Notes on Using Control Variates
for Estimation with Reversible MCMC Samplers

Petros Dellaportas [*]   Ioannis Kontoyiannis [†]

July 2009

Abstract

A general methodology is presented for the construction and effective use of control variates for reversible MCMC samplers. The values of the coefficients of the optimal linear combination of the control variates are computed, and adaptive, consistent MCMC estimators are derived for these optimal coefficients. All methodological and asymptotic arguments are rigorously justified. Numerous MCMC simulation examples from Bayesian inference applications demonstrate that the resulting variance reduction can be quite dramatic.

1   Introduction

Markov chain Monte Carlo (MCMC) methods provide the facility to draw, in an asymptotic sense, a sequence of dependent samples from a very wide class of probability measures in any dimension. This facility – together with the tremendous increase of computer power in recent years – makes MCMC perhaps the main reason for the widespread use of Bayesian statistical modeling across the entire spectrum of quantitative scientific disciplines.

This paper provides a firm methodological foundation for the construction and use of control variates for reversible MCMC samplers. Although popular in the standard Monte Carlo setting, control variates have received little attention in the MCMC literature. The proposed methodology will be shown, in many instances, to reduce the variance of the resulting estimators quite dramatically.

In the simplest Monte Carlo setting, when the goal is to compute the expected value of some function F evaluated on independent and identically distributed (i.i.d.) samples $X_1, X_2, \ldots$, the variance of the standard ergodic averages of the $F(X_i)$ can be reduced by exploiting available zero-mean statistics. If there are one or more functions $U_1, U_2, \ldots, U_k$ – the *control variates* – for which it is known that the expected value of $U(X_i)$ is equal to zero, then adding any linear combination, $\theta_1 U_1(X_i) + \theta_2 U_2(X_i) + \cdots + \theta_k U_k(X_i)$, to the $F(X_i)$ does not change the asymptotic mean of the corresponding ergodic averages. Moreover, if the best constant coefficients $\{\theta_j^*\}$ are used, then the variance of the estimates is no larger than before and often it is much smaller. The standard practice in this setting is to estimate the optimal $\{\theta_j^*\}$ adaptively, based on the same sequence of samples; see, e.g., Liu (2001), Givens and Hoeting (2005), Robert and Casella (2004)

---

[*]Department of Statistics, Athens University of Economics and Business, Patission 76, Athens 10434, Greece. Email: `petros@aueb.gr`.

[†]Department of Informatics, Athens University of Economics and Business, Patission 76, Athens 10434, Greece. Email: `yiannis@aueb.gr`.

for details. Because of the demonstrated effectiveness of this technique, in many important areas of application – e.g., in computational finance where Monte Carlo methods are a basic tool for the approximate computation of expectations, see Glasserman (2004) – a major research effort is devoted to the construction of effective control variates in specific applied problems.

However, up to now little has been established in the way of extending the above methodology to estimators based on MCMC samples, at least in part due to the intrinsic difficulties presented by the Markovian structure. For example, Mengersen et al. (1999) comment that "control variates have been advertised early in the MCMC literature (see, e.g., Green and Han (1992)), but they are difficult to work with because the models are always different and their complexity is such that it is extremely challenging to derive a function with known expectation." Indeed, there are two fundamental difficulties; not only is it hard to find nontrivial functions with known expectation with respect to the stationary distribution of the chain, but also, even in cases where such functions are available, there is no effective way to obtain useful estimates of the corresponding optimal coefficients $\{\theta_j^*\}$. The reason why this is a fundamentally difficult problem is that the MCMC variance of ergodic averages is intrinsically an infinite-dimensional object: It cannot be written in closed form as a function of the transition kernel and the stationary distribution of the chain.

An early reference of variance reduction for Markov chain samplers is Green and Han (1992), who exploit an idea of Barone and Frigessi (1989) and construct antithetic variables that may achieve variance reduction in simple settings but they do not appear to be widely applicable. Andradòttir et al. (1993) focus on finite state space chains, they observe that optimum variance reduction can be achieved via the solution of the associated Poisson equation (see Section 2.1), and they propose numerical algorithms for its solution. Rao-Blackwellisation has been suggested by Gelfand and Smith (1990) and by Robert and Casella (2004) as a way to reduce the variance of MCMC estimators. Also, Philippe and Robert (2001) investigated the use of Riemann sums as a variance reduction tool in MCMC algorithms. An interesting as well as natural control variate that has been used, mainly as a convergence diagnostic, by Fan et al. (2006), is the score statistic. Although Philippe and Robert (2001) mention that it can be used as a control variate, its practical utility has not been investigated. Atchadé and Perron (2005) restrict attention to independent Metropolis samplers and provide an explicit formula for the construction of control variates based on adaptive estimators. Mira et al. (2003) note that a solution to the Poisson equation provides the optimum control variate and they attempt to solve it numerically. Hammer and Håkon (2008) construct control variates for general Metropolis-Hastings samplers by expanding the state space. To estimate the optimal coefficients $\{\theta_j^*\}$ they use the same formula that one obtains for control variates in i.i.d. Monte Carlo sampling, but such estimators are strictly suboptimal; they are briefly discussed in Section 2.3, where we also explore their efficiency.

A more relevant, for our purposes, line of work is that initiated by Henderson (1997), who observed that, for *any* real-valued function G defined on the state space of a Markov chain $\{X_n\}$, the function $U(x) := G(x) - E[G(X_{n+1})|X_n = x]$ has zero mean with respect to the stationary distribution of the chain. Henderson (1997), like some of the other authors mentioned above, also notes that the best choice for the function G would be the solution of the associated Poisson equation, and proceeds to compute approximations of this solution for specific Markov chains, with particular emphasis on models arising in stochastic network theory.

The gist of our approach is to adapt Henderson's idea for the construction of control variates, and use them in conjunction with a new, efficiently implementable and provably optimal esti-

mator for the coefficients $\{\theta_j^*\}$ for reversible chains. The ability to estimate the $\{\theta_j^*\}$ effectively makes these control variates practically relevant in the statistical MCMC context, and allows us to avoid having to compute analytical approximations to the solution of the underlying Poisson equation. Our estimator for $\{\theta_j^*\}$ is adaptive, in the sense that is based on the MCMC output, and it can be used after the sample is obtained, making its actual computation independent of the MCMC algorithm.

This methodology not only generalizes the classical method of control variates to the MCMC setting, it also offers an important advantage: Unlike the case of independent sampling where control variates need to be found in an *ad hoc* manner depending on the specific problem at hand, here the control variates (as well as estimates for the corresponding optimal coefficients) come for free. The only requirement for the application of this method at the post-processing stage is the availability of a function G of the sampled parameters, together with its one-step conditional expectation, $E[G(X_{n+1})|X_n = x]$. As we show in numerous specific examples, these are often readily available; for example, the availability of such expectations is essentially a prerequisite for Gibbs sampling.

For any one particular application, there is, of course, a plethora of functions G (and, consequently, of corresponding control variates U) that can be used, so an important consideration for the effectiveness of this methodology for variance reduction is the careful choice of these functions. This issue is addressed in detail; we provide numerous illustrative examples of estimation problems based on MCMC samplers, motivated primarily by Bayesian inference problems. These examples are chosen as representing different major classes of MCMC samplers commonly used in important applications. In each case, the ideas underlying the choice of the functions G are explained, and these choices are justified either rigorously or heuristically, in connection with the theoretical development we present.

The examples we consider range from the simplest, illustrative samplers, to complex applications of Bayesian models to real data. In all cases, the resulting variance reduction is very significant and often quite large: For all the MCMC-based ergodic estimators we consider, the use of control variates gives variances at least 30 times smaller, and often hundreds or thousands of times smaller.

Presently we focus only on cases of reversible MCMC samplers for which the one-step conditional expectations, $E[G(X_{n+1})|X_n = x]$, of one or more functions G are available analytically in closed form. MCMC algorithms with this property include a vast array of samplers commonly used in practical Bayesian inference problems. In the examples presented in Sections 3 and 6 below we outline the implementation details of our methodology for a representative subset of both simple and complex models. Since our estimators for $\{\theta_j^*\}$ are applicable to reversible chains, we employ random-scan instead of the usual systematic-scan Gibbs or Metropolis-within-Gibbs algorithms. We also investigate the behavior of our estimators on discrete state space, random-walk Metropolis-Hastings samplers, and on Metropolis-within-Gibbs samplers. Although, strictly speaking, our theoretical development does not necessarily require that conditional expectations $E[G(X_{n+1})|X_n = x]$ be analytically available, almost all of the examples presented here do have that property, primarily for the sake of convenience and of clarity of exposition. Further ongoing work by Dellaportas et al. (2008) explores ways in which this same theory can be applied to arbitrary reversible MCMC samplers, including cases where one-step conditional expectations are unavailable.

As mentioned above, Henderson (1997) takes a different path toward optimizing the use of control variates for Markov chain samplers. Considering primarily continuous-time Markov pro-

cesses, an approximation G for the solution to the associated Poisson equation is derived from the so-called "heavy traffic" or "fluid model" approximations of the original process. The motivation and application of this method is primarily related to examples from stochastic networks and queueing theory. Closely related approaches are presented by Henderson and Glynn (2002) and Henderson et al. (2003), where the effectiveness of network control policies of multiclass networks is evaluated via Markovian simulation tools. There, control variates are used for variance reduction, and the optimal parameters $\{\theta_j^*\}$ are estimated via an adaptive, stochastic gradient algorithm. General convergence properties of ergodic estimators using control variates are derived by Henderson and Simon (2004), in the case when the solution to the Poisson equation (either for the original chain or for an approximating chain) is known explicitly. Kim and Henderson (2007) introduce two related adaptive methods for tuning non-linear versions of the parameters $\{\theta_j\}$, when using families of control variates that naturally admit a non-linear parameterization. After deriving asymptotic properties for these estimators, they present numerical examples for a simulation problem related to pricing derivative instruments in computational finance. In the case when the control variate U is defined in terms of a function G that can be taken as a Lyapunov function for the chain $\{X_n\}$, Meyn (2006) derives precise exponential asymptotics for the performance of estimators employing such control variates.

The rest of the paper is organized as follows. Section 2.1 gives the basic definitions that will remain in effect throughout the paper, and motivates the construction of control variates in connection with the Poisson equation. Sections 2.2, 2.3 and 2.4, building on ideas of Henderson (1997), illustrate the use of naive estimators of the optimal coefficient for a single control variate, and develop the theory for two new estimators for reversible chains. In Section 3 we investigate the impact of these estimators on variance reduction in five small MCMC examples, which are representative of a larger class of Bayesian inference problems. Section 4 discusses the effect of our estimators on bias reduction, compares the two estimators and advocates the use of one of them for general purposes. These estimators are generalized in Section 5 to the case of multiple control variates. Four more complex Bayesian inference problems that are implemented via MCMC are visited in Section 6; guidelines for constructing appropriate control variates are given, and their effects on variance reduction are illustrated. Finally, we provide theoretical justifications of our asymptotic arguments in Section 7 and conclude with a short discussion of possible further extensions in Section 8.

2   Control Variates for Markov Chains

2.1   The setting

Suppose $\{X_n\}$ is a discrete-time Markov chain with initial state $X_0 = x$, taking values in the state space X with an associated σ-algebra $\mathcal{B}$. In typical applications, X will often be a (Borel measurable) subset of $\mathbb{R}^d$ together the collection $\mathcal{B}$ of all its (Borel) measurable subsets. [More precise definitions and detailed assumptions will be given in Section 7.] The distribution of $\{X_n\}$ is described by its transition kernel, $P(x, dy)$,

$$P(x, A) := \Pr\{X_{k+1} \in A \,|\, X_k = x\}, \quad x \in \mathsf{X}, \ A \in \mathcal{B}. \tag{1}$$

It is well known that in many applications where it is desirable to compute the expectation $E_\pi(F) := \pi(F) := \int F \, d\pi$ of some function $F : \mathsf{X} \to \mathbb{R}$ with respect to some probability measure π on $(\mathsf{X}, \mathcal{B})$, it turns out that, although the direct computation of $\pi(F)$ is impossible or we

cannot even produce samples from π, we can construct an easy-to-simulate Markov chain $\{X_n\}$ which has π as its unique invariant measure. Under appropriate conditions, the distribution of $\{X_n\}$ converges to π, a fact which can be made precise in several ways. For example, writing PF for the function,

$$PF(x) := E_x[F(X_1)] := E[F(X_1)\,|\,X_0 = x], \quad x \in \mathsf{X},$$

we have that, for any initial condition x,

$$P^n F(x) := E[F(X_n)\,|\,X_0 = x] \to \pi(F), \quad \text{as } n \to \infty,$$

for an appropriate class of functions $F : \mathsf{X} \to \mathbb{R}$. Furthermore, the rate of this convergence can be quantified by the function,

$$\hat{F}(x) = \sum_{n=0}^{\infty} \Big[P^n F(x) - \pi(F) \Big], \tag{2}$$

where $\hat{F}$ is easily seen to satisfy the Poisson equation for F, namely,

$$P\hat{F} - \hat{F} = -F + \pi(F). \tag{3}$$

To see that, at least formally, simply apply P to both sides of (2) and note that the resulting series for $P\hat{F} - \hat{F}$ becomes telescoping and simplifies to $-F + \pi(F)$.

The above results describe how the distribution of X_n converges to π. In terms of estimation, the quantities of interest are the ergodic averages,

$$\mu_n(F) := \frac{1}{n} \sum_{i=0}^{n-1} F(X_i).$$

Again, under appropriate conditions the ergodic theorem holds,

$$\mu_n(F) \to \pi(F), \quad \text{a.s., as } n \to \infty, \tag{4}$$

for an appropriate class of functions F. Moreover, the rate of this convergence is quantified by an associated central limit theorem, which states that,

$$\sqrt{n}[\mu_n(F) - \pi(F)] = \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} [F(X_i) - \pi(F)] \xrightarrow{\mathcal{D}} N(0, \sigma_F^2), \quad \text{as } n \to \infty,$$

where σ_F^2, the asymptotic variance of F, is given by,

$$\sigma_F^2 := \lim_{n \to \infty} \mathrm{Var}_\pi(\sqrt{n}\mu_n(F)) = \lim_{n \to \infty} \mathrm{Var}_\pi\Big(\frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} F(X_i) \Big) = \sum_{n=-\infty}^{\infty} \mathrm{Cov}_\pi(F(X_0), F(X_n)).$$

Alternatively, it can be expressed in terms of the solution $\hat{F}$ to Poisson's equation as,

$$\sigma_F^2 = \pi\Big(\hat{F}^2 - (P\hat{F})^2 \Big). \tag{5}$$

The results in equations (2) and (5) clearly indicate that it is useful to be able to compute the solution $\hat{F}$ to the Poisson equation for F. In general this is a highly nontrivial – or impossible – task; for one thing, it requires knowledge of the mean $\pi(F)$. The following example is one of the rare cases where explicit computations are possible.

Suppose $\{X_n\}$ is a discrete time version of the Ornstein-Uhlenbeck process defined by, $X_0 = x$ and $X_{n+1} = \alpha X_n + Z_n$, where α is a constant in $(0, 1)$ and $\{Z_n\}$ are independent and identically distributed (i.i.d.) standard Normal random variables. Standard methods easily show that the distribution of X_n converges to $\pi := N(0, (1 - \alpha^2)^{-1})$, so if we take $F(x) \equiv x$, then, $\mu_n(F) \to \pi(F) = 0$ a.s., as $n \to \infty$. Moreover, the central limit theorem implies that,

$$\sqrt{n}\mu_n(F) \xrightarrow{\mathcal{D}} N(0, \sigma^2), \quad n \to \infty,$$

where $\sigma^2 = \sigma_F^2$ is given by (5). In order to compute the variance we need to know $\hat{F}$. As a first guess, we take $G(x) = cx + b$ and compute,

$$PG(x) - G(x) = E[cX_1 + b \,|\, X_0 = x] - cx - b = E[c(\alpha x + Z_1)] - cx = -c(1 - \alpha)x.$$

For this to be equal to $-F(x) + \pi(F) = -x$, we need $c = (1 - \alpha)^{-1}$; any b will do. Taking, for simplicity, $b = 0$, yields,

$$\hat{F}(x) = \frac{x}{1 - \alpha} \quad \text{and} \quad P\hat{F}(x) = \frac{\alpha x}{1 - \alpha}, \quad x \in \mathbb{R}.$$

Therefore, writing W for $N(0, (1 - \alpha^2)^{-1})$ random variable,

$$\sigma^2 = \pi\left(\hat{F}^2 - (P\hat{F})^2\right) = E\left[\frac{W^2}{(1 - \alpha)^2} - \frac{\alpha^2 W^2}{(1 - \alpha)^2}\right] = \frac{1}{(1 - \alpha)^2}.$$

2.2 Control variates

Suppose that, for some Markov chain $\{X_n\}$ with transition kernel P and invariant measure π, we use the ergodic averages $\mu_n(F)$ as in (4) to estimate the mean $\pi(F)$ of some function F under π. In many applications, although the estimates $\mu_n(F)$ converge to $\pi(F)$ as $n \to \infty$, the associated asymptotic variance σ_F^2 is large and the convergence is very slow.

In order to reduce the variance, we employ the idea of using control variates, as in the case of simple Monte Carlo with i.i.d. samples; see, for example, the standard texts Robert and Casella (2004); Liu (2001); Givens and Hoeting (2005), or the paper by Glynn and Szechtman (2002) for extensive discussions. Given a function $U : \mathsf{X} \to \mathbb{R}$ for which we know that $\pi(U) = 0$, define,

$$F_\theta = F - \theta U, \tag{6}$$

and consider the modified estimators,

$$\mu_n(F_\theta) = \mu_n(F) - \theta\mu_n(U).$$

We will concentrate exclusively on the the following class of functions U proposed by Henderson (1997). For an arbitrary $G : \mathsf{X} \to \mathbb{R}$ with $\pi(|G|) < \infty$, define,

$$U = G - PG.$$

The invariance of π under P and the integrability of G immediately imply that $\pi(U) = 0$. [See Section 7 for the details, complete assumptions, and full, rigorous results corresponding to this discussion.] Therefore, the ergodic theorem guarantees that the $\{\mu_n(F_\theta)\}$ are consistent with probability one, and it is natural to seek particular choices for U and θ so that the asymptotic variance $\sigma^2_{F_\theta}$ of the modified estimators is significantly smaller that the variance σ^2_F of the standard ergodic averages $\mu_n(F)$.

Suppose, at first, that we have complete freedom in the choice of G, so that we may set $\theta = 1$ without loss of generality. Then we wish to make the asymptotic variance of,

$$F - U = F - G + PG,$$

as small as possible. But, in view of the Poisson equation (3), we see that the choice $G = \hat{F}$ yields,

$$F - U = F - \hat{F} + P\hat{F} = \pi(F), \tag{7}$$

which has zero variance. Therefore, our first rule of thumb for choosing G is:

Choose a control variate $U = G - PG$ with $G \approx \hat{F}$.

As mentioned above, it is typically impossible to compute $\hat{F}$ for realistic models in applications. But it is often possible to come up with a guess G that approximates $\hat{F}$, or at least some G for which heuristics indicate that it would be useful as a control variate. Once such a function is selected, we form the modified estimators $\mu_n(F_\theta)$ with respect to the function F_θ as in (6),

$$F_\theta = F - \theta U = F - \theta G + \theta PG.$$

The next task is to choose θ so that the resulting variance,

$$\sigma^2_\theta := \sigma^2_{F_\theta} = \pi\Big(\hat{F}^2_\theta - (P\hat{F}_\theta)^2\Big),$$

is minimized. Note that, from the definitions,

$$\hat{U} = G \quad \text{and} \quad \hat{F}_\theta = \hat{F} - \theta G. \tag{8}$$

Therefore,

$$\sigma^2_\theta = \pi\Big((\hat{F} - \theta G)^2\Big) - \pi\Big((P\hat{F} - \theta PG)^2\Big).$$

Expanding the above expression as a quadratic in θ, the optimal value for θ is determined as,

$$\theta^* = \frac{\pi\big(\hat{F}G - (P\hat{F})(PG)\big)}{\pi(G^2 - (PG)^2)}. \tag{9}$$

Note that, since $\hat{U} = G$, the denominator is simply σ^2_U. Once again, this expression depends on $\hat{F}$, so it is not immediately clear how to estimate θ^* directly from the data $\{X_n\}$. We consider the issue of estimating θ^* in detail below, but first let us interpret the optimal value of θ^*. Starting from the expression,

$$\sigma^2_\theta = \lim_{n\to\infty} \text{Var}_\pi\Big(\frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} [F(X_i) - \theta U(X_i)]\Big),$$

7

simple calculations lead to,

$$\sigma_\theta^2 = \sigma_F^2 + \theta^2 \sigma_U^2 - 2\theta \sum_{n=-\infty}^{\infty} \mathrm{Cov}_\pi(F(X_0), U(X_n)),$$

so that θ^* can also be expressed as

$$\theta^* = \frac{1}{\sigma_U^2} \sum_{n=-\infty}^{\infty} \mathrm{Cov}_\pi(F(X_0), U(X_n)), \tag{10}$$

leading to the optimal asymptotic variance,

$$\sigma_{\theta^*}^2 = \sigma_F^2 - \frac{1}{\sigma_U^2} \Big[\sum_{n=-\infty}^{\infty} \mathrm{Cov}_\pi(F(X_0), U(X_n)) \Big]^2. \tag{11}$$

Therefore, in order to reduce the variance, we want to have the covariance between F and U to be as large as possible. This leads to our second rule of thumb for selecting control variates:

Choose a control variate $U = G - PG$ so that U and F are highly correlated.

Incidently, note that, since the denominator of (9) equals σ_U^2, comparing the expressions for θ^* in (9) and (11) we see that,

$$\sum_{n=-\infty}^{\infty} \mathrm{Cov}_\pi(F(X_0), U(X_n)) = \pi\big(\hat{F}G - (P\hat{F})(PG)\big). \tag{12}$$

Moreover, the fact that σ_U^2 is always nonnegative, suggests that there should be a way to rewrite the expression $\pi(G^2 - (PG)^2)$ in the denominator of θ^* in a way which makes this nonnegativity obvious. Indeed:

Lemma 1. The asymptotic variance σ_U^2 of the function $U = G - PG$ can be expressed as,

$$\sigma_U^2 = \pi(G^2 - (PG)^2) = E_\pi\Big[\big(G(X_1) - PG(X_0)\big)^2\Big]. \tag{13}$$

PROOF. Starting from the right-hand side of (13),

$$\begin{aligned}
E_\pi\Big[\big(G(X_1) - PG(X_0)\big)^2\Big] &= \pi(G^2) - 2E_\pi[G(X_1)PG(X_0)] + \pi((PG)^2) \\
&= \pi(G^2) - 2E_\pi\Big\{E\Big[G(X_1)PG(X_0) \,\Big|\, X_0\Big]\Big\} + \pi((PG)^2) \\
&= \pi(G^2) - 2E_\pi\Big[E[G(X_1)\,|\,X_0]PG(X_0)\Big] + \pi((PG)^2) \\
&= \pi(G^2 - (PG)^2),
\end{aligned}$$

as claimed. The fact that $\sigma_U^2 = \pi(G^2 - (PG)^2)$ is immediate upon noting that $\hat{U} = G$. $\quad\square$

In view of Lemma 1, θ^* can also be expressed as,

$$\theta^* = \frac{\pi\big(\hat{F}G - (P\hat{F})(PG)\big)}{E_\pi\Big[\big(G(X_1) - PG(X_0)\big)^2\Big]}. \tag{14}$$

8

2.3 A suboptimal empirical estimate of θ^*

Let $\{X_n\}$ be a Markov chain with transition kernel P and invariant measure π. In order to estimate the mean $\pi(F)$ of some function F under π, we replace the ergodic averages $\mu_n(F)$ of (4) by the modified estimates $\mu_n(F_\theta) = \mu_n(F) - \theta\mu_n(U)$, where the control variate $U := G - PG$ for some fixed function G, which, we hope, approximates the solution $\hat{F}$ to the Poisson equation for F, or, at least, is strongly correlated with F. In order to select a "good" value for the coefficient θ – a value that leads to a relatively small asymptotic variance for the estimates $\mu_n(F_\theta)$ – we first consider the following simplistic scheme.

Pretending momentarily that $\{X_n\}$ is a sequence of i.i.d. samples with distribution π, then $\hat{F} = F$ and the optimal coefficient choice for θ becomes,

$$\theta_{\text{iid}}^* = \frac{\text{Cov}_\pi(F, G)}{\text{Var}_\pi(G)} = \frac{\text{Cov}_\pi(F, U)}{\text{Var}_\pi(U)},$$

which can be adaptively estimated by,

$$\hat{\theta}_{n,\text{iid}} = \frac{\mu_n(FU)}{\mu_n(U^2)}.$$

This leads us to the usual adaptive estimator for $\pi(F)$, commonly used in the case of i.i.d. samples,

$$\mu_{n,\text{iid}}(F) := \mu_n(F_{\hat{\theta}_{n,\text{iid}}}) = \mu_n\Big(F - \frac{\mu_n(FU)}{\mu_n(U^2)}U\Big) = \mu_n(F) - \frac{\mu_n(U)\mu_n(FU)}{\mu_n(U^2)}.$$

To examine its performance when used on samples from a Markov chain, we consider an example.

Example 1. A simple Gibbs sampler. Let $\pi(x, y)$ be a bivariate Normal distribution with zero mean, unit variances, and covariance $\rho > 0$. We use the systematic-scan Gibbs sampler to simulate from π. Starting from arbitrary $X_0 = x$ and $Y_0 = y$, X_1 is generated by sampling from $\pi(x|y) \sim N(\rho y, (1-\rho^2))$, and then Y_1 is generated by sampling from $\pi(y|X_1) \sim N(\rho X_1, (1-\rho^2))$. Continuing this way produces a Markov chain $\{(X_n, Y_n)\}$ with distribution converging to π.

Suppose we wish to estimate the expected value of X^2 under π. Letting $F(x, y) = x^2$, the standard estimates $\mu_n(F) \to \pi(F) = E_\pi(X^2) = 1$ a.s., but when ρ is close to 1 the variance is high and the convergence very slow. In this particularly simple example, we can actually solve the Poisson equation for F. Since F is quadratic, we consider a candidate solution of the form $G(x, y) = bx^2 + cy^2$. Direct calculation shows that,

$$PG(x, y) - G(x, y) = -bx^2 + (\rho^4 c + \rho^2 b - c)y^2 + (\rho^2 c + b + c)(1 - \rho^2),$$

and for this to be identically equal to $-F(x, y) + \pi(F) = -x^2 + 1$, it suffices to take $b = 1$ and $c = \rho^2(1 - \rho^4)^{-1}$. Therefore,

$$\hat{F}(x, y) = x^2 + \rho^2(1 - \rho^4)^{-1}y^2.$$

From this we can compute the asymptotic variance σ_F^2 of the estimates $\mu_n(F)$ by substituting $\hat{F}$ in (5), to obtain,

$$\sigma_F^2 = E_\pi[(X^2 + cY^2)^2 - (1 + cY^2)^2] = 2(1 + \rho^4)(1 - \rho^4)^{-1},$$

which is indeed high for $\rho \sim 1$.

And now suppose that, as is typically the case in applications, $\pi(x, y)$ is not available and we cannot obtain an explicit solution for $\hat{F}$. In order to create a control variate U it is natural to start with $G = F$ itself, since we certainly expect that $U = F - PF$ will be strongly correlated with F. But F only depends on x, so, in order to take advantage of the fact that we also produce samples for y, we let $G(x, y) = F(x, y) + F(y, x) = x^2 + y^2$ and define the control variate,

$$U(x, y) = G(x, y) - PG(x, y) = x^2 + (1 - \rho^2 - \rho^4)y^2 - (2 - \rho^2 - \rho^4).$$

We will now compare the performance of three estimators: (i) The standard estimator $\mu_n(F)$; (ii) The suboptimal adaptive estimator $\mu_{n,\mathrm{iid}}(F)$ based on the control variate U defined above; and (iii) The optimal estimator $\mu_n(F_{\theta^*})$ based on the same control variate U, but with respect to the optimal value of θ^*.

Since in this case we know both π and $\hat{F}$ explicitly, for the sake of comparison we compute the theoretically optimum value of θ appearing in (9) as,

$$\theta^* = \frac{1 + 3\rho^2 + 2\rho^4}{(1 - \rho^4)(2 + 4\rho^2 + 3\rho^4 + \rho^6)}.$$

Figure 1 shows a typical realization of the performance of all three estimators for the following parameter values: The correlation $\rho = 0.95$, the number of steps $n = 5000$, the initial values are $x_0 = y_0 = 0$, and the optimal value of $\theta^* \approx 3.273$. In this experiment, the (estimated) variance of the adaptive estimator is smaller than that of the standard estimator by a factor of ≈ 3.13; whereas the variance of the optimal estimator is smaller than that of the standard estimator by a factor of ≈ 18.52.
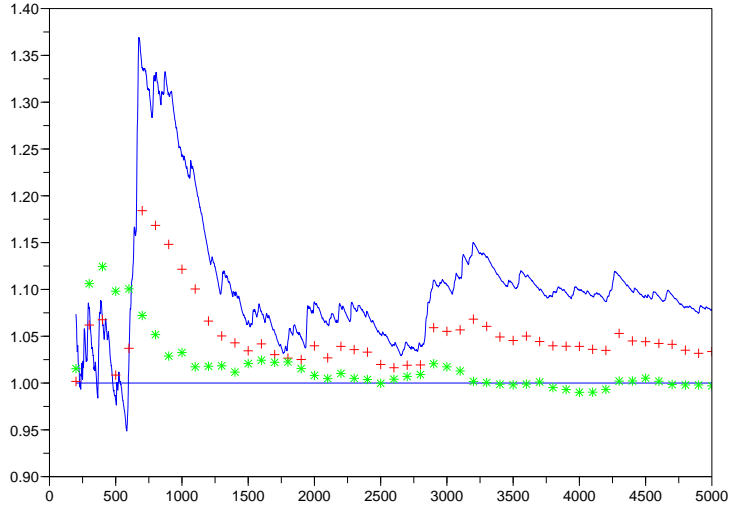

Figure 1: The sequence of the standard ergodic averages is shown as a solid blue line; the suboptimal adaptive estimates $\mu_{n,\mathrm{iid}}(F)$ as red "+" signs; and the optimal estimates $\mu_n(F_{\theta^*})$ are green "*" signs. For visual clarity, the estimates $\mu_{n,\mathrm{iid}}(F)$ and $\mu_n(F_{\theta^*})$ are plotted only every 100 simulation steps.

The reduction in the variance was computed from $T = 100$ independent repetitions of the same experiment. For $\mu_n(F)$, we obtained $T = 100$ different estimates $\mu_n^{(i)}(F)$, for $i = 1, 2, \ldots, T$, and the variance of $\mu_n(F)$ was estimated by,

$$\frac{1}{T-1} \sum_{i=1}^{T} [\mu_n^{(i)}(F) - \bar{\mu}_n(F)]^2, \tag{15}$$

where $\bar{\mu}_n(F)$ is the average of the $\mu_n^{(i)}(F)$. The same procedure was applied to estimate the variance of $\mu_{n,\text{iid}}(F)$ and $\mu_n(F_{\theta^*})$. The factors by which the variance of $\mu_n(F)$ is larger than that of $\mu_{n,\text{iid}}(F)$ and $\mu_n(F_{\theta^*})$, respectively, are shown in Table 1.

Variance reduction factors				
	Simulation steps			
Estimator	$n = 1000$	$n = 2000$	$n = 5000$	$n = 10000$
$\mu_{n,\text{iid}}(F)$	3.16	2.96	3.13	3.01
$\mu_n(F_{\theta^*})$	18.28	16.43	18.52	16.07

Table 1: Estimated factors by which the variance of $\mu_n(F)$ is larger than that of $\mu_{n,\text{iid}}(F)$ and $\mu_n(F_{\theta^*})$, respectively, after $n = 1000, 2000, 5000$ and 10000 simulation steps.

For different values of the number of iterations n, the corresponding variance reduction factors were computed based on independent runs, and are not continuations of shorter runs. Note that the adaptive estimates $\mu_{n,\text{iid}}(F)$ were actually computed in two steps: First the value for the coefficient $\hat{\theta}_{n,\text{iid}}$ was computed, and then the values $\mu_{n,\text{iid}}(F)$ were calculated. In both passes, the same simulation samples were used. We emphasize that *this procedure is used throughout the paper*. Indeed, the fact that the estimators can be computed after the MCMC sample has been obtained is a major advantage of our methodology.

Clearly, although the adaptive estimator $\mu_{n,\text{iid}}(F)$ does offer a significant advantage over $\mu_n(F)$, there is a lot to be gained from obtaining more accurate estimates of the optimal coefficient θ^*. We remark that, instead of treating the samples $\{X_n\}$ as being i.i.d., more accurate estimates for θ^* can be obtained by approximating the expression (10) via averages over blocks. Nevertheless, extensive simulation experiments clearly indicate that the corresponding estimation gains are usually negligible, while the optimal, consistent estimation procedures for θ^* given in the following section make a very significant difference.

2.4   Optimal empirical estimates of θ^* for reversible chains

Let $\Delta = P - I$ denote the generator of a discrete time Markov chain $\{X_n\}$ with transition kernel P. If the chain is reversible, then Δ is a self-adjoint linear operator on the space $L_2(\pi)$. This simply means that,

$$\pi(F \, \Delta G) = \pi(\Delta F \, G),$$

for any two functions $F, G \in L_2(\pi)$. Our central result in terms of the estimation methodology is the observation that, in this case, the optimal coefficient θ^* admits a representation that does not involve the solution to Poisson's equation $\hat{F}$:

Proposition 1. If the chain $\{X_n\}$ is reversible, then the optimal coefficient θ^* for the control variate $U = G - PG$ can be expressed as,

$$\theta^* = \theta_{\text{rev}}^* := \frac{\pi\big((F - \pi(F))(G + PG)\big)}{\pi(G^2 - (PG)^2)}, \tag{16}$$

or, alternatively,

$$\theta^* = \theta_{\text{rev}}^* := \frac{\pi\big((F - \pi(F))(G + PG)\big)}{E_\pi\Big[\big(G(X_1) - PG(X_0)\big)^2\Big]}. \tag{17}$$

PROOF. Let $\bar{F} = F - \pi(F)$ denote the centered version of F, and recall that $\hat{F}$ solves Poisson's equation for F, so $P\hat{F} = \hat{F} - \bar{F}$. Therefore, the numerator in the expression for θ^* in (9) can be expressed as,

$$
\begin{aligned}
\pi\big(\hat{F}G - (P\hat{F})(PG)\big) &= \pi\big(\hat{F}G - (\hat{F} - \bar{F})(PG)\big) \\
&= \pi\big(\bar{F}PG - \hat{F}\Delta G\big) \\
&= \pi\big(\bar{F}PG - \Delta\hat{F}G\big) \\
&= \pi\big(\bar{F}PG + \bar{F}G\big) \\
&= \pi\big(\bar{F}(G + PG)\big).
\end{aligned}
$$

This proves (16), and (17) follows from (14). $\qquad\square$

The expressions (16) and (17) immediately suggest estimating θ^* via,

$$
\begin{aligned}
\hat{\theta}_{n,\text{rev},1} &= \frac{\mu_n(F(G + PG)) - \mu_n(F)\mu_n(G + PG)}{\mu_n(G^2) - \mu_n((PG)^2)} \\
\text{or} \quad \hat{\theta}_{n,\text{rev},2} &= \frac{\mu_n(F(G + PG)) - \mu_n(F)\mu_n(G + PG)}{\frac{1}{n}\sum_{i=0}^{n-1}(G(X_i) - PG(X_{i-1}))^2}.
\end{aligned}
$$

The resulting estimators, $\mu_n(F_{\hat{\theta}_{n,\text{rev},1}})$ and $\mu_n(F_{\hat{\theta}_{n,\text{rev},2}})$ for $\pi(F)$ based on the control variate $U = G - PG$ and the coefficients $\hat{\theta}_{n,\text{rev},1}$ and $\hat{\theta}_{n,\text{rev},2}$, respectively, are denoted,

$$
\begin{aligned}
\mu_{n,\text{rev},1}(F) &:= \mu_n(F_{\hat{\theta}_{n,\text{rev},1}}) = \mu_n(F - \hat{\theta}_{n,\text{rev},1}U) \\
\text{and} \quad \mu_{n,\text{rev},2}(F) &:= \mu_n(F_{\hat{\theta}_{n,\text{rev},2}}) = \mu_n(F - \hat{\theta}_{n,\text{rev},2}U).
\end{aligned}
$$

An alternative way for estimating θ^* adaptively, which also applies to non-reversible chains, was recently developed in Meyn (2007), based on the "temporal difference learning" algorithm. As most of the chains we will consider are reversible and this alternative method is computationally significantly more expensive than our estimates $\hat{\theta}_{n,\text{rev},1}$ and $\hat{\theta}_{n,\text{rev},2}$, it will not be considered further in the present discussion.

A slightly more general case of the earlier example with a bivariate Gaussian density is considered below; the random-scan Gibbs sampler is used to examine the performance of the two new estimators. We note that although the systematic-scan Gibbs sampler in general does not produce a reversible chain, the random-scan Gibbs sampler always does. Also, the back-and-forth version of the systematic-scan Gibbs sampler is reversible, see Roberts (1992).

Example 2. The bivariate Gaussian through the random-scan Gibbs sampler. Let $(X, Y) \sim \pi(x, y)$ be an arbitrary bivariate Normal distribution, where, without loss of generality, we take the expected values of both X and Y to be zero and the variance of X to be equal to one. Let $\mathrm{Var}(Y) = \tau^2$ and the covariance $E(XY) = \rho\tau$ for some $\rho \in (-1, 1)$. Given arbitrary initial values $x_0 = x$ and $y_0 = y$, the Gibbs sampler selects one of the two co-ordinates at random, and either updates y by sampling from $\pi(y|x) \sim N(\rho\tau x, \tau^2(1-\rho^2))$, or x from $N(\frac{\rho}{\tau}y, 1-\rho^2)$. Continuing this way produces a reversible Markov chain $\{(X_n, Y_n)\}$ with distribution converging to π.

To estimate the expected value of X under π, we let $F(x, y) = x$ and $G(x, y) = x + y$, so that,

$$PG(x, y) = \tfrac{1}{2}(1 + \rho\tau)x + \tfrac{1}{2}\left(1 + \frac{\rho}{\tau}\right)y,$$

and the control variate $U = G - PG$ is, $U(x, y) = G(x, y) - PG(x, y) = \frac{1}{2}(1 - \rho\tau)x + \frac{1}{2}(1 - \frac{\rho}{\tau})y$. We will compare the performance of five estimators: (i) The standard estimator $\mu_n(F)$; (ii) The suboptimal adaptive estimator $\mu_{n,\mathrm{iid}}(F)$ based on the control variate $U = G - PG$ defined above; (iii, iv) The two adaptive estimators $\mu_{n,\mathrm{rev},1}(F)$ and $\mu_{n,\mathrm{rev},2}(F)$ for the same control variate U; (v) The *optimal* estimator $\mu_n(F_{\theta^*})$ based on the same control variate, but with respect to the optimal value of θ^*.

In Figure 2 we plot the results of all five estimators, applied to a typical execution of the Gibbs sampler with $n = 5000$ steps and initial values $x_0 = y_0 = 0.1$. The problem parameter values are $\rho = 0.99$ and $\tau^2 = 10$.
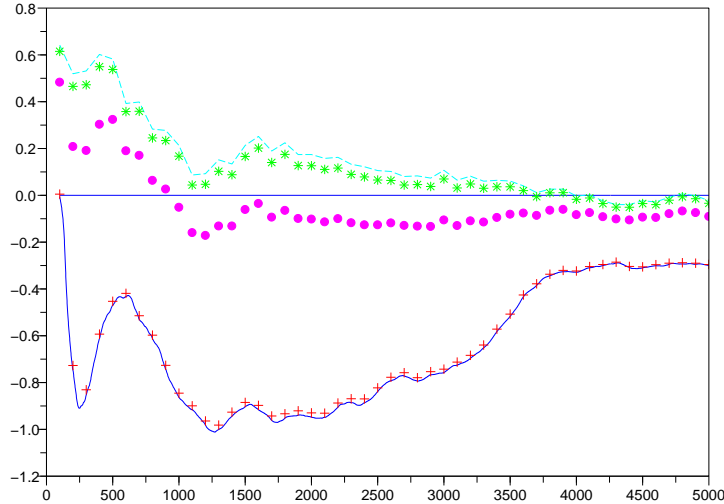

Figure 2: The sequence of the standard ergodic averages $\mu_n(F)$ is shown as a solid blue line; the suboptimal adaptive estimates $\mu_{n,\mathrm{iid}}(F)$ as red "+" signs; the optimal adaptive estimates $\mu_{n,\mathrm{rev},1}(F)$ as bold magenta dots, $\mu_{n,\mathrm{rev},2}(F)$ as a dashed cyan line, and the estimates $\mu_n(F_{\theta^*})$ corresponding to the optimal value of θ^* as green "*" signs. For visual clarity, all estimates except $\mu_n(F)$ are plotted only every 100 simulation steps.

While the optimal estimator $\mu_n(F_{\theta^*})$ offers an obviously large advantage over the standard estimates $\mu_n(F)$, the improvement of the suboptimal estimator $\mu_{n,\mathrm{iid}}(F)$ is rather insignificant.

The adaptive estimators $\mu_{n,\text{rev},1}(F)$ and $\mu_{n,\text{rev},2}(F)$ are similarly very effective, and their performance is fairly close to that of the optimal $\mu_n(F_{\theta^*})$. As in Example 1, we compute the factor by which each of these methods reduces the variance of the standard estimates $\mu_n(F)$, using $T = 200$ independent repetitions of the same experiment; recall equation (15) above. The results are shown in Table 2.

Variance reduction factors					
	Simulation steps				
Estimator	$n = 1000$	$n = 5000$	$n = 10000$	$n = 50000$	$n = 100000$
$\mu_{n,\text{iid}}(F)$	1.04	1.03	1.02	1.02	1.02
$\mu_{n,\text{rev},1}(F)$	2.14	6.25	6.77	8.26	7.50
$\mu_{n,\text{rev},2}(F)$	2.79	5.66	6.58	8.19	7.54
$\mu_n(F_{\theta^*})$	5.23	9.12	8.20	8.25	7.53

Table 2: Estimated factors by which the variance of $\mu_n(F)$ is larger than that of $\mu_{n,\text{iid}}(F)$, $\mu_{n,\text{rev},1}(F)$ $\mu_{n,\text{rev},2}(F)$ and $\mu_n(F_{\theta^*})$, respectively, after $n = 1000, 5000, 10000, 50000$ and 100000 simulation steps.

Clearly, both adaptive estimators $\mu_{n,\text{rev},1}(F)$, $\mu_{n,\text{rev},2}(F)$ perform very well, and their results are reasonably close to those of the optimal estimator $\mu_n(F_{\theta^*})$. A natural way to attempt to obtain an even greater improvement in terms of variance reduction would be to consider a control variate U based on a G of the form $G(x, y) = ax + by$, for coefficients $a \neq b$. But there is no obvious choice for the relationship between these two coefficients, and also we do not want to develop methods that are too model-specific. A generic way to address such problems is to consider two control variates U_1, U_2, based on the two different functions $G_1(x, y) = x$ and $G_2(x, y) = y$. The corresponding methodology for such cases is developed in Section 5, where we also revisit this example.

Another well-known difficulty with the standard estimates in this example (in addition to their high variance) is that they converge very slowly when the initial values of the Gibbs sampler are far from their mean. The above results are from simulations with $x_0 = y_0 = 0.1$, and we also run several examples with different initial values. In those cases we found that a lot of the time the estimators $\mu_{n,\text{rev},1}(F)$ and $\mu_{n,\text{rev},2}(F)$ not only reduced the variance, but also greatly improved the bias. An example with initial values $x_0 = 4$ and $y_0 = 12$ is shown in Figure 3. A more detailed discussion of this issue will be given in Section 4, where we also address the question of choosing between the two adaptive estimators, $\mu_{n,\text{rev},1}(F)$ and $\mu_{n,\text{rev},2}(F)$.

3 Five Simple MCMC Examples

Below we present five examples more closely motivated by problems in statistical inference. Among the vast array of simple MCMC samplers that can be used for illustration purposes, we have chosen a set of representative examples that cover a broad class of real applications. The Gaussian-Gamma posterior in Example 3, as well as the the bivariate Gaussian density of Example 2, are representative of the large class of normal hierarchical models that are analyzed in Gelfand et al. (1990). Similarly, the Gibbs sampler of Example 4 is seen as a simplistic version of a wide class of models that include discrete variables as latent variable indicators or model indices. The discrete state space random-walk Metropolis algorithm in Example 5 is used in model search algorithms in which analytical or approximate integration of all model parameters
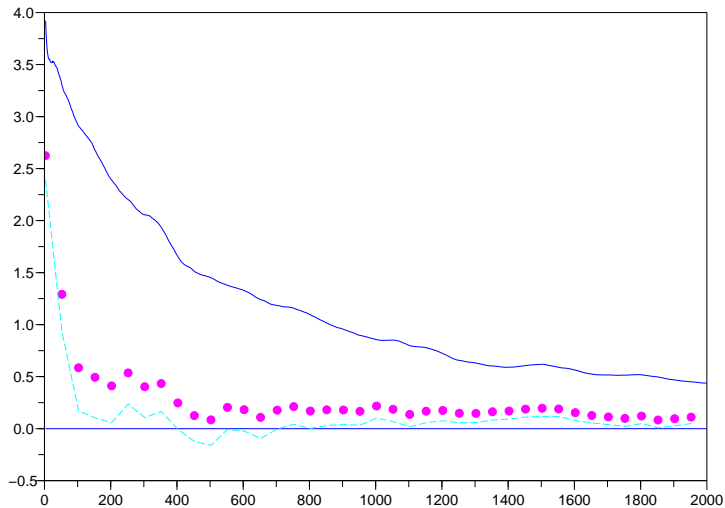
Figure 3: The sequence of the standard ergodic averages is shown as a solid blue line; the adaptive estimates $\mu_{n,\mathrm{rev},1}(F)$, plotted only every 50 simulation steps, are shown as bold magenta dots, and the adaptive estimates $\mu_{n,\mathrm{rev},2}(F)$ as a dashed cyan line.

is first performed; see for example Clyde and George (2004). A simple version of a finite-mixtures mode of Normals is explored in detail in Example 6. This this class of models has been, and still is, one of the most challenging inference problems. Finally, we illustrate our methodology in the case of a "difficult" model where Cauchy priors result in heavy-tailed posterior densities; such densities are commonly met in, for example, spatial statistics; see Dellaportas and Roberts (2003) for an illustrative example.

Example 3. A Gaussian-Gamma posterior. First we consider an example of the random-scan Gibbs sampler applied to simple Bayesian inference problem. The model is a simple two-parameter example of Gilks (1986), in which we begin with observations $x = (x_1, x_2, \ldots, x_N)$ that are independently generated from a $N(\mu, \gamma^{-1})$ distribution, and we place priors $\mu \sim N(0, 1)$ and $\gamma \sim \mathrm{Gamma}(2, 1)$ on the parameters μ and γ, respectively. [Throughout the paper, the paramatrization of the $\mathrm{Gamma}(a, b)$ density is chosen so that it has mean a/b.] It is straightforward that Gibbs sampling from the posterior $\pi(\mu, \gamma | x)$ proceeds by updating μ given γ from a Normal density with mean $(\gamma \sum_i x_i)/(1 + N\gamma)$ and variance $1/(1 + N\gamma)$, and γ given μ from a Gamma density with index $2 + N/2$ and scale $1 + \frac{1}{2} \sum_i (x_i - \mu)^2$. In our simulation, we assume that $N = 10$ and that the data vector $x = (x_i)$ is given by $x = (-23, 27, 12, 17, -8, 2, -18, 17, 7, -33)$, so that the sample mean is zero. We wish to estimate the posterior mean of μ, so we let $F(\mu, \gamma) = \mu$. Although in general this posterior mean is not computable in closed form, here the posterior marginal density of μ is proportional to the product of a Student's t density with mean zero (because the sample mean of x is zero) and the prior $N(0, 1)$ density. Therefore, the resulting density is symmetric around zero, which implies that the posterior mean of μ is actually zero. We compare the performance of the simple empirical averages $\mu_n(F)$ with the adaptive estimators $\mu_{n,\mathrm{rev},1}(F)$ and $\mu_{n,\mathrm{rev},2}(F)$, based on the control variate $U = G - PG$ with $G(\mu, \gamma) = \mu$.

Figure 4 shows a typical random-scan Gibbs sampling run of length $n = 5000$, with starting
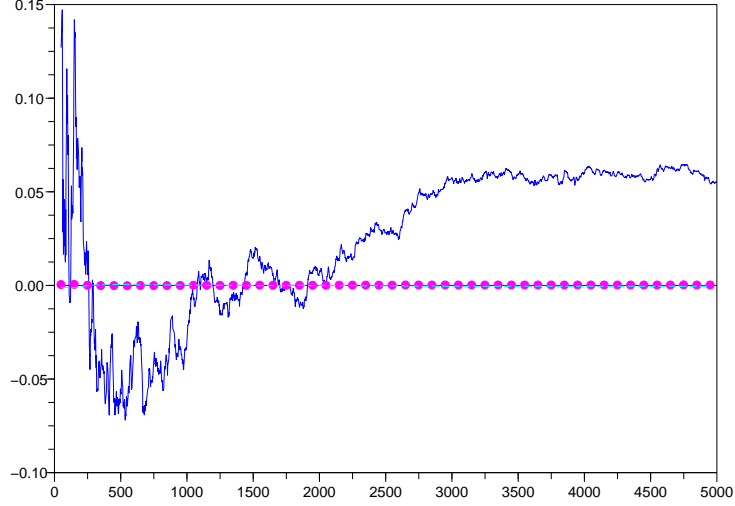
15

Figure 4: The sequence of the standard ergodic averages is shown as a solid blue line; the adaptive estimates $\mu_{n,\mathrm{rev},1}(F)$ as bold magenta dots; and the adaptive estimates $\mu_{n,\mathrm{rev},2}(F)$ as a cyan dashed line. For visual clarity, the values $\mu_{n,\mathrm{rev},1}(F)$ are plotted only every 100 simulation steps.

values $\mu_0 = \gamma_0 = 1$. It is obvious from the plot that both adaptive estimators converge incredibly fast compared to the standard ergodic averages $\mu_n(F)$. The corresponding variance reduction factors, computed from $T = 100$ repetitions of the same experiment, are shown in Table 3.

Variance reduction factors				
	Simulation steps			
Estimator	$n = 1000$	$n = 5000$	$n = 10000$	$n = 50000$
$\mu_{n,\mathrm{rev},1}(F)$	9403	341095	419766	20453186
$\mu_{n,\mathrm{rev},2}(F)$	713	1880	5287	15495

Table 3: Estimated factors by which the variance of $\mu_n(F)$ is larger than that of $\mu_{n,\mathrm{rev},1}(F)$ and $\mu_{n,\mathrm{rev},2}(F)$, after $n = 1000, 5000, 10000$, and 50000 simulation steps.

Given the tremendous effectiveness of the control variate $U = G - PG$ with $G(\mu, \gamma) = \mu$, it is natural to ask if perhaps a multiple of G actually solves the Poisson equation, that is, if $\theta^*(G - PG) = F - \pi(F)$. Since $\pi(F) = 0$ and in the simulation experiments both $\hat{\theta}_{n,\mathrm{rev},1}$ and $\hat{\theta}_{n,\mathrm{rev},2}$ apparently converge to values very close to 2, we examine the relationship, $2(G - PG) = \mu$. Substituting the expressions for G and PG this becomes,

$$\frac{\gamma \sum_i x_i}{1 + N\gamma} \stackrel{?}{=} 0, \tag{18}$$

which, in our case, is indeed an equality, since the empirical mean of our sample x is equal to zero. More generally, this will be an approximate equality (at least for most of the relevant values of μ and γ), as long the empirical mean of the sample is close to zero, or if most of the

16

mass of the posterior on γ is concentrated near zero. In fact, a multiple of $G(\mu, \gamma) = \mu$ will always solve the Poisson equation exactly, as long as the mean of the Gaussian prior on μ instead of zero is taken to be equal to the sample mean of x.

The above discussion explains the effectiveness of the control variate $U = G - PG$ with $G(\mu, \gamma) = \mu$, but it also suggests that if the number of observations N is small, and either: (a) the sample mean of the observations x is not close to zero; or (b) the empirical standard deviation of x is not appropriately "small" (in other words, the posterior on γ is not concentrated near zero); then this G would not be an approximate solution to the Poisson equation, and the corresponding control variate U would be much less effective. Nevertheless, even in the unlikely scenario where the observation vector is $x' = (x_i') = (4.75, 5.09, 4.63, 4.73, 5.08, 4.47, 5.24, 5.06, 4.98, 5.21)$, using the same control variate as before is quite effective; see the corresponding results in Table 4.

Variance reduction factors						
	Simulation steps					
Estimator	$n = 1000$	$n = 5000$	$n = 10000$	$n = 50000$	$n = 100000$	$n = 200000$
$\mu_{n,\mathrm{rev},1}(F)$	0.02	0.01	0.00	0.44	3.43	3.87
$\mu_{n,\mathrm{rev},2}(F)$	0.37	4.46	4.17	8.81	7.88	6.50

Table 4: Estimated factors by which the variance of $\mu_n(F)$ is larger than that of $\mu_{n,\mathrm{rev},1}(F)$ and $\mu_{n,\mathrm{rev},2}(F)$, after $n = 1000, 5000, 10000, 50000, 100000$ and 200000 simulation steps. Here the vector x of observations has sample mean 4.924 and sample standard deviation ≈ 0.068.

The reason this scenario is referred to as being "unlikely" is because the set of observations x' was actually obtained as an i.i.d. sample (rounded off to two decimal places) from the $N(5, 0.09)$ density; it has sample mean equal to 4.924, and sample variance ≈ 0.068. Therefore, having a $N(0, 1)$ prior on the mean of the observations x' that actually vary between 4.47 and 5.24 is an unreasonable choice. Also note that both of the potential sources of concern (a), (b) above apply here. Indeed, for $\gamma = 1/0.068$, the right-hand-side of (18) is ≈ 4.9, which is certainly not close to zero. Still, using the control variate $U = G - PG$ with $G(\mu, \gamma) = \mu$ consistently yields nontrivial variance reduction factors.

Example 4. An example with a discrete variable. Next we construct a bivariate density with a discrete variable z and a continuous variable p, where $z|p \sim \mathrm{Bern}(p)$ and $p \sim \mathrm{Beta}(\alpha, \beta)$. The random-scan Gibbs sampler draws randomly from either $z|p \sim \mathrm{Bern}(p)$ or from $p|z \sim \mathrm{Beta}(\alpha + z, \beta + 1 - z)$.

We wish to estimate the mean of z, so we set $F(z, p) = z$ and examine the performance of the ergodic averages $\mu_n(F)$ and the two adaptive estimators $\mu_{n,\mathrm{rev},1}(F)$ and $\mu_{n,\mathrm{rev},2}(F)$ based on the control variate $U = G - PG$; for G we take, as before, $G(z, p) = z + p$. Figure 5 depicts a typical realization of the random-scan Gibbs sampler, with $\alpha = 2$, $\beta = 1$, starting values $z_0 = p_0 = 1/2$, and $n = 5000$ steps. Here, the true value of $\pi(z)$ is $\alpha/(\alpha + \beta) = 2/3$. The corresponding variance reduction factors, estimated from $T = 100$ repetitions of the same experiment, are shown in Table 5.

Like in Example 3, since the use of the control variate U decreases the MCMC variance dramatically, it is natural to check if perhaps a multiple of G solves the Poisson equation. Direct computation gives,

$$PG(z, p) = p + \frac{\alpha + (\alpha + \beta + 2)z}{2(\alpha + \beta + 1)},$$
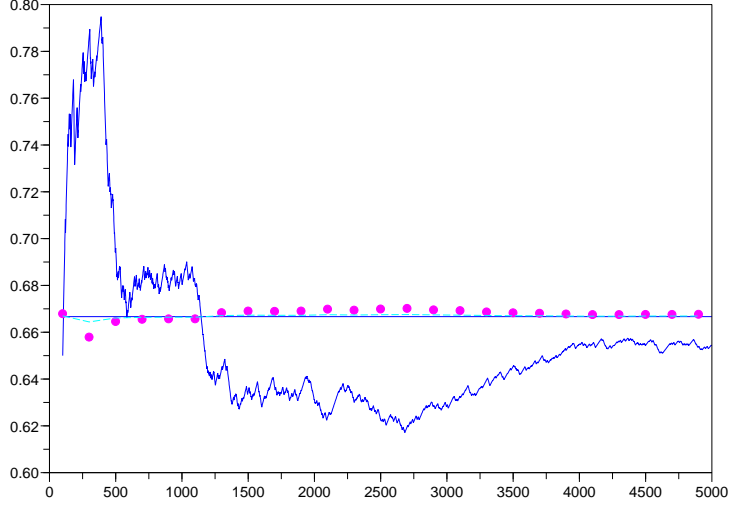
Figure 5: The sequence of the standard ergodic averages is shown as a solid blue line; the adaptive estimates $\mu_{n,\mathrm{rev},1}(F)$ as bold magenta dots; and the adaptive estimates $\mu_{n,\mathrm{rev},2}(F)$ as a cyan dashed line. For visual clarity, the values $\mu_{n,\mathrm{rev},1}(F)$ are plotted only every 200 simulation steps.

Variance reduction factors						
	Simulation steps					
Estimator	$n = 1000$	$n = 5000$	$n = 10000$	$n = 20000$	$n = 50000$	$n = 100000$
$\mu_{n,\mathrm{rev},1}(F)$	5.89	24.50	41.40	212.8	702.5	1721.3
$\mu_{n,\mathrm{rev},2}(F)$	247.4	1286.5	2145.8	4235.4	12066	24777

Table 5: Estimated factors by which the variance of $\mu_n(F)$ is larger than that of $\mu_{n,\mathrm{rev},1}(F)$ and $\mu_{n,\mathrm{rev},2}(F)$, after $n = 1000, 5000, 10000, 20000, 50000$ and 100000 simulation steps.

so that,

$$PG(z,p) - G(z,p) = \frac{\alpha + \beta}{2(\alpha + \beta + 1)}\Big[-z + \frac{\alpha}{\alpha + \beta}\Big] = \frac{\alpha + \beta}{2(\alpha + \beta + 1)}\Big[-F(z,p) + \pi(F)\Big].$$

Indeed, then, G is a multiple of the solution of the Poisson equation for F,

$$\hat{F}(z,p) = \frac{2(\alpha + \beta + 1)}{\alpha + \beta}(z + p) = \frac{2(\alpha + \beta + 1)}{\alpha + \beta}\,G(z,p),$$

and the optimal coefficient for U is,

$$\theta^* = \frac{2(\alpha + \beta + 1)}{\alpha + \beta}.$$

This explains the effectiveness of this particular choice of the function G. Incidently, it is somewhat remarkable that a multiple of the same function $G(z,p) = z + p$ solves the Poisson equation for any choice of the parameter values α, β.

Example 5. Random-walk Metropolis for Poisson generation. Consider the target distribution $\pi \sim \text{Poisson}(\lambda)$. When the mean λ is large, it is hard to sample from π directly and, instead, we consider a random-walk Metropolis sampler, which, given $X_n = x$, proposes a move to $X'_{n+1} = x + Z_n$, where the increments Z_n are i.i.d. and $Z_n = -1$ or $+1$ with probability $1/2$ each. The acceptance probability can be easily computed as,

$$\alpha(x,y) = \begin{cases} \min\{1, \frac{\lambda}{x+1}\} & \text{if } y = x+1, \\ \min\{1, \frac{x}{\lambda}\} & \text{if } y = x-1. \end{cases}$$

Suppose we wish to estimate the mean of $\sqrt{x}$ under π, so let $F(x) = \sqrt{x}$. To use a control variate $U = G - PG$ with respect to some function G on the integers, note that PG is,

$$PG(x) = G(x) + \frac{1}{2}\alpha(x, x+1)[G(x+1) - G(x)] + \frac{1}{2}\alpha(x, x-1)[G(x-1) - G(x)],$$

so that, in particular, taking $G(x) = x$, we have,

$$U(x) = \tfrac{1}{2}\alpha(x, x-1) - \tfrac{1}{2}\alpha(x, x+1).$$

Figure 6 shows a typical realization of the Metropolis sampler, using $G(x) = x$, with initial value $x_0 = 95$, for $n = 10000$ simulation steps. The "true" mean of $\sqrt{X}$ under π is estimated to be ≈ 9.9874, after 3 million Metropolis steps. The corresponding variance reduction factors, estimated from $T = 100$ repetitions of the same experiment, are shown in Table 6.
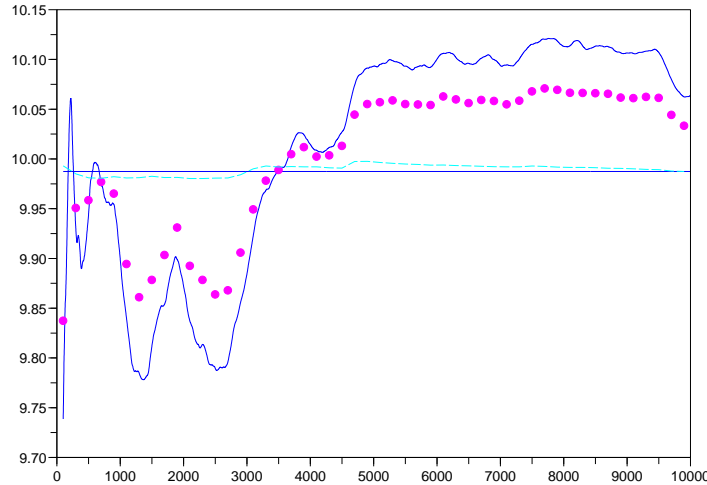

Figure 6: The sequence of the standard ergodic averages is shown as a solid blue line and the adaptive estimates $\mu_{n,\text{rev},2}(F)$ as a dashed cyan line. The adaptive estimates are plotted only every 200 simulation steps.

Example 6. A two-parameter Gaussian mixture posterior. We examine a simple Gaussian mixture example as in Robert and Casella (2004, Example 9.2). Suppose $x = (x_1, x_2, \dots, x_N)$ are independent observations from the mixture $pN(\mu_1, \sigma^2) + (1-p)N(\mu_2, \sigma^2)$; the mixing proportion p and the variance σ^2 are assumed fixed and known, and $N(0, 10\sigma^2)$

19

Variance reduction factors				
	Simulation steps			
Estimator	$n = 1000$	$n = 10000$	$n = 50000$	$n = 100000$
$\mu_{n,\mathrm{rev},1}(F)$	0.91	0.32	0.15	0.20
$\mu_{n,\mathrm{rev},2}(F)$	4.73	39.19	157.5	239.98

Table 6: Estimated factors by which the variance of $\mu_n(F)$ is larger than the corresponding variance of $\mu_{n,\mathrm{rev},1}(F)$ and $\mu_{n,\mathrm{rev},2}(F)$, respectively, after $n = 1000, 10000, 50000$ and 100000 simulation steps.

priors are placed on the means μ_1, μ_2. To facilitate sampling from the posterior, the model can alternatively be described in terms of latent variables $Z = (Z_1, Z_2, \ldots, Z_N)$, where the Z_i are independent with distribution $P(Z_i = 1) = 1 - P(Z_i = 0) = p$, and, conditional on μ_1, μ_2 and Z, each $X_i | Z_i = 1 \sim N(\mu_1, \sigma^2)$, and $X_i | Z_i = 0 \sim N(\mu_2, \sigma^2)$.

Conditional on x and z, the parameters μ_1 and μ_2 are independent, with,

$$
\begin{aligned}
\pi(\mu_1 | x, z) &\sim N\left(\frac{\sum_j z_j x_j}{n_1 + 1/10}, \frac{\sigma^2}{n_1 + 1/10} \right), \\
\pi(\mu_2 | x, z) &\sim N\left(\frac{\sum_j (1 - z_j) x_j}{n_2 + 1/10}, \frac{\sigma^2}{n_2 + 1/10} \right),
\end{aligned}
\tag{19}
$$

respectively, where $n_1 = \sum_j z_j$ is the number of z_i that are equal to 1, and $n_2 = \sum_j (1 - z_j) = N - n_1$ is the number of z_i that are equal to zero. Also, given μ_1, μ_2 and x, the Z_i are independent, and for each $i = 1, 2, \ldots, N$,

$$
\pi(Z_i = 1 | x, \mu_1, \mu_2) = q_i := \frac{p \exp\{-(x_i - \mu_1)^2 / 2\sigma^2\}}{p \exp\{-(x_i - \mu_1)^2 / 2\sigma^2\} + (1 - p) \exp\{-(x_i - \mu_2)^2 / 2\sigma^2\}}.
\tag{20}
$$

The random-scan Gibbs sampler here draws a sample from μ_1, μ_2, or from the entire vector Z, each chosen with probability $1/3$.

We consider an example with the exact same parameter settings as in Robert and Casella (2004, Example 9.2): With $p = 0.7$, $\sigma^2 = 1$, $\mu_1 = 0$ and $\mu_2 = 2.7$, we generated $N = 500$ samples from the mixture $pN(\mu_1, \sigma^2) + (1 - p)N(\mu_2, \sigma^2)$. In order to estimate μ_1, the function F is set to be $F(\mu_1, \mu_2, Z) = \mu_1$. Using a control variate $U = G - PG$ with the simplest choice of $G = F$, yields variance reduction factors around 4, which are significantly smaller than those achieved in some of the earlier examples. For that reason, we also consider $G(\mu_1, \mu_2, Z) = \sum_i Z_i x_i$ as a different candidate G for the control variate U. In fact, we let $G = c\mu + \sum_i Z_i x_i$, and select the value of c so that a multiple of G is as close as possible to a solution of the Poisson equation, that is, $\theta(PG - G) \approx -F + \pi(F)$, for some θ; substituting the values of F, G and PG, and taking $\pi(F)$ to be equal to the prior expectation of F (namely, zero), this becomes,

$$
\sum_i x_i \left[Z_i \left(1 - \frac{c}{n_1 + 1/10} \right) - q_i \right] \approx \left(\frac{3}{\theta} - c \right) \mu_1,
\tag{21}
$$

where the (q_i) are the Bernoulli parameters of the (Z_i), given in (20). A reasonable goal here is to choose c so as to reduce the variability of the left-hand-side as much as possible, since it is not directly related to μ_1. Ideally, this would mean taking $c = n_1 + 1/10$, but since n_1

is itself random, we take c to be equal to the (prior) expectation of that expression, namely, $c = (Np + 1/10)$, so that the resulting function G is,

$$G(\mu_1, \mu_2, Z) = (Np + 1/10)\mu_1 + \sum_i Z_i x_i.$$

A typical realization of the estimates based on $n = 10000$ Gibbs steps is shown in Figure 7, and the corresponding variance reduction factors are displayed in Table 7. The initial values are $\mu_1 = 0$, $\mu_2 = 1$, and the "true" posterior mean of μ_1 is estimated to be ≈ -0.0143, after 10 million Gibbs steps.
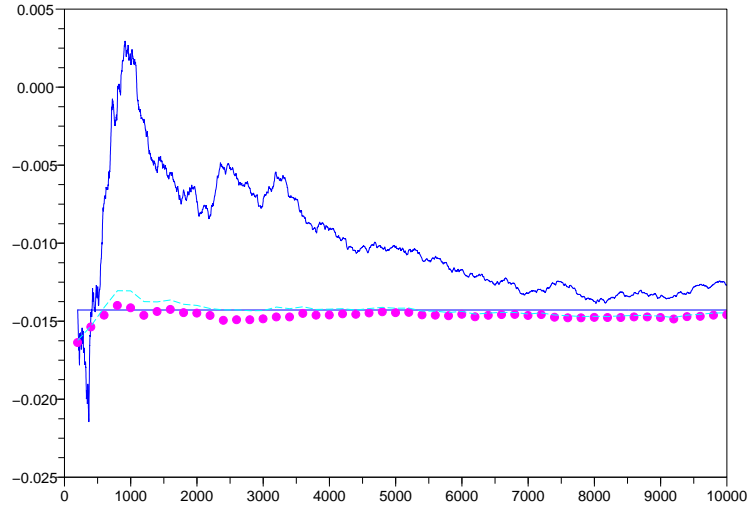

Figure 7: The sequence of the standard ergodic averages is shown as a solid blue line; the adaptive estimates $\mu_{n,\mathrm{rev},1}(F)$ as bold magenta dots; and the adaptive estimates $\mu_{n,\mathrm{rev},2}(F)$ as a cyan dashed line. For visual clarity, the values $\mu_{n,\mathrm{rev},1}(F)$ are plotted only every 200 simulation steps.

Variance reduction factors				
	Simulation steps			
Estimator	$n = 10000$	$n = 50000$	$n = 100000$	$n = 200000$
$\mu_{n,\mathrm{rev},1}(F)$	11.76	15.81	19.02	22.12
$\mu_{n,\mathrm{rev},2}(F)$	11.63	15.44	18.98	21.98

Table 7: Estimated factors by which the variance of $\mu_n(F)$ is larger than the corresponding variances of $\mu_{n,\mathrm{rev},1}(F)$ and $\mu_{n,\mathrm{rev},2}(F)$, respectively, after $n = 10000, 50000, 100000$ and 200000 simulation steps.

Incidently, the above calculation suggests that the optimal value for θ here would be the one that also makes the right-hand-side of (21) vanish, namely, $\theta^* \approx 3/c = 3/(Np + 1/10) \approx 0.009$. In our simulation experiments, the estimates of θ^* produced by both $\hat{\theta}_{n,\mathrm{rev},1}$ and $\hat{\theta}_{n,\mathrm{rev},2}$ are around 0.011, which is indeed quite close.

Finally we note that models of this type often present a difficultly, in that the posterior on (μ_1, μ_2) is bimodal. As a result, if the Gibbs sampler is initialized near the lower mode, it will never visit the neighborhood of the actual mode, at least not in any reasonable amount of time; see Robert and Casella (2004); Diebolt and Robert (1994). A more general Gaussian mixture model that at least partially addresses this issue is explored in Section 6.2.

Example 7. A Metropolis-within-Gibbs sampler. We consider an inference problem motivated by a simplified version of an example in Roberts and Rosenthal (2006). Suppose N i.i.d. observations $x = (x_1, x_2, \ldots, x_N)$ are drawn from a $N(\phi, V)$ distribution, and place independent priors $\phi \sim \text{Cauchy}(0, 1)$ and $V \sim \text{IG}(1, 1)$, on the parameters ϕ, V, respectively. The induced full conditionals of the posterior are easily seen to satisfy,

$$\pi(\phi|V, x) \quad \propto \quad \Big(\frac{1}{1+\phi^2}\Big) \exp\Big\{ -\frac{1}{2V}\sum_i(\phi - x_i)^2 \Big\},$$

$$\text{and} \quad \pi(V|\phi, x) \quad \sim \quad \text{IG}\Big(1 + \frac{N}{2}, 1 + \frac{1}{2}\sum_i(\phi - x_i)^2\Big).$$

Since the distribution $\pi(\phi|V, x)$ is not of standard form, direct Gibbs sampling is not possible. Instead, we use a random-scan Metropolis-within-Gibbs sampler, Müller (1993); Tierney (1994), and either update V from its conditional (Gibbs step), or update ϕ in a random walk-Metropolis step with a $\phi' \sim N(\phi, 1)$ proposal, each case chosen with probability $1/2$. Since both the Cauchy and the inverse Gamma distributions are heavy tailed, we naturally expect that the MCMC samples will be highly variable. Indeed, this was found to be the case in the simulation example we consider next, where the above algorithm is applied to a vector x of $N = 100$ i.i.d. $N(2, 4)$ observations, and with initial values $\phi_0 = 0$ and $V_0 = 1$. As a result of this variability, the standard empirical averages of the values of the two parameters also converge very slowly. Since V is the more variable of the two, we let $F(\phi, V) = V$ and consider the problem of estimating its posterior mean. We will compare the performance of the standard empirical averages $\mu_n(F)$ with that of the two adaptive estimators $\mu_{n,\text{rev},1}(F)$ and $\mu_{n,\text{rev},2}(F)$, with the control variate $U = G - PG$ defined in terms of the function $G = F$. Note that, in this setting, we are restricted in our choice of functions G to those which depend only on V. Since ϕ is updated by an accept/reject Metropolis step, if G depended on ϕ it would not be possible to compute the required one-step expectation PG in closed form.

Figure 8 shows a typical realization of the results of the three estimators, for $n = 10000$ simulation steps. The "true" posterior mean of V is estimated to be ≈ 4.254 after 10 million simulation steps, and the corresponding variance reduction factors, estimated from $T = 100$ repetitions of the same experiment, are shown in Table 8.

Variance reduction factors				
	Simulation steps			
Estimator	$n = 10000$	$n = 50000$	$n = 100000$	$n = 200000$
$\mu_{n,\text{rev},1}(F)$	2.58	7.62	9.34	8.13
$\mu_{n,\text{rev},2}(F)$	7.89	7.48	10.46	8.54

Table 8: Estimated factors by which the variance of $\mu_n(F)$ is larger than the corresponding variances of $\mu_{n,\text{rev},1}(F)$ and $\mu_{n,\text{rev},2}(F)$, respectively, after $n = 10000, 50000, 100000$ and 200000 simulation steps.
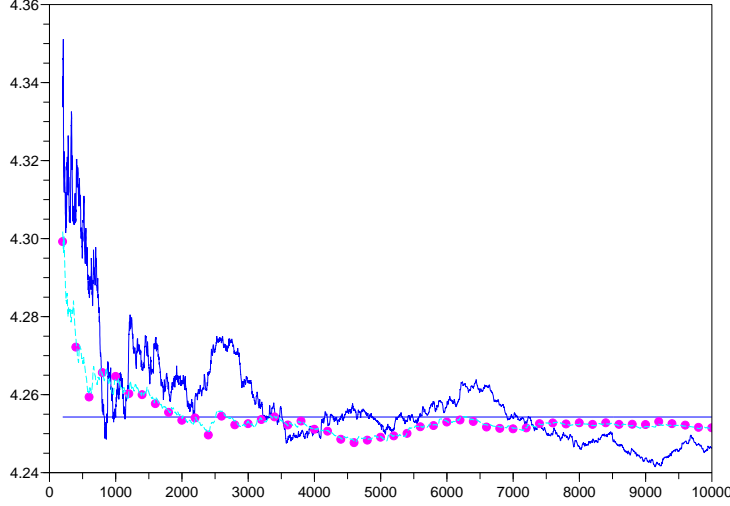
Figure 8: The sequence of the standard ergodic averages is shown as a solid blue line; the adaptive estimates $\mu_{n,\mathrm{rev},1}(F)$ as bold magenta dots; and the adaptive estimates $\mu_{n,\mathrm{rev},2}(F)$ as a cyan dashed line. For visual clarity, the values $\mu_{n,\mathrm{rev},1}(F)$ are plotted only every 200 simulation steps.

4   Variance, Bias and Choosing Between $\mu_{n,\mathbf{rev},1}$ and $\mu_{n,\mathbf{rev},2}$

We examine how the use of the adaptive estimators estimators $\mu_{n,\mathrm{rev},1}(F)$ and $\mu_{n,\mathrm{rev},2}(F)$ can affect the estimation bias, especially in cases where the initial values of the sampler are far from the true mean of the target distribution. Also we briefly discuss the different advantages and disadvantages offered by each of these two estimators, and conclude that, generally, the preferable choice is $\mu_{n,\mathrm{rev},2}(F)$.

4.1   Estimation bias

The primary focus of the present work is on variance reduction, more specifically, on reducing the asymptotic variance σ_F^2 of the estimators $\mu_n(F)$. This variance is a "steady-state" object, in that it characterizes the long-term behavior of the averages $\mu_n(F)$ and depends neither on the initial condition $X_0 = x$ nor on the transient behavior of the chain. The bias, on the other hand, depends heavily on the initial condition, and vanishes asymptotically. Indeed, according to the expression in (2) for the solution $\hat{F}$ of the Poisson equation (3),

$$\mathrm{bias}_x(\mu_n(F)) := E_x[\mu_n(F)] - \pi(F) = \frac{1}{n}\sum_{k=0}^{n-1}[P^k F(x) - \pi(F)] = \frac{1}{n}[\hat{F}(x) + o(1)], \qquad (22)$$

which decays to zero approximately like $\hat{F}(x)/n$, as $n \to \infty$.

If instead of the standard ergodic averages $\mu_n(F)$ we use an estimator of the form $\mu_n(F_\theta)$ based on a control variate $U = G - PG$ for some function G, then, replacing F with F_θ in the

above computation shows that the bias of $\mu_n(F_\theta)$ is,

$$\text{bias}_x(\mu_n(F_\theta)) := E_x[\mu_n(F_\theta)] - \pi(F) = \frac{1}{n}[\hat{F}_\theta(x) + o(1)] = \frac{1}{n}[\hat{F}(x) - \theta G(x) + o(1)], \quad (23)$$

where we used the fact that $\hat{F}_\theta = \hat{F} - \theta G$, as shown in (8). Therefore, the function G that minimizes (to first order) the bias of the estimates $\mu_n(F_\theta)$ is again the solution of the Poisson equation $G = \hat{F}$. Of course this can also be seen directly from the definition of F_θ: As in (7), if $G = \hat{F}$ and $\theta = 1$, then, $F - \theta U \equiv \pi(F)$, leading to an estimator with zero bias and zero variance.

As noted earlier, the solution $\hat{F}$ to the Poisson equation cannot be computed in the vast majority of realistic examples of nontrivial Markov chains appearing in applications, if for no other reason, because it requires knowledge of the mean $\pi(F)$. Instead, a more pragmatic goal is to try and choose a "good" value for the parameter θ, so that the resulting estimator $\mu_n(F_\theta)$ has significantly smaller bias than $\mu_n(F)$. Unlike the variance, the bias depends heavily on the initial condition x, so there is no obvious choice that makes $\theta G(x)$ "close" to $\hat{F}(x)$ for all x. In fact, for good variance reduction results we wish to have $G(x) \approx \hat{F}(x)$ for most values x near the bulk of the target distribution π, whereas for the bias we need to have $G(x) \approx \hat{F}(x)$ at the initial value x of the chain, which may well be out in the tail of π. Nevertheless, it may be natural to expect that taking $\theta \approx \theta^*$ could be a good general substitute. Although θ^* does not eliminate the bias entirely, it does bring θG "as close as possible" to $\hat{F}$, where "closeness" here is measured in the sense of minimizing $\sigma^2_{F_\theta}$.

In order to examine whether the choice $\theta \approx \theta^*$ does indeed offer an advantage in terms of the bias, we revisit Example 2 from Section 2.4, where it was observed that in some cases the adaptive estimators $\mu_{n,\text{rev},1}(F)$ and $\mu_{n,\text{rev},2}(F)$ did offer a significant reduction in the bias.

Example 2 revisited: Bias and MSE. As before, we use the random-scan Gibbs sampler to simulate from a bivariate Normal vector $(X, Y) \sim \pi(x, y)$, where the expected values of both X and Y are zero, $\text{Var}(X) = 1$, $\tau^2 = \text{Var}(Y) = 10$, and their covariance $E(XY) = \rho\tau$ with $\rho = 0.99$. To estimate the expected value of X under π, we let $F(x, y) = x$, $G(x, y) = x + y$ and take the control variate $U = G - PG$.

In Section 2.4 it was noted that, when the initial values of the sampler were relatively far from their mean (so that the samples where initially heavily biased), the adaptive estimators $\mu_{n,\text{rev},1}(F)$ and $\mu_{n,\text{rev},2}(F)$ not only reduced the variance, but also appeared to be correcting for the estimation bias; see Figure 3. This agrees with the intuition obtained by the discussion following the computations in (22) and (23) above. In order to get a more precise idea of the effect of the use of the adaptive estimators $\mu_{n,\text{rev},1}(F)$ and $\mu_{n,\text{rev},2}(F)$ on the bias and the overall estimation error, we estimate the factors by which each of these two estimators improves (i) the bias; (ii) the variance; and (iii) the overall estimation mean-squared error (MSE). The results are shown in Tables 9 and 10; Table 9 shows simulation results for a sampler started from initial values near the true mean of the distribution, $x_0 = y_0 = 0.1$, and Table 10 shows corresponding results with initial values $x_0 = 4$, $y_0 = 12$.

The bias $E_x[\mu_n(F)] - \pi(F)$ of the standard estimators $\mu_n(F)$ was computed from $T = 200$ independent repetitions of the same experiment, in a way similar to that used for the variance in the earlier examples; see the discussion in Example 1. Specifically, for $\mu_n(F)$, $T = 200$ different estimates $\mu_n^{(i)}(F)$, for $i = 1, 2, \ldots, T$, were obtained from $T = 200$ independent runs of the Gibbs

24

Example 2: $x_0 = y_0 = 0.1$					
Bias reduction factors					
Estimator	$n = 1000$	$n = 10000$	$n = 20000$	$n = 50000$	$n = 100000$
$\mu_{n,\mathrm{rev},1}(F)$	0.80	2.06	1.27	1.65	1.00
$\mu_{n,\mathrm{rev},2}(F)$	0.83	1.31	1.02	1.58	0.75
Variance reduction factors					
$\mu_{n,\mathrm{rev},1}(F)$	2.46	7.27	8.06	8.77	9.60
$\mu_{n,\mathrm{rev},2}(F)$	3.51	6.34	6.62	8.15	9.33
MSE reduction factors					
$\mu_{n,\mathrm{rev},1}(F)$	2.45	7.26	8.04	8.64	9.54
$\mu_{n,\mathrm{rev},2}(F)$	3.46	6.29	6.59	8.03	9.29

Table 9: Estimated factors by which the bias, variance, and MSE of $\mu_n(F)$ is larger than that of $\mu_{n,\mathrm{rev},1}(F)$ and $\mu_{n,\mathrm{rev},2}(F)$, respectively, after $n = 1000, 10000, 20000, 50000$ and 100000 simulation steps.

Example 2: $x_0 = 4$, $y_0 = 12$					
Bias reduction factors					
Estimator	$n = 1000$	$n = 10000$	$n = 20000$	$n = 50000$	$n = 100000$
$\mu_{n,\mathrm{rev},1}(F)$	1.95	3.66	4.45	7.97	7.97
$\mu_{n,\mathrm{rev},2}(F)$	6.88	7.39	8.35	13.52	9.22
Variance reduction factors					
$\mu_{n,\mathrm{rev},1}(F)$	1.59	8.04	9.01	8.66	9.08
$\mu_{n,\mathrm{rev},2}(F)$	0.89	7.00	7.91	8.72	8.72
MSE reduction factors					
$\mu_{n,\mathrm{rev},1}(F)$	2.57	8.34	9.75	9.11	9.34
$\mu_{n,\mathrm{rev},2}(F)$	2.57	7.60	9.01	9.22	8.98

Table 10: Estimated factors by which the bias, variance, and MSE of $\mu_n(F)$ is larger than that of $\mu_{n,\mathrm{rev},1}(F)$ and $\mu_{n,\mathrm{rev},2}(F)$, respectively, after $n = 1000, 10000, 20000, 50000$ and 100000 simulation steps.

sampler. Then the bias of $\mu_n(F)$ was estimated by,

$$\bar{\mu}_n(F) - \pi(F) \; := \; \frac{1}{T} \sum_{i=1}^{T} \mu_n^{(i)}(F) - \pi(F), \tag{24}$$

and the same procedure was applied to estimate the bias of $\mu_{n,\mathrm{rev},1}(F)$ and $\mu_{n,\mathrm{rev},2}(F)$. The bias reduction factors shown in the two tables are the ratios of the corresponding (absolute values of the) bias estimates. The variance reduction factors were computed as before, and the MSE reduction factors were computed in an analogous manner.

In both cases, the results clearly show that both estimators $\mu_{n,\mathrm{rev},1}(F)$ and $\mu_{n,\mathrm{rev},2}(F)$ greatly reduce the estimation error, not only in terms of their asymptotic variance, but in terms of the bias and of the overall estimation error as well.

4.2 Choosing between the two estimators

In the simulation examples presented so far as well as in many more experiments, we observed that the overall performance of the two estimators is fairly similar. One difference is that, in cases where the initial values of the sampler were very far from the bulk of the mass of the target distribution π, sometimes $\hat{\theta}_{n,\text{rev},1}$ converged faster than $\hat{\theta}_{n,\text{rev},2}$ and the corresponding estimator $\mu_{n,\text{rev},1}(F)$ gave better results than $\mu_{n,\text{rev},2}(F)$. The reason for this discrepancy is the existence of a the time-lag in the definition of $\hat{\theta}_{n,\text{rev},2}$: When the initial simulation phase produces samples that approach the area near the mode of the distribution approximately monotonically, the denominator of $\hat{\theta}_{n,\text{rev},2}$ accumulates a systematic one-sided error, and therefore takes longer to converge. But this is a transient phenomenon, and can be addressed (and often eliminated) by including a burn-in phase in the simulation.

One the other hand, we observed that the estimator $\hat{\theta}_{n,\text{rev},2}$ was systematically more stable than $\hat{\theta}_{n,\text{rev},1}$, especially in the more complex MCMC scenarios involving multiple control variates. This was particularly pronounced in cases where the denominator of θ^* is near zero. There, because of the inevitable fluctuations in the estimation of this denominator, the values of $\hat{\theta}_{n,\text{rev},1}$ fluctuated wildly between large negative and positive values, whereas the estimates $\hat{\theta}_{n,\text{rev},2}$ were much more reliable since, by definition, the denominator of $\hat{\theta}_{n,\text{rev},2}$ is always nonnegative.

In conclusion, we find that between $\mu_{n,\text{rev},1}$ and $\mu_{n,\text{rev},2}$, the estimator $\mu_{n,\text{rev},2}$ is generally the more reliable, preferable choice. In all the examples that follow, we will restrict attention to this estimator; see also the comments at the end of Section 5.1.

5 Using Multiple Control Variates Simultaneously

5.1 Adaptive estimators with multiple control variates

Starting from the same setting of a Markov chain $\{X_n\}$ with transition kernel P, invariant measure π, and a function $F : \mathsf{X} \to \mathbb{R}$ whose mean under π is to be estimated, suppose that, instead of using a single control variate $U = G - PG$, we wish to use multiple $U_j = G_j - PG_j$, $j = 1, 2, \ldots, k$. One reason for such a choice is so that the optimal $G = \hat{F}$ may potentially be approximated as a linear combination of "basis functions" G_j, namely, $\hat{F} \approx \sum_j \theta_j G_j$.

Formally, let $G : \mathsf{X} \to \mathbb{R}^k$ denote the column vector $G = (G_1, G_2, \ldots, G_k)^t$, where each G_j is a given function $G_j : \mathsf{X} \to \mathbb{R}$, and similarly write $U = (U_1, U_2, \ldots, U_k)^t$ for the column vector of control variates $U_j = G_j - PG_j$. For any coefficient vector $\theta = (\theta_1, \theta_2, \ldots, \theta_k)^t \in \mathbb{R}^k$, we write $F_\theta = F - \langle \theta, U \rangle$ and consider the corresponding modified estimator for $\pi(F)$,

$$\mu_n(F_\theta) = \mu_n(F) - \langle \theta, \mu_n(U) \rangle = \mu_n(F) - \sum_{j=1}^{k} \theta_j \mu_n(U_j).$$

[Here and throughout the paper all vectors are column vectors, and $\langle \cdot, \cdot \rangle$ denotes the usual Euclidean inner product.] Arguing exactly as in the one-dimensional case, the asymptotic variance of F_θ can be expressed as,

$$\sigma_{F_\theta}^2 = \sigma_F^2 - 2\pi \Big(\hat{F} \langle \theta, G \rangle - P\hat{F} \langle \theta, PG \rangle \Big) + \pi \Big(\langle \theta, G \rangle^2 - \langle \theta, PG \rangle^2 \Big), \tag{25}$$

where, PG stands for the vector $(PG_1, PG_2, \ldots, PG_k)^t$.

To find the optimal θ^*, differentiate the quadratic $\sigma_{F_\theta}^2$ with respect to each θ_i and set the derivative equal to zero, to obtain, in matrix notation,

$$\Gamma(G)\theta^* = \pi(\hat{F}G - (P\hat{F})(PG)),$$

where the $k \times k$ matrix $\Gamma(G)$ has entries, $\Gamma(G)_{ij} = \pi(G_i G_j - (PG_i)(PG_j))$. Therefore,

$$\theta^* = \Gamma(G)^{-1}\pi(\hat{F}G - (P\hat{F})(PG)), \tag{26}$$

as long as $\Gamma(G)$ is invertible. Note that this expression is perfectly analogous to the one-dimensional formula for θ^* in (9). Also, in view of equation (12) from Section 2.2, the entries of $\Gamma(G)$ can be expressed as,

$$\Gamma(G)_{ij} = \pi(G_i G_j - (PG_i)(PG_j)) = \pi(\hat{U}_i G_j - (P\hat{U}_i)(PG_j)) = \sum_{n=-\infty}^{\infty} \text{Cov}_\pi(U_i(X_0), G_j(X_n)).$$

This shows that $\Gamma(G)$ has the structure of a covariance matrix and, in particular, it suggests that $\Gamma(G)$ should be positive semidefinite. Indeed, the following lemma states that the entries of $\Gamma(G)$ can be written in a way which makes both of these assertions obvious:

Lemma 2. Let $K(G)$ denote the covariance matrix of the random variables

$$Y_i := G_i(X_1) - PG_i(X_0), \quad i = 1, 2, \ldots, k,$$

where $X_0 \sim \pi$. Then $\Gamma(G) = K(G)$, that is, for all $1 \leq i, j \leq k$,

$$\pi(G_i G_j - (PG_i)(PG_j)) = E_\pi\left[\Big(G_i(X_1) - PG_i(X_0)\Big)\Big(G_j(X_1) - PG_j(X_0)\Big)\right]. \tag{27}$$

PROOF. Expanding the right-hand side of (27) we obtain,

$$\pi(G_i G_j) - E_\pi[G_i(X_1)PG_j(X_0)] - E_\pi[G_j(X_1)PG_i(X_0)] + \pi((PG_i)(PG_j)),$$

and the result follows upon noting that the second and third terms above are both equal to the fourth. To see this, observe that the second term can be rewritten as,

$$E_\pi\Big\{E\Big[G_i(X_1)PG_j(X_0)\,\Big|\,X_0\Big]\Big\} = E_\pi\Big[E[G_i(X_1)\,|\,X_0]PG_j(X_0)\Big] = \pi((PG_i)(PG_j)),$$

and similarly for the third term. $\qquad\square$

Therefore, the optimal coefficient θ^* can also be expressed as,

$$\theta^* = K(G)^{-1}\pi(\hat{F}G - (P\hat{F})(PG)). \tag{28}$$

Proceeding exactly as before, for a reversible chain, starting from the expressions for θ^* in (26) and (28) we obtain:

Proposition 2. If the chain $\{X_n\}$ is reversible, then the optimal coefficient vector θ^* for the control variates $U_i = G_i - PG_i$, $i = 1, 2, \ldots, k$ can be expressed as,

$$\theta^* = \theta_{\text{rev}}^* := \Gamma(G)^{-1}\pi\big((F - \pi(F))(G + PG)\big), \tag{29}$$

or, alternatively,

$$\theta^* = \theta^*_{\text{rev}} := \mathrm{K}(G)^{-1}\pi\big((F - \pi(F))(G + PG)\big). \tag{30}$$

The proof of Proposition 2 is perfectly analogous to that of Proposition 1 in the case of a single control variate.

As before, the expressions (29) and (30) suggest estimating θ^* via,

$$\hat{\theta}_{n,\Gamma} = \Gamma_n(G)^{-1}[\mu_n(F(G + PG)) - \mu_n(F)\mu_n(G + PG)]$$
$$\text{or} \quad \hat{\theta}_{n,\mathrm{K}} = \mathrm{K}_n(G)^{-1}[\mu_n(F(G + PG)) - \mu_n(F)\mu_n(G + PG)],$$

where the $k \times k$ matrices $\Gamma_n(G)$ and $\mathrm{K}_n(G)$ are defined, respectively, by,

$$(\Gamma_n(G))_{ij} = \mu_n(G_i G_j) - \mu_n((PG_i)(PG_j))$$
$$\text{and} \quad (\mathrm{K}_n(G))_{ij} = \frac{1}{n}\sum_{t=0}^{n-1}(G_i(X_t) - PG_i(X_{t-1}))(G_j(X_t) - PG_j(X_{t-1})).$$

The resulting estimators, $\mu_n(F_{\hat{\theta}_{n,\Gamma}})$ and $\mu_n(F_{\hat{\theta}_{n,\mathrm{K}}})$ for $\pi(F)$ based on the vector of control variates $U = G - PG$ and the coefficients $\hat{\theta}_{n,\Gamma}$ and $\hat{\theta}_{n,\mathrm{K}}$, respectively, are defined analogously to the single-control-variate case as,

$$\mu_{n,\Gamma}(F) := \mu_n(F_{\hat{\theta}_{n,\Gamma}}) = \mu_n(F) - \langle\hat{\theta}_{n,\mathrm{K}}, \mu_n(U)\rangle \tag{31}$$
$$\text{and} \quad \mu_{n,\mathrm{K}}(F) := \mu_n(F_{\hat{\theta}_{n,\Gamma}}) = \mu_n(F) - \langle\hat{\theta}_{n,\mathrm{K}}, \mu_n(U)\rangle. \tag{32}$$

Recall that in Section 4.2 we concluded that, for the case of a single control variate, the adaptive estimator $\mu_{n,\text{rev},2}$ was generally preferable to $\mu_{n,\text{rev},1}$. For the same reasons, and also based on the results of extensive simulation experiments with multiple control variates, we similarly conclude that $\mu_{n,\mathrm{K}}(F)$ is more reliable, more stable, and generally preferable to $\mu_{n,\Gamma}(F)$. Therefore, in all of our subsequent examples we restrict attention to the estimator $\mu_{n,\mathrm{K}}(F)$.

5.2 Examples

Here we re-examine two of the earlier examples, and illustrate how the use of multiple control variates can often provide a much greater improvement in estimation accuracy.

Example 2 revisited. Let $(X, Y) \sim \pi(x, y)$ be a zero mean, bivariate Normal distribution, with $\mathrm{Var}(X) = 1$, $\mathrm{Var}(Y) = \tau^2$, and $E(XY) = \rho\tau$ for some $\rho \in (-1, 1)$. To estimate the expected value of X under π we sample from π using a random-scan Gibbs sampler and set $F(x, y) = x$. Instead of the single control variate $U = G - PG$ based on $G(x, y) = x + y$, here we consider two control variates U_1, U_2 defined in terms of $G_1(x, y) = x$ and $G_2(x, y) = y$, respectively. We examine the performance of the adaptive estimator $\mu_{n,\mathrm{K}}(F)$, and compare it with the performance of obtained earlier by the single-control-variate estimator $\mu_{n,\text{rev},2}(F)$.

Figure 9 depicts a typical realization of the sequence of estimates of the standard ergodic averages $\mu_n(F)$, as well as the corresponding estimates obtained by $\mu_{n,\mathrm{K}}(F)$, for $n = 20000$ simulation steps. The parameter values are $\rho = 0.99$ and $\tau^2 = 10$, with initial values $x_0 = y_0 = 0.1$. Table 11 shows the corresponding variance reduction factors, estimated from $T = 200$ repetitions of the same experiment.
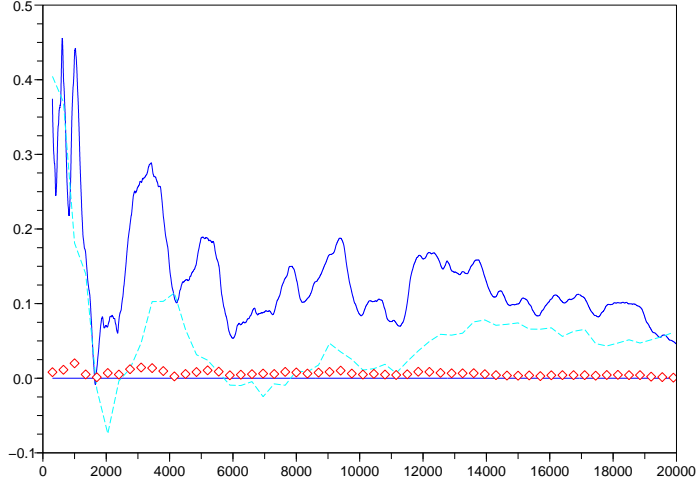
28

Figure 9: The sequence of the standard ergodic averages is shown as a solid blue line; the adaptive estimates $\mu_{n,\mathrm{rev},2}(F)$ with respect to the single control variate U as a cyan dashed line, and the adaptive estimates $\mu_{n,\mathrm{K}}(F)$ with respect to the two control variates U_1, U_2 as red diamonds. For visual clarity, the values $\mu_{n,\mathrm{K}}(F)$ are plotted only every 350 simulation steps.

Variance reduction factors					
	Simulation steps				
Estimator	$n = 1000$	$n = 10000$	$n = 50000$	$n = 100000$	$n = 200000$
$\mu_{n,\mathrm{rev},2}(F)$	2.89	6.17	8.17	7.42	9.96
$\mu_{n,\mathrm{K}}(F)$	4.13	27.91	122.4	262.5	445.0

Table 11: Estimated factors by which the variance of $\mu_n(F)$ is larger than the corresponding variances of $\mu_{n,\mathrm{rev},2}(F)$ and $\mu_{n,\mathrm{K}}(F)$, respectively, after $n = 1000, 10000, 50000, 100000$ and 200000 simulation steps.

Clearly the estimator based on the two control variates is extremely effective, and certainly significantly better than the one based on a single control variate. As in Example 4, this effectiveness is actually explained by the fact that the exact solution of the Poisson equation in this case is of the form $\hat{F}(x,y) = ax + by$. Indeed, it is a simple matter to verify that, $\hat{F}(x,y) = \frac{2}{1-\rho^2}[x + \frac{\rho}{\tau}y]$.

Example 6 revisited. Recall the setting of the inference problem in Example 6 above, where, based on $N = 500$ independent observations $x = (x_1, x_2, \ldots, x_N)$ generated from the mixture $pN(\mu_1, \sigma^2) + (1-p)N(\mu_2, \sigma^2)$, we wish to estimate μ_1. The mixing proportion $p = 0.7$ and the variance $\sigma^2 = 1$ are fixed and known, $N(0, 10\sigma^2)$ priors are placed on μ_1, μ_2, and each of the binary latent variables (Z_i) equals 1 if the corresponding x_i is generated from the first component of the mixture, and $Z_i = 0$ otherwise. We use the random-scan Gibbs sampler, based on the full conditionals of the posterior, given in (19) and (20).

In Section 3, letting $F(\mu_1, \mu_2, Z) = \mu_1$ and using a control variate $U = G - PG$ in terms of the function $G(\mu_1, \mu_2, Z) = (Np + 1/10)\mu_1 + \sum_i Z_i x_i$, we obtained variance reduction factors around

29

20. The natural next step is to repeat the same experiment, this time with two control variates U_1, U_2 defined in terms of the functions $G_1(\mu_1, \mu_2, Z) = \mu_1$ and $G_2(\mu_1, \mu_2, Z) = \sum_i Z_i x_i$. In numerous simulation experiments we observed that, using two control variates in this case offered no apparent performance improvement. This suggests that the ratio of the coefficients of the functions G_2 and G_1, which was earlier chosen as $1/(Np+0.1)$ based on a heuristic computation, must be near-optimal. Indeed, after two million Gibbs steps, the estimated value of the optimal parameter vector θ^* for the two control variates U_1, U_2 was $\approx (3.832, 0.0129)$. The resulting optimal ratio $0.0123/3.832 \approx 0.0034$ is, as expected, quite close to $1/(Np + 0.1) \approx 0.0029$.

Next we consider using four control variates, defined in terms of the functions G_1, G_2 above together with $G_3(\mu_1, \mu_2, Z) = \mu_2$ and $G_4(\mu_1, \mu_2, Z) = \mu_1^2$. In this case, the corresponding variance reduction factors, estimated from $T = 100$ repetitions of the same experiment (with initial values for the sampler $\mu_1 = 0$, $\mu_2 = 1$), are **97.47, 138.39, 91.84** and **103**, after $n = 10000, 50000, 100000$ and 200000 simulation steps, respectively. Compared to the earlier results (variance reduction factors around 20), these results clearly demonstrate the significant improvement in estimation accuracy due to the simultaneous use of multiple control variates.

6 Four More Complex MCMC Examples

This section illustrates our proposed methodology applied to a series of real Bayesian inference problems, providing guidelines on how functions G can be chosen for the construction of effective control variates. The first example is a binary probit model, an early success of MCMC inference through data augmentation; see Albert and Chib (1993). The second example is a simple finite mixture of normals, another early application of data augmentation via Gibbs sampling; see Diebolt and Robert (1994). What makes this problem particularly interesting is the fact that, although we impose an *a priori* restriction on the ordering of means, the control variates methodology can still be applied after a first phase of *unrestricted* MCMC sampling, and after the sample has been ordered at the post-processing stage. If the objective is to estimate the means, the calculation of effective control variates U is still possible, despite the fact that the resulting Markov chain has a particularly complex structure. The third example is of a Bayesian model-determination problem, in which model searching is achieved by a discrete Metropolis algorithm on the space of candidate models. Such applications have recently found tremendous interest, especially in the context of genetics (see, e.g., Bottolo and Richardson (2008)) where the model space is endowed with a multimodal discrete density. Finally, in the case of a simple log-linear model we show that, even when we are forced to consider functions G that are very different from F, the resulting control variates U can still be very effective in terms of variance reduction.

6.1 A binary probit example

Probit models are a well-known and commonly used class of discrete regression models; see, for example, the monograph by Johnson and Albert (1999) and the references therein. Here we illustrate the use of the control variate methodology when a random-scan Gibbs sampler is used for Bayesian inference from the posterior of a binary probit model.

Specifically, we begin with an $N \times k$ matrix $x = (x_1^t, x_2^t, \ldots, x_N^t)^t$ of known covariates, where each x_i is a column vector in $\mathbb{R}^k$. We also have and a vector $Y = (Y_1, Y_2, \ldots, Y_N)$ of known binary responses Y_i, where we assume that the Y_i have,

$$p_i := \Pr\{Y_i = 1\} = 1 - \Pr\{Y_i = 0\} = \Phi(x_i^t \beta), \quad i = 1, 2, \ldots, N,$$

and the unknown parameter vector $\beta \in \mathbb{R}^k$ is to be estimated. To facilitate sampling from the posterior of β, Albert and Chib (1993) introduce independent latent random variables $Z = (Z_1, Z_2, \ldots, Z_N)^t$, where each $Z_i \sim N(x_i^t \beta, 1)$. In other words $Z = x\beta + \epsilon$, where $\epsilon \sim N(0, I)$ is independent noise. Then the Y_i can be expressed, $Y_i = \mathbb{I}_{\{Z_i > 0\}}$, so that, again, $p_i = \Phi(x_i^t \beta)$.

If we place a diffuse prior on β, then $\pi(\beta|x, Y, Z) \sim N((x^t x)^{-1}(x^t Z), (x^t x)^{-1})$, and the Z_i are conditionally independent given x, Y, β, with,

$$\pi(Z_i|x, Y, \beta) \quad \sim \quad N(x_i^t \beta, 1) \quad \text{conditional on } Z_i > 0, \text{ if } Y_i = 1;$$
$$\pi(Z_i|x, Y, \beta) \quad \sim \quad N(x_i^t \beta, 1) \quad \text{conditional on } Z_i < 0, \text{ if } Y_i = 0.$$

We consider a specific example using the "statistics class" data from Johnson and Albert (1999, p. 77). In this case, for $N = 30$ students, each Y_i is the indicator of weather student i passed or failed in a statistics class. There are $k = 2$ covariates (x_{i1}, x_{i2}) for each student, where $x_{i1} = 1$ for all i and x_{i2} is the ith student's SAT Math test score. We place a diffuse prior on the coefficient vector $\beta = (\beta_0, \beta_1)$, and we consider the problem of estimating the posterior mean of β_1. (The parameter β_1 is chosen as the more interesting of the two; the results are very similar for the case of β_0.) To that end, we let $F(\beta, Z) = \beta_1$, and we also consider a vector $U = G - PG$ of control variates based on five-component function G,

$$G(\beta, Z) = (\beta, (x^t x)^{-1} x^t Z, \beta_1^2)^t.$$

For the initial condition of β in the sampler we took its least-squares estimate $\hat{\beta} := (x^t x)^{-1}(x^t y)$, and for Z we simply drew a sample from its full conditional density as above. The choice of the function G for the construction of control variates is pretty self-evident: The variables β_0, β_1 and β_1^2 should obviously be strongly correlated with the target variable β_1, and the vector $(x^t x)^{-1} x^t Z$ is included in an attempt to minimize the effect of the mean of β under its full conditional.

The result of a typical realization of the random-scan Gibbs sampler after 15000 iterations is shown in Figure 10. The horizontal line shows the "true" value of $\pi(F) \approx 0.03759$, the result of $\mu_n(F)$ after 10 million Gibbs iterations. The variance reduction factors obtained by $\mu_{n,\mathrm{K}}(F)$, estimated after $T = 100$ repetitions of this experiment, are **5.08, 34.22, 53.54, 88.37** and **69.72**, after $n = 1000, 10000, 50000, 100000$ and 200000 iterations, respectively.

6.2   Gaussian mixtures

Mixtures of densities provide a versatile class of statistical models that have received a lot of attention from both a theoretical and a practical perspective, for many decades now. Mixtures primarily serve as a means of modelling heterogeneity for classification and discrimination, and as a way of formulating flexible models for density estimation. Although one of the fist major success stories in the MCMC community was the Bayesian implementation of the finite Gaussian mixtures problem, Tanner and Wong (1987); Diebolt and Robert (1994), there are still numerous unresolved issues in inference for finite mixtures, as discussed, for example, in the recent review paper by Marin et al. (2005). These difficulties emanate primarily from the fact that such models are often ill-posed or non-identifiable. In terms of Bayesian inference via MCMC, these issues reflect important problems in prior specifications and label switching. In particular, improper priors are hard to use and proper mixing over all posterior modes requires enforcing label-switching moves through Metropolis steps. Detailed discussions of the dangers emerging from prior specifications and identifiability constraints can be found in Marin et al. (2005); Lee et al. (2008); Jasra et al. (2005).
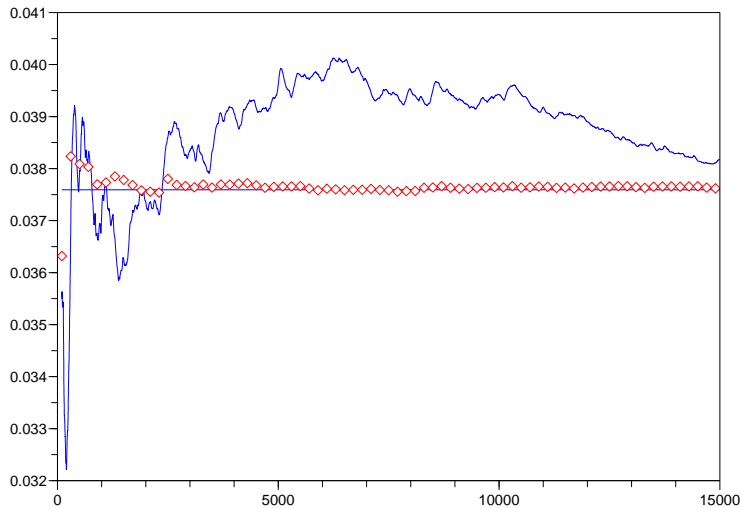
Figure 10: Probit: The sequence of the standard ergodic averages is shown as a solid blue line and the adaptive estimates $\mu_{n,\mathrm{K}}(F)$ as red diamonds. The adaptive estimates are plotted only every 200 simulation steps.

Here we generalize the estimation setting of Example 6 above, by employing a more realistic two-component Gaussian mixtures model as follows. Starting with $N = 500$ data points $x = (x_1, x_2, \ldots, x_N)$ generated from the mixture distribution $0.7N(0, 0.5^2) + 0.3N(0.1, 3^2)$, and assuming that the means, variances and mixing proportions are all unknown, we consider the problem of estimating the two means. The usual Bayesian formulation enriches that of Example 6 by introducing parameters $(\mu_1, \mu_2, \sigma_1, \sigma_2, Z, p)$, as follows. The data are assumed to be i.i.d. from $pN(\mu_1, \sigma_1^2) + (1 - p)N(\mu_2, \sigma_2^2)$, and we place the following priors: $p \sim \mathrm{Dirichlet}(\delta, \delta)$, the two means μ_1, μ_2 are independent with each $\mu_j \sim N(\xi, \kappa^{-1})$, and similarly the variances are independent with each $\sigma_j^{-2} \sim \mathrm{Gamma}(\alpha, \beta)$. We adopt the vague, data-dependent prior structure of Richardson and Green (1997): We set $\delta = 1$, we let ξ equal to the empirical mean of the data x, $\kappa^{-1/2}$ is taken to be equal to the data range, $\alpha = 2$ and $\beta = 0.02\kappa^{-1}$. As before, conditional on the parameters $(\mu_1, \mu_2, \sigma_1, \sigma_2, p)$, the latent variables $Z = (Z_1, Z_2, \ldots, Z_N)$ are i.i.d. with $\Pr\{Z_i = 1\} = 1 - \Pr\{Z_i = 2\} = p$, and, given the entire parameter vector $(\mu_1, \mu_2, \sigma_1, \sigma_2, Z, p)$, the data $x = (x_i)$ are i.i.d. with each x_i having distribution $N(\mu_j, \sigma_j^2)$ if $Z_i = j$, for $i = 1, 2, \ldots, N$, $j = 1, 2$.

In order to estimate the mean vector (μ_1, μ_2) we sample from the posterior via a standard random-scan Gibbs sampler, and we also introduce the *a priori* restriction that $\mu_1 < \mu_2$. In terms of the sampling itself, as noted by Stephens (1997), it is preferable to first obtain draws from the unconstrained posterior distribution and then to impose the identifiability constraint at the post-processing stage. In each iteration, the random-scan Gibbs sampler selects one of the four parameter blocks (μ_1, μ_2), (σ_1, σ_2), Z or p, each with probability $1/4$, and draws a sample from the corresponding full conditional density. These densities are all of standard form and easy to sample from; see, for example, Richardson and Green (1997). In particular, the two

means are conditionally independent with each,

$$\mu_j \sim N\left(\frac{\sigma_j^{-2}\sum_i(y_i + \kappa\xi)\mathbb{I}_{\{Z_i=j\}}}{\sigma_j^{-2}n_j + \kappa}, \frac{1}{\sigma_j^{-2}n_j + \kappa}\right),\tag{33}$$

where $n_j = \#\{i : Z_i = j\}$, for $j = 1, 2$. Note that the data x have been generated so that the two means are very close, which results in frequent label switching throughout the MCMC run and in near-identical marginal densities of μ_1 and μ_2.

We perform a post-processing relabelling of the sampled values according to the above restriction, and we denote the ordered sampled vector by $(\mu_1^o, \mu_2^o, \sigma_1^o, \sigma_2^o, Z^o, p^o)$. In order to estimate the posterior mean of the smaller of the two means, we let,

$$F(\mu_1, \mu_2, \sigma_1, \sigma_2, Z, p) := \mu_1^o = \min\{\mu_1, \mu_2\}.$$

To reduce the variance of the estimator $\mu_n(F)$ we consider a bivariate control variate $U = G - PG$, where the function $G = (G_1, G_2)^t$ is selected as follows. For G_1 we take the obvious choice, $G_1(\mu_1, \mu_2, \sigma_1, \sigma_2, Z, p) = \mu_1^o$, so that, PG_1, the expected value of $\min\{\mu_1, \mu_2\}$ under (33), is easily seen to be,

$$\begin{aligned}PG_1(\mu_1, \mu_2, \sigma_1, \sigma_2, Z, p) &= \frac{3}{4}G_1(\mu_1, \mu_2, \sigma_1, \sigma_2, Z, p)\\ &+ \frac{\nu_1}{4}\Phi\left(\frac{\nu_2 - \nu_1}{\sqrt{\tau_1^2 + \tau_2^2}}\right) + \frac{\nu_2}{4}\Phi\left(\frac{\nu_1 - \nu_2}{\sqrt{\tau_1^2 + \tau_2^2}}\right) - \frac{1}{4}\sqrt{\tau_1^2 + \tau_2^2}\,\phi\left(\frac{\nu_2 - \nu_1}{\sqrt{\tau_1^2 + \tau_2^2}}\right),\end{aligned}\tag{34}$$

where ν_j and τ_j^2 are the means and variances of μ_j, respectively, for $j = 1, 2$, under the full conditional densities in (33); see, for example, Cain (1994). Clearly this introduces a significant amount of unwanted variability in $U_1 = G_1 - PG_1$, so, in order to cancel it out, we choose G_2 to approximately cancel out the last three terms of the above expression. Since the nonlinear terms involving ϕ and Φ are hard to handle analytically and are also bounded, and since we expect the dependence on the mean vector to be taken care of by G_1, we focus on approximating the $\sqrt{\tau_1^2 + \tau_2^2}$ factor. Since κ will be typically small compared to n_1 and n_2, we approximate τ_1^2 by $\sigma_1^2/(Np)$ and τ_2^2 by $\sigma_2^2/(N(1-p))$. And since we expect the influence of σ_1^o to be dominant over that of σ_2^o with respect to μ_1^o, a straightforward first-order Taylor expansion shows that the dominant linear term is σ_1^o, suggesting the choice $G_2(\mu_1, \mu_2, \sigma_1, \sigma_2, Z, p) = \sigma_1^o$.

To compute PG_2, we first calculate the probability $p(\text{order})$ that $\mu_1 < \mu_2$ under (33),

$$p(\text{order}) = \frac{\Phi\big(E(\mu_2|\cdots) - E(\mu_1|\cdots)\big)}{\sqrt{E(\sigma_1^2|\cdots) + E(\sigma_2^2|\cdots)}},$$

where all four expectations above are taken under the corresponding full conditional densities, and, since the full conditional of each σ_j^{-2} is a Gamma density, the expectations of σ_1, σ_2, σ_1^2, and σ_2^2, are all available in closed form. Therefore, $p(\text{order})$ can be computed explicitly, and,

$$\begin{aligned}PG_2(\mu_1, \mu_2, \sigma_1, \sigma_2, Z, p) &= \frac{1}{2}G_2(\mu_1, \mu_2, \sigma_1, \sigma_2, Z, p)\\ &+ \frac{1}{4}\big[\mathbb{I}_{\{\mu_1<\mu_2\}}E(\sigma_1|\cdots) + \mathbb{I}_{\{\mu_1>\mu_2\}}E(\sigma_2|\cdots)\big] + \frac{1}{4}\big[p(\text{order})\sigma_1 + (1 - p(\text{order}))\sigma_2\big],\end{aligned}$$

where, again, the expectations are taken under the corresponding full conditional densities.

With this choice for $G = (G_1, G_2)^t$, the variance reduction factors obtained by $\mu_{n,\mathrm{K}}(F)$ (estimated from $T = 100$ repetitions) are **16.17, 25.36, 38.99, 44.5** and **36.16**, after $n = 1000, 10000, 50000, 100000$ and 200000 simulation steps, respectively. Figure 11 shows the results of a typical simulation run. The initial values of the sampler were taken after a 1000-iteration burn-in period, and the horizontal line in the graph depicting the "true" value of the posterior mean of F was obtained after 5 million Gibbs iterations.
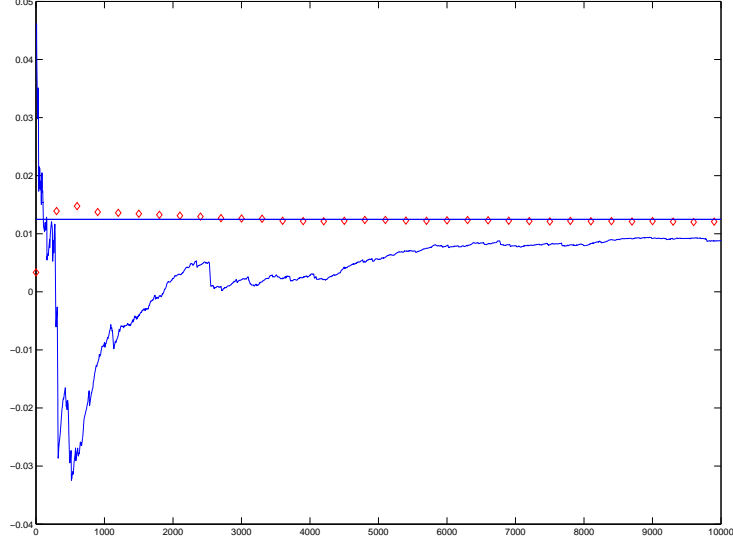

Figure 11: Two-component Gaussian mixture model: The sequence of the standard ergodic averages for μ_1 is shown as a solid blue line and the adaptive estimates $\mu_{n,\mathrm{K}}(F)$, reported every 300 iterations, as red diamonds.

6.3 A two-threshold autoregressive model

We revisit the monthly U.S. 3-month treasury bill rates, from January 1962 until December 1999, previously analyzed by Dellaportas et al. (2007) using flexible volatility threshold models. The time series has $N = 456$ points and is denoted by $r = (r_t) = (r_t \; ; \; t = 1, 2, \ldots, N)$. Here we model these data in terms of a self-exciting threshold autoregressive model, with two regimes; it is one of the models proposed by Pfann et al. (1996), and it is defined as,

$$\Delta r_t = \left\{ \begin{array}{ll} \alpha_{10} + \alpha_{11} r_{t-1} & r_{t-1} < c_1 \\ \alpha_{20} + \alpha_{21} r_{t-1} & r_{t-1} \geq c_1 \end{array} \right\} + \left\{ \begin{array}{ll} \sigma_1 \epsilon_t & r_{t-1} < c_2 \\ \sigma_2 \epsilon_t & r_{t-1} \geq c_2 \end{array} \right\}, \tag{35}$$

where $\Delta r_t = r_t - r_{t-1}$, and the parameters c_1, c_2 are the thresholds where mean or volatility regime shifts occur. Instead, we re-write the model as,

$$\Delta r_t = \left\{ \begin{array}{ll} \alpha_{10} + \alpha_{11} r_{t-1} & r_{t-1} < c_1 \\ \alpha_{20} + \alpha_{21} r_{t-1} & r_{t-1} \geq c_1 \end{array} \right\} + \left\{ \begin{array}{ll} \sigma \epsilon_t & r_{t-1} < c_2 \\ \sigma (1 + \gamma)^{1/2} \epsilon_t & r_{t-1} \geq c_2 \end{array} \right\}, \tag{36}$$

where $\gamma \geq -1$ characterizes the jump in σ^2 between the two volatility regimes. Whereas Pfann et al. (1996) use a Gibbs sampler to estimate the parameters of the model in (35), we exploit

the parameterization (36) as follows. We adopt independent improper conjugate priors for the variance, $\pi(\sigma^2) \propto \sigma^{-2}$, and for the regression coefficients, $\pi(\alpha_{ij}) \propto 1$. We take the prior for each of c_1 and c_2 to be a discrete uniform over the distinct values of $\{r_t\}$, except the two smallest and largest values of $\{r_t\}$ so that identifiability is obtained; and the prior for γ to be an exponential density with mean one, shifted to -1.

Our goal is to estimate the posterior probability $\pi(c_1^1, c_2^1 | r)$ of the most likely model, that is, of the model corresponding to the pair of thresholds (c_1^1, c_2^1) maximizing $\pi(c_1, c_2 | r)$. In the above formulation, (36) can be written equivalently as, $R = X\alpha + \epsilon$, where $R = (\Delta r_2, \ldots, \Delta r_T)^t$, $\alpha = (\alpha_{10}, \alpha_{11}, \alpha_{20}, \alpha_{21})$, ϵ is a zero-mean Gaussian vector with covariance matrix Σ, and X is the design matrix with row t given by $(1\ r_{t-1}\ 0\ 0)$, if $r_{t-1} \le c_1$, and by $(0\ 0\ 1\ r_{t-1})$ otherwise. The covariance matrix of the errors, Σ, is diagonal with $\Sigma_{tt} = \sigma^2$ if $r_{t-1} \le c_2$, and $\Sigma_{tt} = (1 + \gamma)\sigma^2$, otherwise. Integrating out the parameters α and σ, the marginal likelihood of the data r with known c_1, c_2 and γ is,

$$p(r|\gamma, c_1, c_2) \propto \exp\left\{ -\frac{1}{2}\Big[|\Sigma| + \log|X^t\Sigma^{-1}X| + N\log(R^t\Sigma^{-1}R - \hat{\alpha}^t X^t\Sigma^{-1}X\hat{\alpha})\Big]\right\},$$

where $\hat{\alpha} = (X^tX)^{-1}X^tR$ is the least-squares estimate of α; see, for example, O'Hagan and Forster (2004). After further performing a one-dimensional numerical integration over γ by numerical quadrature, we can write the marginal posterior distribution of (c_1, c_2) explicitly as $\pi(c_1, c_2 | r) \propto p(r | c_1, c_2)$. Therefore, we can sample from the posterior of the thresholds (c_1, c_2) by employing a discrete Metropolis-Hastings algorithm on (c_1, c_2), where the thresholds c_1, c_2 take values on the lattice of all the observed values of the rates (r_t) except the two farthermost at each end. This way, we replace the 8-dimensional Gibbs sampler of Pfann et al. (1996) for (35), by a five-dimensional analytical integration over α and σ, a numerical integration over γ, and a Metropolis-Hastings algorithm over (c_1, c_2).

Note that this algorithm is computationally less expensive, and also more reliable since Gibbs sampling across a discrete and continuous product space may encounter 'sticky patches' in the parameter space. The discrete Metropolis-Hastings sampler we employ is based on symmetric random walk steps, with vertical or horizontal increments of size up to ten, over the lattice of all possible values. In other words, the proposed pair (c_1', c_2') given the current values (c_1, c_2) is one of the forty "neighboring" pairs (c_1', c_2') of (c_1, c_2), where two pairs are neighbors if they differ in exactly one co-ordinate, and by a distance of at most ten locations. Clearly, here we do not touch upon the finer issues of efficient model searching, as these would possibly require more sophisticated MCMC algorithms.

After a preliminary, exploratory simulation stage, we identified the three *a posteriori* most likely pairs of thresholds as being $(c_1^1, c_2^1) = (13.63, 2.72)$, $(c_1^2, c_2^2) = (13.89, 2.72)$, $(c_1^3, c_2^3) = (13.78, 2.72)$. To estimate the actual posterior probability of the most likely model, $\pi(c_1^1, c_2^1 | r)$, we define $G_j(c_1, c_2) = \mathbb{I}_{\{(c_1, c_2)=(c_1^j, c_2^j)\}}$, for $j = 1, 2, 3$, we set $F = G_1$, and we use the control variate $U = G - PG$ based on $G = (G_1, G_2, G_3)^t$. Writing $(c_1, c_2) \sim (c_1', c_2')$ when (c_1, c_2) and (c_1', c_2') are neighboring pairs, PG_j can be expressed, for $j = 1, 2, 3$, as,

$$PG_j(c_1, c_2) = \begin{cases} 1 - \frac{1}{40}\sum_{(c_1', c_2') \sim (c_1, c_2)} \min\left\{1, \frac{p(r|c_1', c_2')}{p(r|c_1, c_2)}\right\}, & \text{if } (c_1, c_2) = (c_1^j, c_2^j); \\ \frac{1}{40}\min\left\{1, \frac{p(r|c_1^j, c_2^j)}{p(r|c_1, c_2)}\right\}, & \text{if } (c_1, c_2) \sim (c_1^j, c_2^j); \\ 0, & \text{otherwise.} \end{cases}$$

The resulting variance reduction factors obtained by $\mu_{n,K}(F)$, estimated from $T = 100$

35

repetitions, are **125.16, 32.83, 36.76, 30.90** and **30.11**, after $n = 10000, 20000, 50000, 100000$ and 200000 simulation steps, respectively. Figure 12 shows a typical simulation run. All MCMC chains were initiated at (c_1^1, c_2^1).
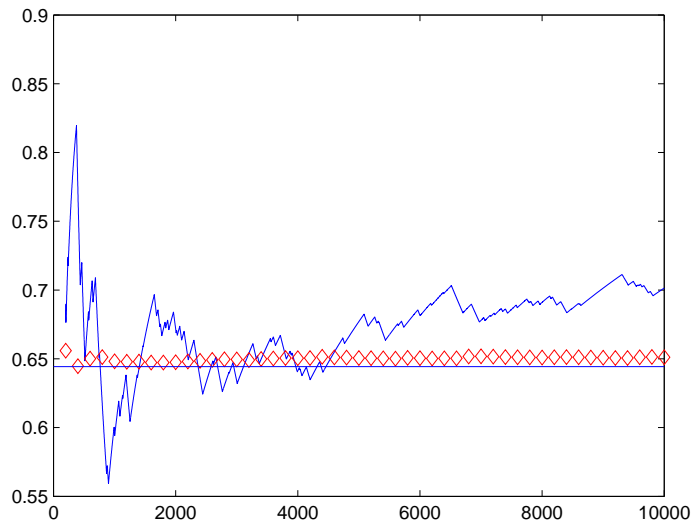

Figure 12: Threshold autoregressive model: The sequence of the standard ergodic averages $\mu_n(F)$ is shown as a solid blue line and the adaptive estimates $\mu_{n,\mathrm{K}}(F)$, reported every 500 iterations, as red diamonds. The straight horizontal line represents the posterior model probability obtained after 50 million iterations.

6.4 A log-linear model

We consider the $2 \times 3 \times 4$ table presented by Knuiman and Speed (1988), where 491 subjects were classified according to hypertension (yes, no), obesity (low, average, high) and alcohol consumption (0, 1-2, 3-5, or 6+ drinks per day). We choose to estimate the parameters of the log-linear model with three main effects and no interactions, specified as,

$$y_i \sim \text{Poisson}(\mu_i), \quad \log(\mu_i) = x_i^t \beta, \quad i = 1, 2 \ldots, 24,$$

where the y_i denote the cell frequencies, modeled as Poisson variables with corresponding means μ_i, each x_i is the ith row of the 24×7 design matrix x, based on sum-to-zero constraints, and $\beta = (\beta_1, \beta_2, \ldots, \beta_7)^t$ is the parameter vector. In Dellaportas and Forster (1999) this model was identified as having the highest posterior probability among all log-linear interaction models, under various prior specifications.

Assuming a flat prior on β, standard Bayesian inference via MCMC can be performed either by a Gibbs sampler that exploits the log-concavity of full conditional densities as in Dellaportas and Smith (1993), or by a multivariate, random walk Metropolis-Hastings sampler, in which an initial maximum likelihood estimate of the covariance matrix gives guidance as to the form of the proposal density. Instead, here we use a simple random-scan Gibbs sampler, noting that a sample from the full conditional density of each β_j can be obtained directly as the logarithm of a Gamma random variable with density,

$$\text{Gamma}\left(\sum_i y_i x_{ij}, \sum_{i:x_{ij}=1} \exp\left\{\sum_{\ell \neq j} \beta_\ell x_{i\ell}\right\}\right). \tag{37}$$

36

In order to estimate the posterior mean of all seven components of β, we set $F_j(\beta) = \beta_j$ for all j, and we use the same seven control variates $U_1, U_2, \ldots, U_7$ for each F_j, where the $U_\ell = G_\ell - PG_\ell$ are defined in terms of the functions, $G_\ell(\beta) = \exp(\beta_\ell)$, $\ell = 1, 2, \ldots, 7$. The computation of PG_ℓ is straightforward since, in view of (37), the mean of $\exp(\beta_j)$ under the full conditional density of β_j is,

$$\frac{\sum_i y_i x_{ij}}{\sum_{i:x_{ij}=1} \exp\left(\sum_{\ell \neq j} \beta_\ell x_{i\ell}\right)}.$$

The variance reduction factors obtained by $\mu_{n,\mathrm{K}}(F_j)$ after $n = 100000$ simulation steps range between 57.16 and 170.34, for different parameters β_j. More precisely, averaging over $T = 100$ repetitions, the variance reduction factors obtained by $\mu_{n,\mathrm{K}}(F_j)$ are in the range, **3.55–5.57, 38.2–57.69, 66.20–135.51, 57.16–170.34** and **85.41–179.11**, after $n = 1000, 10000, 50000$, 100000 and 200000 simulation steps, respectively. Figure 13 shows an example of a sequence of ergodic averages for β_7. All MCMC chains were initiated from the corresponding maximum likelihood estimates.
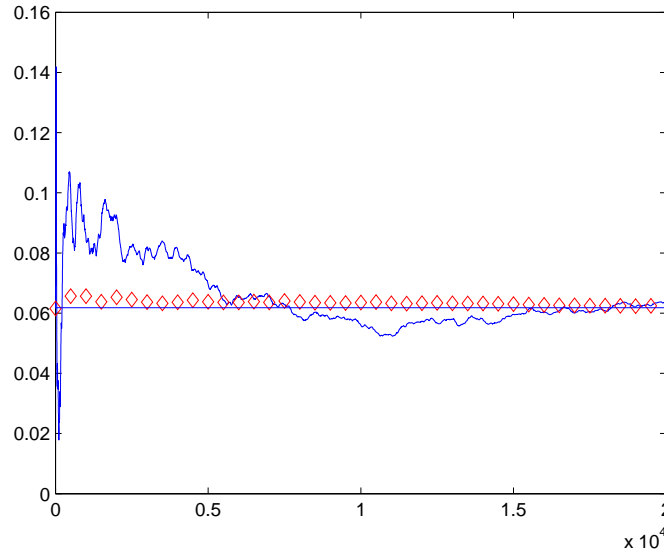

Figure 13: Log-linear model: The sequence of the standard ergodic averages $\mu_n(F_7)$ for β_7 is shown as a solid blue line and the adaptive estimates $\mu_{n,\mathrm{K}}(F_7)$, plotted every 500, iterations, as red diamonds. The straight horizontal line represents the estimate obtained after 5 million iterations.

7   Theory

In this section we give precise conditions under which the asymptotics developed in Sections 2 and 5 are rigorously justified. The results together with their detailed assumptions are stated below and the proofs are contained in the appendix. Note that, since the first two estimators we considered, $\mu_{n,\mathrm{rev},1}(F)$ and $\mu_{n,\mathrm{rev},2}(F)$, are special cases of the estimators $\mu_{n,\Gamma}(F)$ and $\mu_{n,\mathrm{K}}(F)$ introduced in Section 5, here we concentrate on the more general estimators $\mu_{n,\Gamma}(F)$, $\mu_{n,\mathrm{K}}(F)$.

First we recall the basic setting from Section 2. We take $\{X_n\}$ to be a Markov chain with values in a general measurable space X equipped with a σ-algebra $\mathcal{B}$. The distribution

of $\{X_n\}$ is described by its initial state $X_0 = x \in \mathsf{X}$ and its transition kernel, $P(x, dy)$, as in (1). The kernel P, as well as any of its powers P^n, acts linearly on functions $F : \mathsf{X} \to \mathbb{R}$ via, $PF(x) = E[F(X_1)|X_0 = x]$.

Our first assumption on the chain $\{X_n\}$ is that ψ-*irreducible* and *aperiodic*. This means that there is a σ-finite measure ψ on $(\mathsf{X}, \mathcal{B})$ such that, for any $A \in \mathcal{B}$ satisfying $\psi(A) > 0$ and any initial condition x,

$$P^n(x, A) > 0, \qquad \text{for all } n \text{ sufficiently large.}$$

Without loss of generality, ψ is assumed to be *maximal* in the sense that any other such ψ' is absolutely continuous with respect to ψ.

Our second, and stronger, assumption, is an essentially minimal ergodicity condition; cf. Meyn and Tweedie (1993): We assume that there are functions $V : \mathsf{X} \to [0, \infty)$, $W : \mathsf{X} \to [1, \infty)$, a "small" set $C \in \mathcal{B}$, and a finite constant $b > 0$ such that the Lyapunov drift condition (V3) holds:

$$PV - V \leq -W + b\mathbb{I}_C. \tag{V3}$$

Recall that a set $C \in \mathcal{B}$ is small if there exists an integer $m \geq 1$, a $\delta > 0$ and a probability measure ν on $(\mathsf{X}, \mathcal{B})$ such that,

$$P^m(x, B) \geq \delta\nu(B) \quad \text{for all } x \in C, \ B \in \mathcal{B}.$$

Under (V3), we are assured that the chain is positive recurrent, and that it possesses a unique invariant (probability) measure π. Our final assumption on the chain is that the Lyapunov function V in (V3) satisfies, $\pi(V^2) < \infty$.

These assumptions are summarized as follows:

$$\left.\begin{array}{l} \text{The chain } \{X_n\} \text{ is } \psi\text{-irreducible and aperiodic, with unique invariant} \\ \text{measure } \pi, \text{ and there exist functions } V : \mathsf{X} \to [0, \infty), W : \mathsf{X} \to [1, \infty), \\ \text{a small set } C \in \mathcal{B}, \text{ and a finite constant } b > 0, \text{ such that (V3) holds} \\ \text{and } \pi(V^2) < \infty. \end{array}\right\} \tag{A}$$

Although these conditions may seem somewhat involved, their verification is generally straightforward; see the texts Meyn and Tweedie (1993); Robert and Casella (2004), as well as some of the examples developed in Roberts and Tweedie (1996); Hobert and Geyer (1998); Jarner and Hansen (2000); Fort et al. (2003); Roberts and Rosenthal (2004). In fact, it is often possible to avoid having to verify (V3) directly, by appealing to the property of geometric ergodicity, which is essentially equivalent to the requirement that (V3) holds with W being a multiple of the Lyapunov function V. For large classes of MCMC samplers, geometric ergodicity has been established in the above papers, among others. Moreover, geometrically ergodic chains, especially in the reversible case, have many attractive properties, as discussed, for example, by Roberts and Rosenthal (1998).

In the interest of generality, the main results of this section are stated in terms of the weaker (and essentially minimal) assumptions in (A). Some details on general strategies for their verification can be found in the references above.

Apart from conditions on the Markov chain $\{X_n\}$, the asymptotic results stated earlier also require some assumptions on the function $F : \mathsf{X} \to \mathbb{R}$ whose mean under π is to be estimated, and on the (possibly vector-valued) function $G : \mathsf{X} \to \mathbb{R}^k$ which is used for the control variate $U = G - PG$. These assumptions are most conveniently stated within the weighted-L_∞

framework of Meyn and Tweedie (1993). Given an arbitrary function $W : \mathsf{X} \to [1, \infty)$, the weighted-$L_\infty$ space L_∞^W is the Banach space,

$$L_\infty^W := \Big\{ \text{functions } F : \mathsf{X} \to \mathbb{R} \text{ s.t. } \|F\|_W := \sup_{x \in \mathsf{X}} \frac{|F(x)|}{W(x)} < \infty \Big\}.$$

With a slight abuse of notation, we say that a vector-valued function $G = (G_1, G_2, \ldots, G_k)^t$ is in L_∞^W if $F_j \in L_\infty^W$ for each j.

Theorem 1. *Suppose the chain $\{X_n\}$ satisfies conditions* (A), *and let $\{\theta_n\}$ be any sequence of random vectors in $\mathbb{R}^k$ such that θ_n converge to some constant $\theta \in \mathbb{R}^k$ a.s., as $n \to \infty$. Then:*

(i) [ERGODICITY] *The chain is positive Harris recurrent, it has a unique invariant (probability) measure π, and it converges in distribution to π, in that for any $x \in \mathsf{X}$ and $A \in \mathcal{B}$,*

$$P^n(x, A) \to \pi(A), \quad as \ n \to \infty.$$

In fact, there exists a finite constant B such that,

$$\sum_{n=0}^{\infty} |P^n F(x) - \pi(F)| \le B(V(x) + 1), \tag{38}$$

uniformly over all initial states $x \in \mathsf{X}$ and all function F such that $|F| \le W$.

(ii) [LLN] *For any $F, G \in L_\infty^W$ and any $\vartheta \in \mathbb{R}^k$, write $U = G - PG$ and $F_\vartheta := F - \langle \vartheta, U \rangle$. Then the ergodic averages $\mu_n(F)$, as well as the adaptive averages $\mu_n(F_{\theta_n})$, both converge to $\pi(F)$ a.s., as $n \to \infty$.*

(iii) [POISSON EQUATION] *If $F \in L_\infty^W$, then there exists a solution $\hat{F} \in L_\infty^{V+1}$ to the Poisson equation, $P\hat{F} - \hat{F} = -F + \pi(F)$, and $\hat{F}$ is unique up to an additive constant.*

(iv) [CLT FOR $\mu_n(F)$] *If $F \in L_\infty^W$ and the variance, $\sigma_F^2 := \pi(\hat{F}^2 - (P\hat{F})^2)$ is nonzero, then the normalized ergodic averages $\sqrt{n}[\mu_n(F) - \pi(F)]$ converge in distribution to $N(0, \sigma_F^2)$, as $n \to \infty$.*

(v) [CLT FOR $\mu_n(F_{\theta_n})$] *If $F, G \in L_\infty^W$, and the variances, $\sigma_{F_\theta}^2 := \pi(\hat{F}_\theta^2 - (P\hat{F}_\theta)^2)$ and $\sigma_{U_j}^2 := \pi(\hat{U}_j^2 - (P\hat{U}_j)^2)$, $j = 1, 2, \ldots, k$ are all nonzero, then the normalized adaptive averages $\sqrt{n}[\mu_n(F_{\theta_n}) - \pi(F)]$ converge in distribution to $N(0, \sigma_{F_\theta}^2)$, as $n \to \infty$.*

Suppose the chain $\{X_n\}$ satisfies conditions (A) above, and that the functions F and $G = (G_1, G_2, \ldots, G_k)^t$ are in L_∞^W. Theorem 1 states that the ergodic averages $\mu_n(F)$ as well as the modified averages $\mu_n(F_\theta)$ based on the vector of control variates $U = G - PG$ both converge to $\pi(F)$, and both are asymptotically Normal. Next we examine the choice of the parameter vector $\theta = \theta^*$ which minimizes the limiting variance $\sigma_{F_\theta}^2$ of the modified averages, and the asymptotic behavior of the estimators $\hat{\theta}_{n,\Gamma}$ and $\hat{\theta}_{n,\mathrm{K}}$ for θ^*.

As in Section 5, let $\Gamma(G)$ denote the $k \times k$ matrix with entries, $\Gamma(G)_{ij} = \pi(G_i G_j - (PG_i)(PG_j))$, and recall that, according to Theorem 1, there exists a solution $\hat{F}$ to the Poisson equation for F. The simple computation outlined in Section 5 (and justified in the proof of Theorem 2) leading to equation (26) shows that the variance $\sigma_{F_\theta}^2$ is minimized by the choice,

$$\theta^* = \Gamma(G)^{-1} \pi(\hat{F}G - (P\hat{F})(PG)),$$

39

as long as the matrix $\Gamma(G)$ is invertible. Our next result establishes the a.s. consistency of the estimators,

$$
\begin{aligned}
\hat{\theta}_{n,\Gamma} &= \Gamma_n(G)^{-1}[\mu_n(F(G+PG)) - \mu_n(F)\mu_n(G+PG)] \\
\hat{\theta}_{n,K} &= K_n(G)^{-1}[\mu_n(F(G+PG)) - \mu_n(F)\mu_n(G+PG)],
\end{aligned}
$$

where the $k \times k$ matrices $\Gamma_n(G)$ and $K_n(G)$ are defined, respectively, by,

$$
\begin{aligned}
(\Gamma_n(G))_{ij} &= \mu_n(G_i G_j) - \mu_n((PG_i)(PG_j)) \\
\text{and} \quad (K_n(G))_{ij} &= \frac{1}{n}\sum_{t=0}^{n-1}(G_i(X_t) - PG_i(X_{t-1}))(G_j(X_t) - PG_j(X_{t-1})).
\end{aligned}
$$

Theorem 2. *Suppose that the chain $\{X_n\}$ is reversible and satisfies conditions (A). If the functions F, G are both in L_∞^W and the matrix $\Gamma(G)$ is nonsingular, then both the adaptive estimators for θ^* are a.s. consistent:*

$$
\begin{aligned}
\hat{\theta}_{n,\Gamma} &\to \theta^* &\quad \text{a.s., as } n \to \infty; \\
\hat{\theta}_{n,K} &\to \theta^* &\quad \text{a.s., as } n \to \infty.
\end{aligned}
$$

Recall the definitions of the two estimators $\mu_{n,\Gamma}(F)$ and $\mu_{n,K}(F)$ from equations (31) and (32) in Section 5. Combining the two theorems, yields the desired asymptotic properties of the two estimators:

Corollary 1. *Suppose that the chain $\{X_n\}$ is reversible and satisfies conditions (A). If the functions F, G are both in L_∞^W and the matrix $\Gamma(G)$ is nonsingular, then the adaptive estimators $\mu_{n,\Gamma}(F)$ and $\mu_{n,K}(F)$ for $\pi(F)$ satisfy:*

(i) [LLN] *The adaptive estimators $\mu_{n,\Gamma}(F)$, $\mu_{n,K}(F)$ both converge to $\pi(F)$ a.s., as $n \to \infty$.*

(ii) [CLT] *If $\sigma_{F_{\theta^*}}^2 := \pi(\hat{F}_{\theta^*}^2 - (P\hat{F}_{\theta^*})^2)$ is nonzero, then the normalized adaptive averages $\sqrt{n}[\mu_{n,\Gamma}(F) - \pi(F)]$ and $\sqrt{n}[\mu_{n,K}(F) - \pi(F)]$ converge in distribution to $N(0, \sigma_{F_{\theta^*}}^2)$, as $n \to \infty$, where the variance $\sigma_{\theta^*}^2$ is minimal among all estimators based on the control variate $U = G - PG$, in that $\sigma_{\theta^*}^2 = \min_{\theta \in \mathbb{R}^k} \sigma_\theta^2$.*

Some additional results on the long-term behavior of estimators similar to the ones considered above can be found in Meyn's recent work in Meyn (2006) and Meyn (2007, Chapter 11).

8 Extensions

We have presented a series of small and large simulation experiments motivated by important classes of Bayesian inference problems, and we have repeatedly observed that generally straightforward choices for functions G in the construction of control variates $U = G - PG$ provide very effective variance reduction results. Moreover, the methodology utilizing these control variates can be implemented as an essentially black-box, post-processing algorithm.

The theory presented is applicable to any reversible Markov chain. Our focus here has been primarily on cases of samplers for which we can find *some* functions G such that the one-step conditional expectations required for the computation of PG are available in closed form. These are readily available in all conjugate Gibbs and discrete Metropolis algorithms, as well as in most Markovian models for stochastic networks; see Meyn (2007) and the references therein.

Most of our experiments were performed using random-scan Gibbs samplers, in order to maintain reversibility; this is not necessarily a restrictive choice, since the convergence properties of random-scan algorithms are comparable to those of systematic-scan samplers; see Roberts and Sahu (1997). Moreover, any implementation technique that can facilitate or speed up the MCMC convergence (such as blocking schemes, transformations, other reversible chains, and so on), can be used, as long as reversibility is maintained.

There are many other Gibbs sampling algorithms in which full conditional density expectations are analytically available and, therefore, our proposed methodology is immediately applicable. Apart from the natural extensions of the examples in Section 3, we emphasize that conjugate Gibbs sampling is the key ingredient in Bayesian inference for dynamic linear models, see Reis et al. (2006); slice Gibbs auxiliary variables applications, see Damien et al. (1999); Dirichlet processes, see MacEachern and Muller (1998); and spatial regression models, see Gamermanan et al. (2003).

Metropolis-Hastings algorithms were used in Example 5 and in the two-threshold autoregressive model in Section 6.3. In both cases, the samplers operate on a discrete state space, making it possible to compute PG directly in closed form. It may be worth emphasizing that for any discrete Metropolis-Hastings sampler where the number of possible proposed moves is not prohibitively large, the function PG can be easily analytically obtained for any choice of G.

In closing, we note that the main obstacle in the immediate applicability of our methodology is the presence of the accept/reject probability in Metropolis-Hastings steps with continuous proposals. The ways in which this methodology can be applied in such cases are explored in ongoing work that investigates this issue in detail, and which will be reported in Dellaportas et al. (2008).

Acknowledgments

Appendix: Proofs of Theorems 1 and 2

PROOF OF THEOREM 1. Since any small set is petite, Meyn and Tweedie (1993, Section 5.5.2), the f-norm ergodic theorem of Meyn and Tweedie (1993) implies that $\{X_n\}$ is positive recurrent with a unique invariant measure π such that (38) holds, and Meyn and Tweedie (1993, Theorem 11.3.4) proves the Harris property, giving (i).

From Meyn and Tweedie (1993, Theorem 14.0.1) we have that, under (V3), $\pi(W) < \infty$. Since F is in L_∞^W, $\pi(|F|)$ is finite, and since $G \in L_\infty^W$, Jensen's inequality guarantees that $\pi(|U|)$ is finite. The invariance of π then implies that $\pi(U) = 0$; therefore, Meyn and Tweedie (1993, Theorem 17.0.1) shows that $\mu_n(F) \to \pi(F)$ and $\mu_n(U) \to 0$ a.s. as $n \to \infty$, and since $\theta_n \to \theta$ by assumption, $\mu_n(F_\theta)$ also converges to $\pi(F)$ a.s., proving (ii).

The existence of a solution $\hat{F}$ to the Poisson equation in (iii) follows from Meyn and Tweedie (1993, Theorem 17.4.2), and its uniqueness from Meyn and Tweedie (1993, Theorem 17.4.1). The CLT in (iv) is a consequence of Meyn and Tweedie (1993, Theorem 17.4.4).

Finally, since $F, G \in L_\infty^W$, the functions U and F_θ are in L_∞^W too, so $\hat{U}_j$ and $\hat{F}_\theta$ exist for each $j = 1, 2, \ldots, k$. As in (iv), the scaled averages $\sqrt{n}[\mu_n(F_\theta) - \pi(F)]$ and $\sqrt{n}\mu_n(U_j)$ converge in distribution to $N(0, \sigma_{F_\theta}^2)$ and $N(0, \sigma_{U_j}^2)$, respectively, for each j, where the variances $\sigma_{F_\theta}^2$ and $\sigma_{U_j}^2$ are as in (iii). Writing $\theta = (\theta_1, \theta_2, \ldots, \theta_k)^t$ and $\theta_n = (\theta_{n,1}, \theta_{n,2}, \ldots, \theta_{n,k})^t$, we can express,

$$\sqrt{n}[\mu_n(F_{\theta_n}) - \pi(F)] = \sqrt{n}[\mu_n(F_\theta) - \pi(F)] + \sum_{j=1}^k \left\{ (\theta_{n,j} - \theta_j)\sqrt{n}\mu_n(U_j) \right\}.$$

Each of the terms in the second sum on the right-hand-side above converges to zero in probability, since and $\sqrt{n}\mu_n(U_j)$ converges to a Normal distribution and $(\theta_{n,j} - \theta_j) \to 0$ a.s. Therefore, the sum converges to zero in probability, and the CLT in (v) follows from (iv). $\qquad\square$

Note that the assumption $\sigma_{U_j}^2 \neq 0$ in the theorem is not necessary, since the case $\sigma_{U_j}^2 = 0$ is trivial in view of Kontoyiannis and Meyn (2003, Proposition 2.4), which implies that, then, $\sqrt{n}\mu_n(U_j) \to 0$ in probability, as $n \to \infty$.

PROOF OF THEOREM 2.   We begin by justifying the computations in Section 5. Define $\sigma_{F_\theta}^2 = \pi(\hat{F}_\theta^2 - (P\hat{F}_\theta)^2)$, where $\hat{F}$ exists by Theorem 1. Since $\hat{F}$ solves the Poisson equation for F, it is easy to check that $\hat{F}_\theta := \hat{F} - \langle \theta, G \rangle$ solves the Poisson equation for F_θ. Substituting this in the above expression for $\sigma_{F_\theta}^2$ yields (25). To see that all the functions in (25) are indeed integrable recall that $\hat{F} \in L_\infty^{V+1}$ and note that, since V is nonnegative, (V3) implies that $1 \leq W \leq V + b\mathbb{I}_C$, hence $\pi(W^2)$ is finite since $\pi(V^2)$ is finite by assumption. Therefore, since $G \in L_\infty^W$, $\hat{F}$ and G are both in $L_2(\pi)$, and Hölder's inequality implies that $\pi(\hat{F}\langle \theta, G \rangle)$ is finite. Finally, Jensen's inequality implies that $P\hat{F}$ and PG are also in $L_2(\pi)$, so that $\pi(P\hat{F}\langle \theta, PG \rangle) < \infty$. And, for the same reasons, all the functions appearing in the computations leading to the results of Lemma 2 and Proposition 2 are also integrable.

The expression for the optimal θ^* in (26) is simply the solution for the minimum of the quadratic in (25). Again, note that $\hat{F}, G, P\hat{F}$ and PG are all in $L_2(\pi)$ so θ^* is well-defined.

The consistency proofs follow from repeated applications of the ergodic theorems established in Theorem 1. First note that, since $G \in L_\infty^W$ and $\pi(W^2) < \infty$ as remarked above, the product G_iG_j is π-integrable, and by Jensen's inequality so is any product of the form $(PG_i)(PG_j)$. Therefore, the ergodic theorem of Meyn and Tweedie (1993, Theorem 17.0.1) implies that $\Gamma_n(G) \to \Gamma(G)$ a.s. Similarly, the functions F, G, PG, FG and FPG are all π-integrable, so that the same ergodic theorem implies that $\hat{\theta}_{n,\Gamma}$ indeed converges to θ^* a.s., as $n \to \infty$.

To establish the corresponding result for $\hat{\theta}_{n,K}$, it suffices to show that $K_n(G) \to K(G)$ a.s., and to that end we consider the bivariate chain $Y_n = (X_n, X_{n+1})$ on the state space $\mathsf{X} \times \mathsf{X}$. Since $\{X_n\}$ is ψ-irreducible and aperiodic, $\{Y_n\}$ is $\psi^{(2)}$-irreducible and aperiodic with respect to the bivariate measure $\psi^{(2)}(dx, dx') := \psi(dx)P(x, dy)$. Given functions W, V a small set C and a constant b so that (V3) holds, it is immediate that (V3) also holds for $\{Y_n\}$ with respect to the functions $V^{(2)}(x, x') = V(x')$, $W^{(2)}(x, x') = W(x')$, the small set $\mathsf{X} \times C$, and the same b. The unique invariant measure of $\{Y_n\}$ is then $\pi^{(2)}(dx, dx') := \pi(dx)P(x, dy)$, and $\pi^{(2)}(V^{(2)})$ is finite. Therefore, assumptions (A) hold for $\{X_n\}$ and, for each pair $1 \leq i, j \leq k$ we can invoke

the ergodic theorem Meyn and Tweedie (1993, Theorem 17.0.1) for the $\pi^{(2)}$-integrable function,

$$H(x, x') := (G_i(x') - PG_i(x))(G_j(x') - PG_j(x)),$$

to obtain that, indeed, $\mathrm{K}_n(G) \to \mathrm{K}(G)$ a.s. $\qquad\square$

PROOF OF COROLLARY 1. The ergodic theorems in (i) are immediate consequences of Theorem 1 (ii) combined with Theorem 2. The computation in Section 5 which shows that θ^* in (26) indeed minimizes $\sigma^2_{F_\theta}$ (justified in the proof of Theorem 2) shows that $\sigma^2_{\theta^*} = \min_{\theta \in \mathbb{R}^k} \sigma^2_\theta$. Finally, the assumption that $\Gamma(G)$ is nonsingular combined with Lemma 2, imply that all the variances $\sigma^2_{U_j}$ must be nonzero. Therefore, Theorem 2 combined with the central limit theorems in parts (iv) and (v) of Theorem 1, prove part (ii) of the Corollary. $\qquad\square$

References

Albert, J. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc. 88*(422), 669–679.

Andradòttir, S., D. Heyman, and T. Ott (1993). Variance reduction through smoothing and control variates for Markov Chain simulations. *ACM Transactions on Modeling and Computer Simulation 3*, 167–189.

Atchadé, Y. F. and F. Perron (2005). Improving on the independent Metropolis-Hastings algorithm. *Statistica Sinica 15*, 3–18.

Barone, P. and A. Frigessi (1989). Improving stochastic relaxation for Gaussian random fields. *Probability in the Engineering and Informational sciences 4*, 369–389.

Bottolo, L. and S. Richardson (2008). Evolutionary stochastic search. *Preprint*.

Cain, M. (1994). The moment-generating function of the minimum of bivariate normal random variables. *The American Statistician 48*, 124–125.

Clyde, M. and E. I. George (2004). Model uncertainty. *Statist. Sci 19*, 81–94.

Damien, P., J. Wakefield, and S. Walker (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 61*, 331–344.

Dellaportas, P., D. Denison, and C. Holmes (2007). Flexible threshold models for modelling interest rate volatility. *Econometric reviews 26*, 419–437.

Dellaportas, P. and J. J. Forster (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika 86*, 615–633.

Dellaportas, P., I. Kontoyiannis, S. Meyn, and Z. Tsourti (2008). *Manuscript in preparation*.

Dellaportas, P. and G. Roberts (2003). *An introduction to MCMC*, Volume Spatial Statistics and Computational methods,, pp. 1–42. Springer-Verlag, NY. J. Møller.

Dellaportas, P. and A. F. M. Smith (1993). Bayesian inference for generalised linear and proportional hazards models via Gibbs sampling. *Applied Statistics 42*, 443–459.

Diebolt, J. and C. Robert (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Soc. Ser. B 56*(2), 363–375.

Fan, Y., S. Brooks, and A. Gelman (2006). Output assessment for Monte Carlo simulations via the score statistic. *Journal of Computational and Graphical Statistics 15*, 178–206.

Fort, G., E. Moulines, G. Roberts, and J. Rosenthal (2003). On the geometric ergodicity of hybrid samplers. *J. Appl. Probab. 40*(1), 123–146.

Gamermanan, D., A. R. Moreirab, and H. Rue (2003). Space-varying regression models: specifications and simulation. *Computational Statistics and data analysis 42*, 513–533.

Gelfand, A. E., S. E. Hills, A. Racine-Poon, and A. F. M. Smith (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association 85*(412), 972–985.

Gelfand, A. E. and A. F. M. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association 85*, 398–409.

Gilks, W. (1986). Full conditional distributions. In *Markov Chain Monte Carlo in Practice (Gilks, W.R., Richardson, S. and Spiegelhalter, D., Edts.)*, pp. 75–88. London: Chapman and Hall.

Givens, G. and J. Hoeting (2005). *Computational Statistics*. Hoboken, NJ: John Wiley & Sons.

Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*. New York: Springer-Verlag.

Glynn, P. and R. Szechtman (2002). Some new perspectives on the method of control variates. In *Monte Carlo and quasi-Monte Carlo methods, 2000 (Hong Kong)*, pp. 27–49. Berlin: Springer.

Green, P. and X. Han (1992). *Metropolis methods, Gaussian proposals, and antithetic variables*, Volume 74, Lecture Notes in Statistics, Stochastic Methods and Algorithms in Image Analysis, pp. 142–164. Spinger Verlag, Berlin. Eds: P. Barone, A. Frigessi and M. Piccioni.

Hammer, H. and Håkon (2008). Control variates for the Metropolis-Hastings algorithm. *Scandinavian Journal of Statistics 35*, 400–414.

Henderson, S. (1997). *Variance Reduction Via an Approximating Markov Process*. Ph. D. thesis, Department of Operations Research, Stanford University, Stanford, CA.

Henderson, S. and P. Glynn (2002). Approximating martingales for variance reduction in Markov process simulation. *Math. Oper. Res. 27*(2), 253–271.

Henderson, S., S. Meyn, and V. Tadić (2003). Performance evaluation and policy selection in multiclass networks. *Discrete Event Dyn. Syst. 13*(1-2), 149–189. Special issue on learning, optimization and decision making.

Henderson, S. and B. Simon (2004). Adaptive simulation using perfect control variates. *J. Appl. Probab. 41*(3), 859–876.

Hobert, J. and C. Geyer (1998). Geometric ergodicity of Gibbs and block Gibbs samplers for a hierarchical random effects model. *J. Multivariate Anal. 67*(2), 414–430.

Jarner, S. and E. Hansen (2000). Geometric ergodicity of Metropolis algorithms. *Stochastic Process. Appl. 85*(2), 341–361.

Jasra, A., C. Holmes, and D. Stephens (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modelling. *Statistical Science 20*, 50–67.

Johnson, V. and J. Albert (1999). *Ordinal Data Modeling*. New York: Springer-Verlag.

Kim, S. and S. Henderson (2007). Adaptive control variates for finite-horizon simulation. *Math. Oper. Res. 32*(3), 508–527.

Knuiman, M. and T. P. Speed (1988). Incorporating prior information into the analysis of contingency tables. *Biometrics 44*, 1061–1071.

Kontoyiannis, I. and S. Meyn (2003). Spectral theory and limit theorems for geometrically ergodic Markov processes. *Ann. Appl. Probab. 13*, 304–362.

Lee, K., J.-M. Marin, K. Mengersen, and C. Robert (2008). *Bayesian Inference on Mixtures of Distributions*, Volume Proceedings of the Platinum Jubilee of the Indian Statistical Institute (to appear). Elsevier, NY. D. Dey and C.R. Rao.

Liu, J. (2001). *Monte Carlo Strategies in Scientific Computing*. New York: Springer-Verlag.

MacEachern, S. N. and P. Muller (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics 7*, 223–238.

Marin, J.-M., K. Mengersen, and C. Robert (2005). *Bayesian modelling and inference on mixtures of distributions*, Volume Handbook of Statistics: Volume 25. Elsevier, NY. D. Dey and C.R. Rao.

Mengersen, K. L., C. P. Robert, and C. Guihenneuc-Jouyaux (1999). MCMC convergence diagnostics: A reviewww. In A. P. D. J. M. Bernardo, J. O. Berger and A. F. M. Smith (Eds.), *Bayesian Statistics 6*, pp. 415–440. Oxford University Press, UK.

Meyn, S. (2006). Large deviation asymptotics and control variates for simulating large functions. *Ann. Appl. Probab. 16*(1), 310–339.

Meyn, S. (2007). *Control Techniques for Complex Networks*. Cambridge University Press.

Meyn, S. and R. Tweedie (1993). *Markov Chains and Stochastic Stability*. London: Springer-Verlag.

Mira, A., P. Tenconi, and D. Bressanini (2003). Variance reduction for MCMC. *Technical report 2003/29, Universitá degli Studi dell' Insubria, Italy*.

Müller, P. (1993). Alternatives to the Gibbs sampling scheme. Technical report, Institute of Statistics and Decision Sciences, Duke University.

O'Hagan, A. and J. Forster (2004). *Bayesian Inference*. Arnold, London.

Pfann, G., S. P.C., and R. Tschernig (1996). Nonlinear interest rate dynamics and implications for the term structure. *Journal of Econometrics 74*, 149–176.

Philippe, A. and C. Robert (2001). Riemann sums for MCMC estimation and convergence monitoring. *Statistics and Computing 11*, 103–115.

Reis, E. A., E. Salazar, and D. Gamerman (2006). Comparison of sampling schemes for dynamic linear models. *International Statistics Review 74*, 203–214.

Richardson, S. and P. J. Green (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B 59*, 731–792.

Robert, C. and G. Casella (2004). *Monte Carlo Statistical Methods* (Second ed.). New York: Springer-Verlag.

Roberts, G. (1992). Convergence diagnostics of the Gibbs sampler. In J. M. Bernardo, J. Berger, A. Dawid, and A. Smith (Eds.), *Bayesian Statistics 4*, pp. 775–782. Oxford University Press, UK.

Roberts, G. and J. Rosenthal (1998). Markov-chain Monte Carlo: some practical implications of theoretical results. *Canad. J. Statist. 26*(1), 5–31. With discussion by Hemant Ishwaran and Neal Madras and a rejoinder by the authors.

Roberts, G. and J. Rosenthal (2004). General state space Markov chains and MCMC algorithms. *Probab. Surv. 1*, 20–71 (electronic).

Roberts, G. and J. Rosenthal (2006). Examples of adaptive MCMC. *Preprint*. [Available online at: probability.ca/jeff/].

Roberts, G. and R. Tweedie (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika 83*(1), 95–110.

Roberts, G. O. and S. K. Sahu (1997). Updating schemes, correlation structure, blocking and parameterisation for the Gibbs sampler. *Journal of the Royal Statistical Society, Series B 59*, 291–397.

Stephens, M. (1997). *Bayesian Methods for Mixtures of Normal Distributions*. Ph. D. thesis, University of Oxford, UK.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association 82*, 528–540.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist. 22*(4), 1701–1762. With discussion.