf{i}}(y,x_1,\mathbf{X}_0^{n-}), \quad \text{(A.7)}$$

it follows that

$$\begin{aligned} & \mathbb{E}[(H(y,x_1,\mathbf{X}_0^{n-}) - \mathbb{E}_{\mathbf{X}_0^{n-}}[H(y,x_1,\mathbf{X}_0^{n-})])^4] \\ = & \binom{n-1}{k-1}^{-4} \sum_{\mathbf{i}_1 \in \mathcal{I}_n} \sum_{\mathbf{i}_2 \in \mathcal{I}_n} \sum_{\mathbf{i}_3 \in \mathcal{I}_n} \sum_{\mathbf{i}_4 \in \mathcal{I}_n} \mathbb{E}\left[\prod_{j=1}^4 (\hat{H}_{\mathbf{i}_j}(y,x_1,\mathbf{X}_0^{n-}) - \mathbb{E}[\hat{H}_{\mathbf{i}_j}(y,x_1,\mathbf{X}_0^{n-})])\right]. \end{aligned} \quad \text{(A.8)}$$

Note that if $\mathbf{i},\mathbf{j} \in \mathcal{I}_n$ have no elements in common then $\hat{H}_{\mathbf{i}}(y,x_1,\mathbf{X}_0^{n-})$ and $\hat{H}_{\mathbf{j}}(y,x_1,\mathbf{X}_0^{n-})$ are independent. Therefore $\mathbb{E}[\prod_{j=1}^4 (\hat{H}_{\mathbf{i}_j}(y,x_1,\mathbf{X}_0^{n-}) - \mathbb{E}[\hat{H}_{\mathbf{i}_j}(y,x_1,\mathbf{X}_0^{n-})])]$ is only non-zero if and only if for $j = 1,2,3,4$, $\mathbf{i}_j$ has at least an element in common with one the other indices. Moreover, $|\mathbb{E}[\prod_{j=1}^4 (\hat{H}_{\mathbf{i}_j}(y,x_1,\mathbf{X}_0^{n-}) - \mathbb{E}[\hat{H}_{\mathbf{i}_j}(y,x_1,\mathbf{X}_0^{n-})])]| \leq 1$.

The number of combinations of $\mathbf{i}_1,\mathbf{i}_2 \in \mathcal{I}_n$ such that $\mathbf{i}_1$ and $\mathbf{i}_2$ have at least one element in common is

$$\binom{n-1}{k-1}\left\{\binom{n-1}{k-1} - \binom{n-k}{k-1}\right\}, \quad \text{(A.9)}$$

which is bounded above by $n^{2k-3}/\{(k-2)!\}^2$ for all sufficiently large n. Similarly, the number of combinations of $\mathbf{i}_1,\mathbf{i}_2,\rangle_3,\rangle_4 \in \mathcal{I}_n$ such that $\mathbf{i}_2$, $\mathbf{i}_3$ and $\mathbf{i}_4$ all have at least one element in common with $\mathbf{i}_1$ is

$$\binom{n-1}{k-1}\left\{\binom{n-1}{k-1} - \binom{n-k}{k-1}\right\}^3, \quad \text{(A.10)}$$

which is bounded above by $(k-1)^2 n^{4k-7}/\{(k-2)!\}^4$ for all sufficiently large n. Now $\mathbb{E}[\prod_{j=1}^4 (\hat{H}_{\mathbf{i}_j}(y, x_1, \mathbf{X}_0^{n-}) - \mathbb{E}[\hat{H}_{\mathbf{i}_j}(y, x_1, \mathbf{X}_0^{n-})])]$ is only non-zero if either $\mathbf{i}_1, \mathbf{i}_2, \rangle_3, \rangle_4 \in \mathcal{I}_n$ can be grouped into two pairs such that both pairs have at least one element in common or if three of the components all have at least one element in common with the fourth. (Note that there is overlap between these two classifications.) Thus using (A.9) and (A.10), it is straightforward to combine with (A.8) to show that

$$\begin{aligned} & \mathbb{E}[(H(y, x_1, \mathbf{X}_0^{n-}) - \mathbb{E}_{\mathbf{X}_0^{n-}}[H(y, x_1, \mathbf{X}_0^{n-})])^4] \\ \leq & \binom{n-1}{k-1}^{-4} \left\{ 3 \left(\frac{n^{2k-3}}{\{(k-2)!\}^2} \right)^2 + 4 \frac{(k-1)^2 n^{4k-7}}{\{(k-2)!\}^4} \right\} \\ \leq & \frac{(k-1)^4}{(n-k)^{4k-4}} \left\{ 3n^{4k-6} + 4(k-1)^2 n^{4k-7} \right\}. \end{aligned} \tag{A.11}$$

Since the bound obtained in (A.11), holds for all $y, x_1 \in \mathbb{R}$, it follows from (A.4) and (A.5) that

$$\begin{aligned} n\mathbb{P}(\mathbf{X}_0^n \notin \mathcal{A}_n) & \leq n\sqrt{n} \frac{(k-1)^4}{(n-k)^{4k-4}} \left\{ 3n^{4k-6} + 4(k-1)^2 n^{4k-7} \right\} \\ & \to 0 \quad \text{as } n \to \infty. \end{aligned} \tag{A.12}$$

The lemma immediately follows by combining (A.12) and (A.1).