Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2012)
74, *Part* 1, *pp.* 133–161

Control variates for estimation based on reversible Markov chain Monte Carlo samplers

Petros Dellaportas and Ioannis Kontoyiannis

Athens University of Economics and Business, Greece

[Received May 2011. Revised July 2011]

Summary. A general methodology is introduced for the construction and effective application of control variates to estimation problems involving data from reversible Markov chain Monte Carlo samplers. We propose the use of a specific class of functions as control variates, and we introduce a new consistent estimator for the values of the coefficients of the optimal linear combination of these functions. For a specific Markov chain Monte Carlo scenario, the form and proposed construction of the control variates is shown to provide an exact solution of the associated Poisson equation. This implies that the estimation variance in this case (in the central limit theorem regime) is exactly zero. The new estimator is derived from a novel, finite dimensional, explicit representation for the optimal coefficients. The resulting variance reduction methodology is primarily (though certainly not exclusively) applicable when the simulated data are generated by a random-scan Gibbs sampler. Markov chain Monte Carlo examples of Bayesian inference problems demonstrate that the corresponding reduction in the estimation variance is significant, and that in some cases it can be quite dramatic. Extensions of this methodology are discussed and simulation examples are presented illustrating the utility of the methods proposed. All methodological and asymptotic arguments are rigorously justified under essentially minimal conditions.

Keywords: Bayesian inference; Control variates; Hierarchical normal linear model; Log-linear model; Markov chain Monte Carlo methods; Mixtures of normals; Poisson equation; Threshold auto-regressive model; Variance reduction

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 methods perhaps the main reason for the widespread use of Bayesian statistical modelling and inference across the spectrum of quantitative scientific disciplines.

This work provides a methodological foundation for the construction and use of control variates in conjunction with reversible MCMC samplers. Although popular in the standard Monte Carlo setting, control variates have received much less attention in the MCMC literature. The methodology proposed will be shown, both via theoretical results and simulation examples, to reduce the variance of the resulting estimators significantly, and sometimes 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 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

Address for correspondence: Ioannis Kontoyiannis, Department of Informatics, Athens University of Economics and Business, Kontrikton Campus, Patission 76, Athens 10434, Greece.
E-mail: yiannis@aueb.gr

 1369–7412/12/74133

which it is known that the expected value of each $U_j(X_1)$ is equal to 0, then subtracting any linear combination $\theta_1 U_1(X_i) + \theta_2 U_2(X_i) + \ldots + \theta_k U_k(X_i)$ from 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 $\{\theta_j^*\}$ based on the same sequence of samples; see, for example, Liu (2001), Robert and Casella (2004) or Givens and Hoerting (2005). 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 has been devoted to the construction of effective control variates in specific applied problems.

The main difficulty in extending the above methodology to estimators based on MCMC samples is probably due to the intrinsic complexities that are presented by the Markovian structure. However, it is difficult to find non-trivial useful functions with known expectation with respect to the stationary distribution of the chain (for example, Mengersen *et al.* (1999) commented that 'control variates have been advertised early in the MCMC literature (see, for example, 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'), and, even in cases where such functions are available, there has been no effective way to obtain consistent estimates of the corresponding optimal coefficients $\{\theta_j^*\}$. An important underlying reason for both of these difficulties is the basic fact that the MCMC variance of ergodic averages is intrinsically an infinite dimensional object: it cannot be expressed in closed form as a function of the transition kernel and the stationary distribution of the chain.

An early reference for variance reduction for Markov chain samplers is Green and Han (1992), who exploited an idea of Barone and Frigessi (1989) and constructed antithetic variables that may achieve variance reduction in simple settings but do not appear to be widely applicable. Andradòttir *et al.* (1993) focused on finite state space chains, they observeed that optimum variance reduction can be achieved via the solution of the associated *Poisson equation* (see equation (3) below and Section 2.1 for details) and they proposed numerical algorithms for its solution. Rao–Blackwellization 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) mentioned that it can be used as a control variate, its practical utility has not been investigated. Atchadé and Perron (2005) restricted attention to independent Metropolis samplers and provided an explicit formula for the construction of control variates. Hammer and Tjelmeland (2008) constructed control variates for general Metropolis–Hastings samplers by expanding the state space.

In a different context, and partly motivated by considerations from statistical mechanics, Assaraf and Caffarel (1999) introduced a family of control variates that they called 'zero-variance estimators'. The Assaraf–Cafarel estimators have been adapted and applied to problems in statistics by Mira *et al.* (2003, 2010) and Dalla Valle and Leisen (2010).

In most of the works cited above, the method that was used for the estimation of the optimal coefficients $\{\theta_j^*\}$ is either based on the same formula as that obtained for control variates in independent identically distributed Monte Carlo sampling, or on the method of *batch means*, but such estimators are strictly suboptimal and generally ineffective; see Section 6 for details.

For our purposes, a more relevant 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 researchers mentioned above, also noted that the best choice for the function G would be the solution of the associated Poisson equation and proceeded 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 and to use the resulting control variates in conjunction with a new, efficiently implementable and provably optimal estimator for the coefficients $\{\theta_j^*\}$. The ability to estimate the $\{\theta_j^*\}$ effectively makes these control variates practically relevant in the statistical MCMC context and avoids the need to compute analytical approximations to the solution of the underlying Poisson equation.

1.1. Outline of the proposed basic methodology

Section 2.1 introduces the general setting within which all the subsequent results are developed. A sample of size n from an ergodic Markov chain $\{X_n\}$ is used to estimate the mean $E_\pi[F] = \int F \, d\pi$ of a function F, under the unique invariant measure π of the chain. The associated *Poisson equation* is introduced, and it is shown that its solution can be used to quantify, in an essential way, the rate at which the chain converges to equilibrium.

In Section 2.2 we examine the variance of the standard ergodic averages,

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

and we compare it with the variance of the modified estimators,

$$\frac{1}{n} \sum_{i=0}^{n-1} \{F(X_i) - \theta_1 U_1(X_i) - \theta_2 U_2(X_i) - \ldots - \theta_k U_k(X_i)\}. \tag{2}$$

Here and throughout the subsequent discussion, the control variates $U_1, U_2, \ldots, U_k$, are constructed as above via $U_j(x) := G_j(x) - PG_j(x)$, where $PG(x)$ denotes the one-step expectation $E[G(X_{n+1})|X_n = x]$, for particular choices of the functions G_j, $j = 1, 2, \ldots, k$.

The two central methodological issues that are addressed in this work are

(a) the problem of estimating the optimal coefficient vector $\{\theta_j^*\}$ that minimizes the variance of the modified estimators (2) and
(b) the choice of the functions $\{G_j\}$, so that the corresponding functions $\{U_j\}$ will be effective as control variates in specific MCMC scenarios that arise from common families of Bayesian inference problems.

For the first issue, in Section 3, we derive new representations for the optimal coefficient vector $\{\theta_j^*\}$, under the assumption that the chain $\{X_n\}$ is *reversible*; see proposition 2 there. These representations lead to our first main result, namely a new estimator for $\{\theta_j^*\}$; see equations (25) and (26) in Section 3. This estimator is based on the same MCMC output and it can be used after the sample has been obtained, making its computation independent of the MCMC algorithm that is used.

The second problem, that of selecting an effective collection of functions $\{G_j\}$ for the construction of the control variates $\{U_j\}$, is more complex and it is dealt with in stages. First, in Section 2 we recall that there is always a single choice of a function G that actually makes the estimation variance equal to 0: if G satisfies

$$U := G - PG = F - E_\pi[F], \tag{3}$$

then with this control variate and with $\theta = 1$ the modified estimates in expression (2) are *equal* to the required expectation $E_\pi[F]$ for *all* n. A function G satisfying condition (3) is often called a solution to the *Poisson equation for* F (or *Green's function*). But solving the Poisson equation even for simple functions F is a highly non-trivial task, and for chains arising in typical applications it is, for all practical purposes, impossible; see, for example the relevant comments in Henderson (1997) and Meyn (2007). Therefore, as a first *rule of thumb*, we propose that a class of functions $\{G_j\}$ be chosen such that the solution to the Poisson equation (3) can be accurately approximated by a linear combination $\Sigma_j \theta_j G_j$ of the $\{G_j\}$. For this reason we call the $\{G_j\}$ *basis functions*.

Clearly there are many possible choices for the basis functions $\{G_j\}$, and the effectiveness of the resulting control variates depends on the particular choice. In Section 4 we propose a specific and immediately applicable class of $\{G_j\}$, leading to the following proposal, which is the basic methodological contribution of this work.

Suppose that $\pi(x) = \pi\{(x^{(1)}, x^{(2)}, \ldots, x^{(k)})^{\mathrm{T}}\}$ is a multivariate posterior distribution for which MCMC samples are obtained by a reversible Markov chain $\{X_n\}$. To estimate the posterior mean $\mu^{(i)}$ of the ith co-ordinate $x^{(i)}$, let $F(x) = x^{(i)}$, define basis functions G_j as the co-ordinate functions $G_j(x) = x^{(j)}$ for all components j for which $PG_j(x) = E[X_{n+1}^{(j)} | X_n = x]$ is explicitly computable and form the control variates $U_j = G_j - PG_j$.

Then estimate the optimal coefficient vector $\theta^* = \{\theta_j^*\}$ by the estimator $\hat{\theta} = \hat{\theta}_{n,K}$ given in equation (25), and estimate the posterior mean of interest $\mu^{(i)}$ by the modified estimators given in equation (27) in Section 3:

$$\mu_{n,K}(F) := \frac{1}{n} \sum_{i=0}^{n-1} \{F(X_i) - \hat{\theta}_1 U_1(X_i) - \hat{\theta}_2 U_2(X_i) - \ldots - \hat{\theta}_k U_k(X_i)\}. \tag{4}$$

As shown by the results at the end of Section 2, except in degenerate cases (when the resulting control variates $U_j(X)$ are perfectly uncorrelated with $F(X)$ when $X \sim \pi$), this methodology will always lead to estimates with a smaller variance (in the central limit theorem regime) than the standard ergodic averages $\mu_n(F)$ as in expression (1). Moreover, as we discuss next, in a particular MCMC scenario, the estimates $\mu_{n,K}(F)$ have asymptotic variance equal to 0.

There are two basic requirements for the immediate applicability of the methodology described so far; the underlying chain needs to be reversible for the estimates of the coefficient vector $\{\theta_j^*\}$ that is introduced in Section 3 to be consistent and, also, the one-step expectations $PG_j(x) := E[G_j(X_{n+1}) | X_n = x]$ that are necessary for the construction of the control variates U_j need to be explicitly computable.

Since the most commonly used class of MCMC algorithms satisfying both of these requirements is that of conditionally conjugate random-scan Gibbs samplers (following standard parlance, we call a Gibbs sampler 'conditionally conjugate' if the full conditionals of the target distribution are all known and of standard form; this, of course, is unrelated to the notion of a conjugate prior structure in the underlying Bayesian formulation), and since the most commonly used general approximation of the target distribution π arising in Bayesian inference problems is a general multivariate Gaussian, in Section 4 we examine this MCMC problem in detail and obtain our second main result: suppose that we wish to estimate the mean of one of the co-ordinates of a k-dimensional Gaussian distribution π, based on samples $X_i = (X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(k)})^{\mathrm{T}}$ generated by the random-scan Gibbs algorithm. In theorem 1 in Section 4 we show that the solution of the associated Poisson equation can always be expressed as a linear combination of the k co-ordinate functions $G_j(x) := x^{(j)}$, $x = (x^{(1)}, x^{(2)}, \ldots, x^{(k)})^{\mathrm{T}} \in \mathbb{R}^k$. Equivalently, the estimator $\mu_{n,K}(F)$ that is proposed in the basic methodology has *zero* variance in the central limit

theorem regime. This is perhaps the single most interesting case of a Markov chain arising in applications for which an explicit solution to the Poisson equation has ever been obtained.

Section 5 contains three MCMC examples using this methodology. Example 1 is a brief illustration of the result of theorem 1 in the case of a bivariate Gaussian distribution. As expected, the modified estimators (4) are seen to be much more effective than the standard ergodic averages (1), in that their variance is smaller by a factor ranging approximately between 4 and 1000, depending on the sample size. Example 2 contains an analysis of a realistic Bayesian inference problem via MCMC sampling, for a 66-parameter hierarchical normal linear model. There, we consider all 66 problems of estimating the posterior means of all the parameters, and we find that in most cases the reduction in variance resulting from the use of control variates as above is typically by a factor ranging between 5 and 30. The third example illustrates the use of the basic methodology in the case of Metropolis-within-Gibbs sampling. Even though the one-step expectation PG_j can be computed for only one of the two model parameters, we still find that the variance is reduced by a factor ranging approximately between 7 and 10.

1.1.1. Domain of applicability

The present development not only generalizes the classical method of control variates to the MCMC setting, but it also offers an important advantage. In the case of independent sampling, the control variates for each specific application need to be identified from scratch, often in an *ad hoc* fashion. In fact, for most Monte Carlo estimation problems there are *no* known functions that can be used as effective control variates. In contrast, the basic methodology that was described above provides a way of constructing a family of control variates that are immediately applicable to a wide range of MCMC problems, as long as the sampling algorithm produces a reversible chain for which the one-step expectations $PG(x) := E[G(X_{n+1})|X_n = x]$ can be explicitly computed for some simple linear functions G. MCMC algorithms with these properties form a large collection of samplers that are commonly used in Bayesian inference, including, among others, all conditionally conjugate random-scan Gibbs samplers (the main MCMC class that is considered in this work), certain versions of hybrid Metropolis-within-Gibbs algorithms (following the terminology of, for example, Robert and Casella (2004)), and certain types of Metropolis–Hastings samplers on discrete state spaces.

1.1.2. Extensions

In further work, we shall discuss extensions of the basic methodology along two directions that, in some cases, go beyond the above class of samplers, including examples of

(a) MCMC scenarios where it is more effective to use a set of basis functions that is different from those proposed in the basic methodology,
(b) non-conditionally conjugate samplers, where the conditional expectations $PG(x)$ for the class of linear basis functions cannot be computed in closed form, and
(c) certain classes of Metropolis–Hastings sampler when the basic methodology can be applied.

1.1.3. Further results

In Section 6 we first briefly discuss two other consistent estimators for the optimal coefficient vector $\{\theta_j^*\}$. One is a modified version of our earlier estimator $\hat{\theta}_{n,K}$ that is derived in Section 3, and the other was recently developed by Meyn (2007) on the basis of the so-called 'temporal difference learning' algorithm. Then in Section 6.2 we examine the most common

estimator for the optimal coefficient vector $\{\theta^*\}$ that has been used in the literature, which as mentioned earlier is based on the method of batch means. In proposition 3 in Section 6.2 we show that the resulting estimator for $\pi(F)$ is typically strictly suboptimal, and that the amount by which its variance is *larger* than the variance of our modified estimators $\mu_{n,K}(F)$ is potentially unbounded. Moreover, the batch means estimator is computationally more expensive and generally rather ineffective, often severely so. This is illustrated by revisiting the most interesting of the MCMC examples of Section 5, and comparing the performance of the batch means estimator with that of the simple ergodic averages (1) and of the modified estimator $\mu_{n,K}(F)$ in expression (4).

Section 7 provides the theoretical justifications of the asymptotic arguments in Sections 2, 3 and 6. Finally we conclude with a short summary of our results and a brief discussion of possible further extensions in Section 8, with particular emphasis on implementational issues and on the difficulties of applying the present methodology to general Metropolis–Hastings samplers.

1.1.4. Related work

We close this introduction with a few more remarks on previous related work. As mentioned earlier, Henderson (1997) took a different path towards optimizing the use of control variates for Markov chain samplers. Considering primarily continuous time processes, an approximation 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 mostly related to examples from stochastic network theory and queuing theory. Closely related approaches have been presented by Henderson and Glynn (2002) and Henderson *et al.* (2003), where the effectiveness of multiclass network control policies is evaluated via Markovian simulation. Control variates are used for variance reduction, and the optimal coefficients $\{\theta_j^*\}$ are estimated via an adaptive, stochastic gradient algorithm. General convergence properties of ergodic estimators using control variates were 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) introduced two related adaptive methods for tuning non-linear versions of the coefficients $\{\theta_j\}$, when using families of control variates that naturally admit a non-linear parameterization. They derived asymptotic properties for these estimators and presented numerical simulation results.

When the control variate $U = G - PG$ is defined in terms of a function G that can be taken as a Lyapunov function for the chain $\{X_n\}$, Meyn (2006) derived precise exponential asymptotics for the associated modified estimators. Also, Meyn (2007), chapter 11, gave a development of the general control variates methodology for Markov chain data that parallels certain parts of our presentation in Section 2 and discussed numerous related asymptotic results and implementational issues.

In a different direction, Stein *et al.* (2004) drew a connection between the use of control variates in MCMC methods and the 'exchangeable pairs' construction that is used in Stein's method for distribution approximation. They considered a natural class of functions as their control variates, and they estimated the associated coefficients $\{\theta_j\}$ by a simple version of the batch means method that is described in Section 6.2. Finally, the recent work by Delmas and Jourdain (2009) examines a particular case of Henderson's construction of a control variate in the context of Metropolis–Hastings sampling. Like Hammer and Tjelmeland (2008), Delmas and Jourdain expanded the state space to include the proposals and they first took $G = F$ and $\theta = 1$ (which, in part, explains why their waste recycling algorithm is sometimes worse than plain Metropolis sampling). They identified the solution of the Poisson equation as the optimal choice for a basis function and they sought analytical approximations. Then a general linear coefficient θ was

introduced, and for a particular version of the Metropolis algorithm the optimal value θ^* was identified analytically.

2. Control variates for Markov chains

2.1. The setting

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

$$P(x, A) := \Pr(X_{k+1} \in A | X_k = x), \qquad x \in \mathbf{X}, A \in \mathcal{B}. \tag{5}$$

As is well known, in many applications where it is desirable to compute the expectation $E_\pi[F] := \pi(F) := \int F \, \mathrm{d}\pi$ of some function $F : \mathbf{X} \to \mathbb{R}$ with respect to some probability measure π on $(\mathbf{X}, \mathcal{B})$, although the direct computation of $\pi(F)$ is impossible and we cannot even produce samples from π, it is possible to construct an easy-to-simulate Markov chain $\{X_n\}$ which has π as its unique invariant measure. Under appropriate conditions (see Section 7), 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

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

then, for any initial state x,

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

for an appropriate class of functions $F : \mathbf{X} \to \mathbb{R}$ (see Section 7). Furthermore, the rate of this convergence can be quantified by the function

$$\hat{F}(x) = \sum_{n=0}^{\infty} \{P^n F(x) - \pi(F)\}, \tag{6}$$

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

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

(To see this, at least formally, apply P to both sides of equation (6) and note that the resulting series for $P\hat{F} - \hat{F}$ becomes telescoping. Also note the usual convention that $P^0 = I$, the identity kernel.)

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,

$$\mu_n(F) := \frac{1}{n} \sum_{i=0}^{n-1} F(X_i). \tag{8}$$

Again, under appropriate conditions the ergodic theorem holds,

$$\mu_n(F) \to \pi(F), \qquad \text{almost surely, as } n \to \infty, \tag{9}$$

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

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

where σ_F^2, the asymptotic variance of F, is given by $\sigma_F^2:=\lim_{n\to\infty}\mathrm{var}_\pi\{\mu_n(F)\sqrt{n}\}$. Alternatively, it can be expressed in terms of $\hat{F}$ as

$$\sigma_F^2=\pi\{\hat{F}^2-(P\hat{F})^2\}. \tag{10}$$

The results in equations (6) and (10) 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 non-trivial task, and, for chains arising in typical applications, it is impossible for all practical purposes; see, for example, the relevant comments in Henderson (1997) and Meyn (2007). Nevertheless, the function $\hat{F}$ will play a central role throughout our subsequent development.

2.2. Control variates

Suppose that, for some Markov chain $\{X_n\}$ with transition kernel P and invariant measure π, the ergodic averages $\mu_n(F)$ as in expression (8) are used to estimate the mean $\pi(F)=\int F\,\mathrm{d}\pi$ 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.

To reduce the variance, we employ the idea of using control variates, as in the case of simple Monte Carlo sampling with independent and identically distributed samples; see, for example, the standard texts of Robert and Casella (2004), Liu (2001) and Givens and Hoeting (2005) or Glynn and Szechtman (2002) for extensive discussions. Given one or more functions $U_1,U_2,\ldots,U_k$, the *control variates*, such that $U_j:\mathbf{X}\to\mathbb{R}$ and $\pi(U_j)=0$ for all $j=1,2,\ldots,k$, let $\theta=(\theta_1,\theta_2,\ldots,\theta_k)^{\mathrm{T}}$ be an arbitrary, constant vector in $\mathbb{R}^k$, and define

$$F_\theta:=F-\langle\theta,U\rangle=F-\sum_{j=1}^{k}\theta_jU_j, \tag{11}$$

where $U:\mathbf{X}\to\mathbb{R}^k$ denotes the column vector, $U=(U_1,U_2,\ldots,U_k)^{\mathrm{T}}$. (Here and throughout the paper all vectors are column vectors unless explicitly stated otherwise, and $\langle\cdot,\cdot\rangle$ denotes the usual Euclidean inner product.)

We consider the modified estimators

$$\mu_n(F_\theta)=\mu_n(F)-\langle\theta,\mu_n(U)\rangle=\mu_n(F)-\sum_{j=1}^{k}\theta_j\mu_n(U_j), \tag{12}$$

for $\pi(F)$. The ergodic theorem (9) guarantees that the estimators $\{\mu_n(F_\theta)\}$ are consistent with probability 1, and it is natural to seek particular choices for U and θ so that the asymptotic variance $\sigma_{F_\theta}^2$ of the modified estimators is significantly smaller than the variance σ_F^2 of the standard ergodic averages $\mu_n(F)$.

Throughout this work, we shall concentrate exclusively on the following class of control variates U proposed by Henderson (1997). For arbitrary (π-integrable) functions $G_j:\mathbf{X}\to\mathbb{R}$ define

$$U_j:=G_j-PG_j,\qquad j=1,2,\ldots,k.$$

Then the invariance of π under P and the integrability of G_j guarantee that $\pi(U_j)=0$.

In the remainder of this section we derive some simple, general guidelines for choosing functions $\{G_j\}$ that produce effective control variates $\{U_j\}$. This issue is revisited in more detail in the Bayesian MCMC context in Section 4.

Suppose, at first, that we have complete freedom in the choice of the functions $\{G_j\}$, so that we may take $k=1$, a single $U=G-PG$ and $\theta=1$ without loss of generality. Then the goal is to make the asymptotic variance of $F-U=F-G+PG$ as small as possible. But, in view of the Poisson equation (7), we see that the choice $G=\hat{F}$ yields

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

which has zero variance. Therefore, the general principle for selecting a single function G is

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

As mentioned above, it is typically impossible to compute $\hat{F}$ for realistic models that are used in applications. But it is often possible to come up with a guess G that approximates $\hat{F}$, or at least with a collection of functions $\{G_j\}$ such that $\hat{F}$ can be approximated as a linear combination of the $\{G_j\}$. Thus, our first concrete rule of thumb for choosing $\{G_j\}$ states

> choose control variates $U_j=G_j-PG_j$, $j=1,2,\ldots,k$, with respect to a collection of *basis functions* $\{G_j\}$, such that $\hat{F}$ can be approximately expressed as a linear combination of the $\{G_j\}$.

The terminology *basis functions* for the $\{G_j\}$ is meant to emphasize the fact that, although $\hat{F}$ is not known, it is expected that it can be approximately expressed in terms of the $\{G_j\}$ via a linear expansion of the form $\hat{F}\approx\Sigma_{j=1}^{k}\theta_j G_j$.

Once the basis functions $\{G_j\}$ have been selected, we form the modified estimators $\mu_n(F_\theta)$ with respect to the function F_θ as in expression (11):

$$F_\theta=F-\langle\theta,U\rangle=F-\langle\theta,G\rangle+\langle\theta,PG\rangle,$$

where, for a vector of functions $G=(G_1,G_2,\ldots,G_k)^{\mathrm{T}}$, we write PG for the corresponding vector $(PG_1,PG_2,\ldots,PG_k)^{\mathrm{T}}$. The next task is to choose θ so that the resulting variance

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

is minimized. From the definition of U and the statement of the Poisson equation, it is immediate that

$$\hat{U}_j=G_j \qquad \text{for each } j$$

and

$$\hat{F}_\theta=\hat{F}-\langle\theta,G\rangle.$$

Therefore, by equation (10) and linearity,

$$\sigma_\theta^2=\sigma_F^2-2\pi\{\hat{F}\langle\theta,G\rangle-P\hat{F}\langle\theta,PG\rangle\}+\pi\{\langle\theta,G\rangle^2-\langle\theta,PG\rangle^2\}. \tag{13}$$

To find the optimal θ^* which minimizes the variance σ_θ^2, differentiating the quadratic σ_θ^2 with respect to each θ_j and setting the derivative equal to 0, yields, in matrix notation,

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

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

$$\theta^*=\Gamma(G)^{-1}\pi\{\hat{F}G-(P\hat{F})(PG)\}, \tag{14}$$

as long as $\Gamma(G)$ is invertible. Once again, this expression depends on $\hat{F}$, so it is not immediately clear how to estimate θ^* directly from the data $\{X_n\}$. The issue of estimating the optimal coefficient vector θ^* is addressed in detail in Section 3; but first let us interpret θ^*.

For simplicity, consider again the case of a single control variate $U = G - PG$ based on a single function G. Then the value of θ^* in equation (14) simplifies to

$$\theta^* = \frac{\pi\{\hat{F}G - (P\hat{F})(PG)\}}{\pi\{G^2 - (PG)^2\}} = \frac{\pi\{\hat{F}G - (P\hat{F})(PG)\}}{\sigma_U^2}, \tag{15}$$

where the second equality follows from the earlier observation that $\hat{U} = G$. Alternatively, starting from the expression $\sigma_\theta^2 = \lim_{n\to\infty} \mathrm{var}_\pi\{\mu_n(F_\theta)\sqrt{n}\}$, simple calculations lead to

$$\sigma_\theta^2 = \sigma_F^2 + \theta^2\sigma_U^2 - 2\theta \sum_{n=-\infty}^{\infty} \mathrm{cov}_\pi\{F(X_0), U(X_n)\}, \tag{16}$$

so θ^* can also be expressed as

$$\theta^* = \frac{1}{\sigma_U^2} \sum_{n=-\infty}^{\infty} \mathrm{cov}_\pi\{F(X_0), U(X_n)\}, \tag{17}$$

where cov_π denotes the covariance for the stationary version of the chain, i.e., since $\pi(U) = 0$, we have $\mathrm{cov}_\pi\{F(X_0), U(X_n)\} = E_\pi[F(X_0)\,U(X_n)]$, where $X_0 \sim \pi$. Then equation (17) leads to the optimal asymptotic variance,

$$\sigma_{\theta^*}^2 = \sigma_F^2 - \frac{1}{\sigma_U^2}\left[\sum_{n=-\infty}^{\infty} \mathrm{cov}_\pi\{F(X_0), U(X_n)\}\right]^2. \tag{18}$$

Therefore, to reduce the variance, it is desirable that the correlation between F and U be as large as possible. This leads to our second rule of thumb for selecting basis functions:

choose control variates $U = G - PG$ so that each U_j is highly correlated with F.

Incidentally, note that comparing the expressions for θ^* in equations (15) and (17) implies that

$$\sum_{n=-\infty}^{\infty} \mathrm{cov}_\pi\{F(X_0), U(X_n)\} = \pi\{\hat{F}G - (P\hat{F})(PG)\}. \tag{19}$$

3. Estimating the optimal coefficient vector θ^*

Consider, as before, the problem of estimating the mean $\pi(F)$ of a function $F : \mathbf{X} \to \mathbb{R}$ based on samples from an ergodic Markov chain $\{X_n\}$ with unique invariant measure π and transition kernel P. Instead of using the ergodic averages $\mu_n(F)$ as in expression (8), we select a collection of *basis functions* $\{G_j\}$ and form the control variates $U_j = G_j - PG_j$, $j = 1, 2, \ldots, k$. The mean $\pi(F)$ is then estimated by the modified estimators $\mu_n(F_\theta)$ as in equation (12), for a given coefficient vector $\theta \in \mathbb{R}^k$.

Under the additional assumption of reversibility, in this section we introduce a consistent procedure for estimating the optimal coefficient vector θ^* on the basis of the same sample $\{X_n\}$. Then in Section 4 we give more detailed guidelines for choosing the $\{G_j\}$.

Recall that, once the basis functions $\{G_j\}$ have been selected, the optimal coefficient vector θ^* was expressed in equation (14) as $\theta^* = \Gamma(G)^{-1}\pi\{\hat{F}G - (P\hat{F})(PG)\}$, where $\Gamma(G)_{ij} = \pi\{G_iG_j - (PG_i)(PG_j)\}$, $1 \leqslant i, j \leqslant k$. But, in view of equation (19) derived above, the entries $\Gamma(G)_{ij}$ can also be written

$$\begin{aligned}\Gamma(G)_{ij} &:= \pi\{G_iG_j - (PG_i)(PG_j)\} \\ &= \pi\{\hat{U}_iG_j - (P\hat{U}_i)(PG_j)\} = \sum_{n=-\infty}^{\infty} \mathrm{cov}_\pi\{U_i(X_0), G_j(X_n)\}.\end{aligned} \tag{20}$$

This indicates that $\Gamma(G)$ has the structure of a covariance matrix and, in particular, it suggests that $\Gamma(G)$ should be positive semidefinite. Indeed we have the following proposition.

Proposition 1. Let $K(G)$ denote the covariance matrix of the random variables

$$Y_j := G_j(X_1) - PG_j(X_0), \qquad j = 1, 2, \ldots, k,$$

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

$$\pi\{G_i G_j - (PG_i)(PG_j)\} = K(G)_{ij} := E_\pi[\{G_i(X_1) - PG_i(X_0)\}\{G_j(X_1) - PG_j(X_0)\}]. \tag{21}$$

Proof. Expanding the right-hand side of equation (21) yields

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

and the result follows on 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

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

and similarly for the third term. □

Therefore, using proposition 1 the optimal coefficient vector θ^* can also be expressed as

$$\theta^* = K(G)^{-1} \pi\{\hat{F}G - (P\hat{F})(PG)\}. \tag{22}$$

Now assume that the chain $\{X_n\}$ is reversible. Writing $\Delta = P - I$ for the *generator* of $\{X_n\}$, reversibility is equivalent to the statement that Δ is self-adjoint as a linear operator on the space $L_2(\pi)$. In other words,

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

for any two functions $F, G \in L_2(\pi)$. Our first main theoretical result is that the optimal coefficient vector θ^* admits a representation that does not involve the solution $\hat{F}$ of the associated Poisson equation for F, as follows.

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

$$\theta^* = \theta^*_{\text{rev}} = \Gamma(G)^{-1} \pi[\{F - \pi(F)\}(G + PG)], \tag{23}$$

or, alternatively,

$$\theta^*_{\text{rev}} = K(G)^{-1} \pi[\{F - \pi(F)\}(G + PG)], \tag{24}$$

where the matrices $\Gamma(G)$ and $K(G)$ are defined in expressions (20) and (21) respectively.

Proof. Let $\bar{F} = F - \pi(F)$ denote the centred version of F, and recall that $\hat{F}$ solves Poisson's equation for F, so $P\hat{F} = \hat{F} - \bar{F}$. Therefore, using the fact that Δ is self-adjoint on each component of G,

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

Combining this with equations (14) and (22) respectively proves the two claims (23) and (24) of the proposition. □

Expression (24) suggests estimating θ^* via

$$\hat{\theta}_{n,K} = K_n(G)^{-1}[\mu_n\{F(G+PG)\} - \mu_n(F)\,\mu_n(G+PG)], \tag{25}$$

where the empirical $k \times k$ matrix $K_n(G)$ is defined by

$$(K_n(G))_{ij} = \frac{1}{n-1}\sum_{t=1}^{n-1}\{G_i(X_t) - PG_i(X_{t-1})\}\{G_j(X_t) - PG_j(X_{t-1})\}. \tag{26}$$

The resulting estimator $\mu_n(F_{\hat{\theta}_{n,K}})$ for $\pi(F)$ based on the vector of control variates $U = G - PG$ and the estimated coefficients $\hat{\theta}_{n,K}$ is defined as

$$\mu_{n,K}(F) := \mu_n(F_{\hat{\theta}_{n,K}}) = \mu_n(F) - \langle \hat{\theta}_{n,K}, \mu_n(U)\rangle. \tag{27}$$

This will be the main estimator that is used in the remainder of the paper.

4. Choice of basis functions and the basic methodology

Let $\{X_n\}$ be an ergodic, reversible Markov chain with invariant measure π, and let F be a function whose mean $\pi(F)$ is to be estimated on the basis of samples from the chain. In the previous section we showed that, for *any* vector of control variates $U = G - PG$, the variance of the modified estimates $\mu_n(F_\theta) = \mu_n(F) - \langle \theta, U\rangle$ will be smaller than that of the simple ergodic averages $\mu_n(F)$. This is always so, except for the degenerate case where all the control variates U are perfectly uncorrelated with F, in the sense that the infinite series in equation (18) is identically zero.

In Table 1 we make a concrete proposal for the choice of the basis functions $\{G_j\}$ that are used to define the control variates $\{U_j\}$.

It is perhaps somewhat surprising that, unlike in Monte Carlo estimation based on independent identically distributed samples, for MCMC-based estimation the choice of control variates in Table 1 provides effective variance reduction in a wide range of MCMC scenarios stemming from Bayesian inference problems—primarily those where inference is performed via a conditionally conjugate random-scan Gibbs sampler. Examples of the application of this basic

Table 1. Outline of the basic methodology

(a) Given:
(i) a multivariate posterior distribution $\pi(x) = \pi(x^{(1)}, x^{(2)}, \ldots, x^{(d)})$
(ii) a reversible Markov chain $\{X_n\}$ with stationary distribution π
(iii) a sample of length n from the chain $\{X_n\}$
(b) Goal:
estimate the posterior mean $\mu^{(i)}$ of $x^{(i)}$
(c) Define:
(i) $F(x) = x^{(i)}$
(ii) basis functions as the co-ordinate functions $G_j(x) = x^{(j)}$ for all j for which $PG_j(x) := E[X^{(j)}_{n+1} \mid X_n = x]$ is computable in closed form
(iii) the corresponding control variates $U_j = G_j - PG_j$
(d) Estimate:
(i) the optimal coefficient vector θ^* by $\hat{\theta}_{n,K}$ as in equation (25)
(ii) the quantity of interest $\mu^{(i)}$ by the modified estimators $\mu_{n,K}(F)$ as in expression (27)

methodology are presented in Section 5. These examples give strong empirical evidence for the effectiveness of the methods proposed. In the remainder of this section we examine an idealized MCMC estimation scenario, and we show that, in that 'limiting' case, the basic methodology leads to estimators with *zero* asymptotic variance.

To apply the results developed so far, the MCMC sampler at hand needs to be reversible so that the estimator $\hat{\theta}_{n,K}$ in equation (25) for the optimal coefficient vector θ^* can be used, and also it is necessary that the one-step expectations $PG_j(x) = E[G_j(X_{n+1})|X_n = x]$ of the basis functions G_j should be explicitly computable in closed form. The most natural general family of MCMC algorithms that satisfy these two requirements is the collection of conditionally conjugate random-scan Gibbs samplers, with a target distribution π arising as the posterior density of the parameters in a Bayesian inference study. As for π, we focus on the case where it is multivariate normal. In addition to its mathematical tractability, this choice is also motivated by the fact that the Gaussian distribution is the distribution that is most commonly used as a general approximation of the target distribution π arising in Bayesian inference problems.

According to the discussion in Section 2.2, the main goal in choosing the basis functions $G = \{G_j\}$ is that it should be possible to approximate effectively the solution $\hat{F}$ of the Poisson equation for F as a linear combination of the G_j, i.e. $\hat{F} \approx \Sigma_{j=1}^{k} \theta_j G_j$. In the case of a random-scan Gibbs sampler with a Gaussian target density, the Poisson equation can be solved explicitly, and its solution is of a particularly simple form.

Theorem 1. Let $\{X_n\}$ denote the Markov chain constructed from the random-scan Gibbs sampler used to simulate from an arbitrary (non-degenerate) multivariate normal distribution $\pi \sim N(\mu, \Sigma)$ in $\mathbb{R}^k$. If the goal is to estimate the mean of the first component of π, then, letting $F(x) = x^{(1)}$ for $x = (x^{(1)}, x^{(2)}, \ldots, x^{(k)})^{\mathrm{T}} \in \mathbb{R}^k$, the solution $\hat{F}$ of the Poisson equation for F can be expressed as a linear combination of the co-ordinate basis functions $G_j(x) := x^{(j)}$, $x \in \mathbb{R}^k$, $1 \leqslant j \leqslant k$,

$$\hat{F} = \sum_{j=1}^{k} \theta_j G_j. \tag{28}$$

Moreover, writing $Q = \Sigma^{-1}$, the coefficient vector θ in equation (28) is given by the first row of the matrix $k(I - A)^{-1}$, where A has entries $A_{ij} = -Q_{ij}/Q_{ii}$, $1 \leqslant i \neq j \leqslant k$, $A_{ii} = 0$ for all i, and $I - A$ is always invertible.

Proof. Let H denote the candidate solution to the Poisson equation $H(x) = \Sigma_j \theta_j x^{(j)}$ and write $X = (X^{(1)}, X^{(2)}, \ldots, X^{(k)})$ for a random vector with distribution $\pi \sim N(\mu, \Sigma)$. Since π is non-degenerate, Σ is non-singular and so the precision matrix Q exists and its diagonal entries are non-zero. Since the conditional expectation of a component $X^{(j)}$ of X given the values of the remaining $X^{(-j)} := (X^{(1)}, \ldots, X^{(j-1)}, X^{(j+1)}, \ldots, X^{(k)})$ is $\mu^{(j)} + \Sigma_l A_{jl} (X^{(l)} - \mu^{(l)})$, we have

$$PH(x) = \sum_j \theta_j \left[\frac{k-1}{k} x^{(j)} + \frac{1}{k} \left\{ \mu^{(j)} + \sum_l A_{jl} (x^{(l)} - \mu^{(l)}) \right\} \right],$$

so that

$$\begin{aligned} PH(x) - H(x) &= -\frac{1}{k} \sum_j \theta_j \left\{ (x^{(j)} - \mu^{(j)}) - \sum_l A_{jl} (x^{(l)} - \mu^{(l)}) \right\} \\ &= -\frac{1}{k} \theta^{\mathrm{T}} (I - A)(x - \mu), \end{aligned}$$

where we have used the fact that $\Sigma_l A_{jl}(x^{(j)} - \mu^{(j)}) = \Sigma_{l \neq j} A_{jl}(x^{(j)} - \mu^{(j)})$, since the diagonal entries of A are all 0. For this to be equal to $-F(x) + \pi(F) = -(x^{(1)} - \mu^{(1)})$ for all x, it suffices to choose θ such that $\theta^{\mathrm{T}}(I - A) = (k, 0, \ldots, 0)$, as claimed. Finally, to see that $I - A$ is non-singular (and hence invertible), note that its determinant is equal to $(\Pi_j 1/Q_{jj}) \det(Q)$, which is non-zero since Σ is non-singular by assumption. □

In terms of estimation, theorem 1 states that, if samples from a multivariate Gaussian distribution are simulated via random-scan Gibbs sampling to estimate the mean of one of its components, then, using the linear basis functions $G_j(x) = x^{(j)}$ to construct a vector of control variates $U = G - PG$, the modified estimator $\mu_{n,K}(F)$ not only has smaller variance than the standard ergodic averages $\mu_n(F)$, but also its variance in the central limit theorem is in fact equal to *zero*.

Before examining the performance of this methodology in practice, we emphasize that it is applicable (and, as the examples presented below indicate, generally very effectively so) in a wide range of MCMC estimation scenarios, certainly not limited to approximately Gaussian target distributions and to the random-scan Gibbs sampler.

5. Markov chain Monte Carlo examples of the basic methodology

Here we present three examples of the application of the basic methodology that was outlined in the previous section. The examples below are chosen as representative cases covering a broad class of real applications of Bayesian inference.

In example 1, a bivariate normal density is simulated by random-scan Gibbs sampling. This setting is considered primarily as an illustration of the result of theorem 1 and, as expected, the variance of the modified estimators in this case is dramatically smaller. Example 2 considers a case of a fairly realistic Bayesian inference study via MCMC sampling, with a large hierarchical normal linear model, and it is found that the basic methodology of the previous section provides very significant variance reduction. Example 3 illustrates the use of the basic methodology in the case of Metropolis-within-Gibbs sampling from the posterior of a model where the use of heavy-tailed prior densities results in highly variable MCMC samples. (Such densities are commonly met in, for example, spatial statistics; see Dellaportas and Roberts (2003) for an illustrative example.) We find that the use of control variates is again quite effective in reducing the estimation variance. This example illustrates the point that, often, not all basis functions G_j can be easily used in the construction of control variates.

5.1. Example 1: 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 0 and the variance of X to be equal to 1. 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 random-scan 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 $\pi(x|y) \sim N\{(\rho/\tau)y, 1 - \rho^2\}$. 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 set $F(x, y) = x$ and define the basis functions $G_1(x, y) = x$ and $G_2(x, y) = y$. The corresponding functions PG_1 and PG_2 are easily computed as

$$PG_1(x, y) = \frac{1}{2}\left(x + \frac{\rho y}{\tau}\right)$$

and

Table 2. Estimated factors by which the variance of $\mu_n(F)$ is larger than the variance of $\mu_{n,K}(F)$, after $n = 1000, 10000, 50000, 100000, 200000, 500000$ simulation steps

Estimator	*Variance reduction factors for the following simulation steps:*					
	$n=1000$	$n=10000$	$n=50000$	$n=100000$	$n=200000$	$n=500000$
$\mu_{n,K}(F)$	4.13	27.91	122.4	262.5	445.0	1196.6

$$PG_2(x, y) = \tfrac{1}{2}(y + \rho\tau x).$$

The parameter values are chosen as $\rho = 0.99$ and $\tau^2 = 10$, so that the two components are highly correlated and the sampler converges slowly, making the variance of the standard estimates $\mu_n(F)$ large. Using the samples that are produced by the resulting chain with initial values $x_0 = y_0 = 0.5$, we examine the performance of the modified estimator $\mu_{n,K}(F)$ and compare it with the performance of the standard ergodic averages $\mu_n(F)$.

The factors by which the variance of $\mu_n(F)$ is larger than that of $\mu_{n,K}(F)$ are shown in Table 2. In view of theorem 1, it is not surprising that the estimator $\mu_{n,K}(F)$ is clearly much more effective than $\mu_n(F)$.

In this and in all subsequent examples, the reduction in the variance was computed from independent repetitions of the same experiment: here, for $\mu_n(F)$, $T = 200$ different estimates $\mu_n^{(i)}(F)$, for $i = 1, 2, \ldots, T$, were obtained, and the variance of $\mu_n(F)$ was estimated by

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

where $\bar{\mu}_n(F)$ is the average of the $\mu_n^{(i)}(F)$. The same procedure was used to estimate the variance of $\mu_{n,K}(F)$.

5.2. Example 2: hierarchical normal linear model

In an early application of MCMC methods in Bayesian statistics, Gelfand *et al.* (1990) illustrated the use of Gibbs sampling for inference in a large hierarchical normal linear model. The data consist of $N = 5$ weekly weight measurements of $l = 30$ young rats, whose weight is assumed to increase linearly in time, so that

$$Y_{ij} \sim N(\alpha_i + \beta_i x_{ij}, \sigma_c^2), \qquad 1 \leqslant i \leqslant l, \quad 1 \leqslant j \leqslant N,$$

where the Y_{ij} are the measured weights and the x_{ij} denote the corresponding rats' ages (in days). The population structure and the conjugate prior specification are assumed to be of the customary normal–Wishart–inverse gamma form: for $i = 1, 2, \ldots, l$,

$$\phi_i = \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim N(\mu_c, \Sigma_c),$$

$$\mu_c = \begin{pmatrix} \alpha_c \\ \beta_c \end{pmatrix} \sim N(\eta, C),$$

$$\Sigma_c^{-1} \sim \mathrm{W}\{(\rho R)^{-1}, \rho\},$$

$$\sigma_c^2 \sim \mathrm{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0\tau_0^2}{2}\right),$$

with known values for η, C, ν_0, ρ, R and τ_0.

The posterior π has $k := 2l + 2 + 3 + 1 = 66$ parameters, and MCMC samples from $((\phi_i), \mu_c, \Sigma_c, \sigma_c^2) \sim \pi$ can be generated via conditionally conjugate Gibbs sampling since the full conditional densities of all four parameter blocks (ϕ_i), μ_c, Σ_c and σ_c^2 are easily identified explicitly in terms of standard distributions; see Gelfand *et al.* (1990). For example, conditional on (ϕ_i), Σ_c, σ_c^2 and the observations (Y_{ij}), the means μ_c have a bivariate normal distribution with covariance matrix $V := (l\Sigma_c^{-1} + C^{-1})^{-1}$ and mean

$$V\left(\Sigma_c^{-1} \sum_i \phi_i + C^{-1}\eta\right). \tag{29}$$

Suppose, first, that we wish to estimate the posterior mean of α_c. We use a four-block, random-scan Gibbs sampler, which at each step selects one of the four blocks at random and replaces the current values of the parameter(s) in that block with a draw from the corresponding full conditional density. We set $F\{(\phi_i), \mu_c, \Sigma_c, \sigma_c^2\} = \alpha_c$ and construct control variates according to the basic methodology by first defining $k = 66$ basis functions G_j and then computing the one-step expectations PG_j. For example, numbering each G_j with the corresponding index in the order in which it appears above, we have $G_{61}\{(\phi_i), \mu_c, \Sigma_c, \sigma_c^2\} = \alpha_c$, and from expression (29) we obtain

$$PG_{61}\{(\phi_i), \mu_c, \Sigma_c, \sigma_c^2\} = \tfrac{3}{4}\alpha_c + \tfrac{1}{4}\left[(l\Sigma_c^{-1} + C^{-1})^{-1}\left\{\Sigma_c^{-1}\left(\sum_i \phi_i + C^{-1}\eta\right)\right\}\right]_1,$$

where the notation $[\cdots]_1$ indicates the first component of the vector $[\cdots]$.

Fig. 1 shows a typical realization of the sequence of estimates obtained by the standard estimators $\mu_n(F)$ and by the modified estimators $\mu_{n,K}(F)$, for $n = 50000$ simulation steps. The variance of $\mu_{n,K}(F)$ was found to be approximately a *30th* of that of $\mu_n(F)$. The second row of Table 3 shows the estimated variance reduction factors obtained at various stages of the MCMC simulation, based on $T = 100$ repetitions of the same experiment, performed as in example 1.

The initial values of the sampler were chosen as follows. For the (ϕ_i) we used the ordinary least squares estimates obtained from $l = 30$ independent regressions; their sample mean and covariance matrix provided starting values for μ_c and Σ_c respectively, and a pooled variance estimate of the individual regression errors provided the initial value of σ_c^2. The observed data (Y_{ij}) and known parameter values for η, C, ν_0, ρ, R and τ_0 are as in Gelfand *et al.* (1990).

Table 3. Estimated factors by which the variance of $\mu_n(F_j)$ is larger than the variance of $\mu_{n,K}(F)$, after $n = 1000, 10000, 20000, 50000, 100000, 200000$ simulation steps†

Parameter	*Variance reduction factors for the following simulation steps:*					
	$n = 1000$	$n = 10000$	$n = 20000$	$n = 50000$	$n = 100000$	$n = 200000$
(ϕ_i)	1.59–3.58	9.12–31.02	11.73–61.08	10.04–81.36	12.44–85.99	9.38–109.2
α_c	2.99	15.49	32.28	31.14	28.82	36.48
β_c	3.05	19.96	34.05	39.22	32.33	36.04
Σ_c	1.15–1.38	4.92–5.74	5.36–7.60	3.88–5.12	4.91–5.34	3.65–6.50
σ_c^2	2.01	5.06	5.23	5.17	4.75	5.79

†A different function F_j is defined for each of the $k = 66$ scalar parameters in the vector $((\phi_i), \mu_c, \Sigma_c, \sigma_c^2)$, and the same vector of control variates is used for all of them, as specified by the basic methodology of Section 4. In the first row, instead of individual variance reduction factors, we state the range of variance reduction factors obtained on the 60 individual parameters (ϕ_i), and similarly for the three parameters of Σ_c in the fourth row.

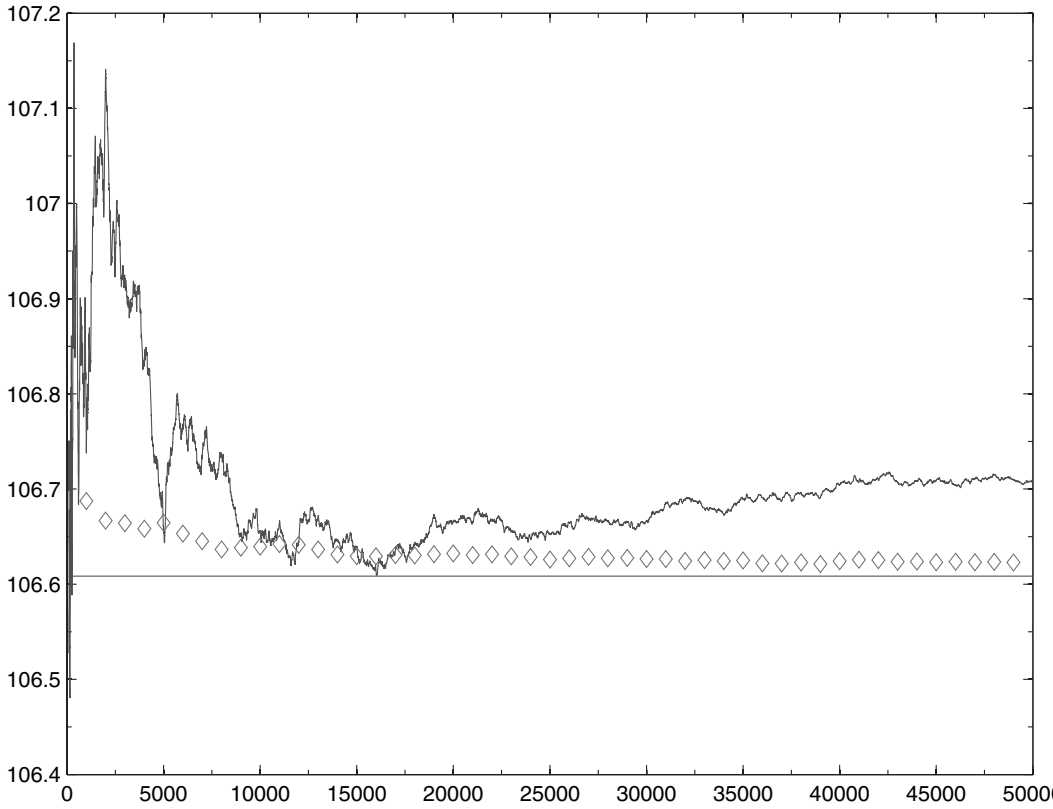

Fig. 1. Sequence of the standard ergodic averages (———) and the modified estimates (◇): for visual clarity, the values $\mu_{n,K}(F)$ are plotted only every 1000 simulation steps; the 'true' value of the posterior mean of α_C (———) was computed after $n = 10^7$ Gibbs steps and taken to be approximately 106.6084

More generally, in such a study we would be interested in the posterior mean of all the $k = 66$ model parameters. The same experiment as above was performed simultaneously for all the parameters. Specifically, 66 different functions F_j, $j = 1, 2, \ldots, 66$, were defined, one for each scalar component of the parameter vector $((\phi_i), \mu_c, \Sigma_c, \sigma_c^2)$, and the modified estimators $\mu_{n,K}(F_j)$ were used for each parameter, with respect to the same collection of control variates as before. The resulting variance reduction factors (again estimated from $T = 100$ repetitions) are shown in Table 3.

5.3. Example 3: Metropolis-within-Gibbs sampling

We consider an inference problem motivated by a simplified version of an example in Roberts and Rosenthal (2009). Suppose that N independent and identically distributed observations $y = (y_1, y_2, \ldots, y_N)$ are drawn from an $N(\phi, V)$ distribution, and place independent priors $\phi \sim$ Cauchy$(0, 1)$ and $V \sim$ IG$(1, 1)$ on the parameters ϕ and V respectively. The induced full conditionals of the posterior are easily seen to satisfy

$$\pi(\phi|V, y) \propto \left(\frac{1}{1+\phi^2}\right) \exp\left\{-\frac{1}{2V} \sum_i (\phi - y_i)^2\right\},$$

and

$$\pi(V|\phi, y) \sim \mathrm{IG}\Big\{1+\frac{N}{2}, 1+\frac{1}{2}\sum_i (\phi - y_i)^2\Big\}.$$

Since the distribution $\pi(\phi|V, y)$ is not of standard form, direct Gibbs sampling is not possible. Instead, we use a random-scan Metropolis-within-Gibbs sampler (see Müller (1993) and Tierney (1994)) and either update V from its full conditional density (Gibbs step) or update ϕ in a random-walk Metropolis step with a $\phi' \sim N(\phi, 1)$ proposal, each case chosen with probability $\frac{1}{2}$. Because 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 so in the simulation example that was considered, where the above algorithm is applied to a vector y of $N = 100$ independent and identically distributed $N(2, 4)$ 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 compare the performance of the standard empirical averages $\mu_n(F)$ with that of the modified estimators $\mu_{n,K}(F)$. As dictated by the basic methodology, we define $G_1(\phi, V) = \phi$ and $G_2(\phi, V) = V$, but we note that the one-step expectation $PG_1(\phi, V)$ cannot be obtained analytically because of the presence of the Metropolis step. Therefore, we use a single control variate $U = G - PG$ defined in terms of the basis function $G(\phi, V) = V$.

The resulting variance reduction factors, estimated from $T = 100$ repetitions of the same experiment, are *7.89*, *7.48*, *10.46* and *8.54*, after $n = 10000$, 50000, 100000, 200000 MCMC steps respectively.

6. Further methodological issues

6.1. Alternative consistent estimators for θ^*

Recall that the estimator $\hat{\theta}_{n,K}$ for the optimal coefficient vector $\theta^* = (\theta_1^*, \theta_2^*, \ldots, \theta_k^*)^{\mathrm{T}}$ that was defined in Section 3 was motivated by the new representation for θ^* derived in equation (24) of proposition 2. But we also derived an alternative expression for θ^* in equation (23), as $\theta^* = \Gamma(G)^{-1}\pi[\{F - \pi(F)\}(G + PG)]$, where $\Gamma(G)_{ij} = \pi\{G_iG_j - (PG_i)(PG_j)\}$, $1 \leqslant i, j \leqslant k$. This suggests that θ^* can alternatively be estimated via

$$\hat{\theta}_{n,\Gamma} := \Gamma_n(G)^{-1}[\mu_n\{F(G + PG)\} - \mu_n(F)\,\mu_n(G + PG)],$$

where the empirical $k \times k$ matrix $\Gamma_n(G)$ is defined by $(\Gamma_n(G))_{ij} = \mu_n\{G_iG_j - (PG_i)(PG_j)\}$, $1 \leqslant i, j \leqslant k$. Then $\hat{\theta}_{n,\Gamma}$ can in turn be used in conjunction with the vector of control variates $U = G - PG$ to estimate $\pi(F)$ via

$$\mu_{n,\Gamma}(F) := \mu_n(F_{\hat{\theta}_{n,\Gamma}}) = \mu_n(F) - \langle\hat{\theta}_{n,\Gamma}, \mu_n(U)\rangle = \frac{1}{n}\sum_{i=0}^{n-1}\{F(X_i) - \langle\hat{\theta}_{n,\Gamma}, U\rangle\}. \tag{30}$$

In theory, the estimators $\hat{\theta}_{n,\Gamma}$ and $\mu_{n,\Gamma}(F)$ enjoy the exact same asymptotic consistency and optimality properties as their earlier counterparts $\hat{\theta}_{n,K}$ and $\mu_{n,K}(F)$ respectively; these are established in Section 7. Also, the overall empirical performance of $\hat{\theta}_{n,\Gamma}$ and $\mu_{n,\Gamma}(F)$ was found in practice to be very similar to that of $\hat{\theta}_{n,K}$ and $\mu_{n,K}(F)$. This was observed in numerous simulation experiments that we conducted, some of which are reported in the unpublished notes of Dellaportas and Kontoyiannis (2009). A small difference between the two estimators was observed in experiments where the initial values of the sampler were quite far from the bulk of the mass of the target distribution π. There $\hat{\theta}_{n,\Gamma}$ sometimes appeared to converge faster than $\hat{\theta}_{n,K}$, and the corresponding estimator $\mu_{n,\Gamma}(F)$ often gave somewhat better results than

$\mu_{n,K}(F)$. The reason for this discrepancy is the existence of a time lag in the definition of $\hat{\theta}_{n,K}$: when the initial simulation phase produces samples that approach the area near the mode of π approximately monotonically, the entries of the matrix $K_n(G)$ accumulate a systematic one-sided error and consequently $K_n(G)$ takes longer to converge than $\Gamma_n(G)$. But this is a transient phenomenon that can be easily eliminated by including a burn-in phase in the simulation.

We systematically observed that the estimator $\hat{\theta}_{n,K}$ was more stable than $\hat{\theta}_{n,\Gamma}$, especially so in more complex MCMC scenarios involving a larger number of control variates. This difference was particularly pronounced in cases where one or more of the entries on the diagonal of $\Gamma(G) = K(G)$ were near 0. There, because of the inevitable fluctuations in the estimation process, the values of some of the entries of $\hat{\theta}_{n,\Gamma}$ fluctuated wildly between large negative and large positive values, whereas the corresponding entries of $\hat{\theta}_{n,K}$ were much more reliable since, by definition, $K(G)$ is positive semidefinite.

In conclusion, we found that, of the two estimators, $\mu_{n,K}(F)$ was consistently the more reliable and preferable choice.

We also briefly mention that a different method for consistently estimating θ^* was recently developed in Meyn (2007), based on the 'temporal difference learning' algorithm. Although this method also applies to non-reversible chains, it is computationally significantly more expensive than the estimates $\hat{\theta}_{n,K}$ and $\hat{\theta}_{n,\Gamma}$, and its applicability is restricted to discrete state space chains (or, more generally, to chains containing an atom). It may be possible to extend this idea to more general classes of chains by a simulated construction that is analogous to Nummelin's 'split chain' technique (see Nummelin (1984)), but we have not pursued this direction further.

6.2. Estimating θ^* via batch means

As noted in Section 2.2, the main difficulty in estimating the optimal coefficient vector θ^* in expression (14) was that it involves the solution $\hat{F}$ to the Poisson equation. Various researchers have observed in the past (see the references below) that one possible way to overcome this problem is to note that θ^* (like $\hat{F}$ itself) can alternatively be written in terms of an infinite series. Restricting attention, for simplicity, to the case of a single control variate $U = G - PG$ based on a single function $G : \mathbf{X} \to \mathbb{R}$, from equation (17) we have

$$\theta^* = \frac{1}{\sigma_U^2} \sum_{j=-\infty}^{\infty} E_\pi[F(X_0)\, U(X_j)] = \frac{1}{\pi\{G^2 - (PG)^2\}} \sum_{j=-\infty}^{\infty} E_\pi[F(X_0)\, U(X_j)]. \tag{31}$$

This suggests the following simple strategy: truncate the series in equation (31) to a finite sum, from $j = -M$ to $j = M$, say, and estimate an approximation to θ^* via

$$\tilde{\theta}_{n,M} = \frac{1}{\mu_n\{G^2 - (PG)^2\}} \sum_{j=-M}^{M} \frac{1}{n-2M} \sum_{i=M+1}^{n-M} F(X_i)\, U(X_{i+j}). \tag{32}$$

Then, $\tilde{\theta}_{n,M}$ converges almost surely to

$$\tilde{\theta}_M := \frac{1}{\sigma_U^2} \sum_{j=-M}^{M} E_\pi[F(X_0)\, U(X_j)], \tag{33}$$

as $n \to \infty$ (see corollary 2 in Section 7), and one would hope that $\tilde{\theta}_M \approx \theta^*$ for 'sufficiently large' M. Using the estimated coefficient $\tilde{\theta}_{n,M}$ in conjunction with the control variate $U = G - PG$, $\pi(F)$ can then be estimated by the corresponding modified averages

$$\tilde{\mu}_{n,M}(F) := \mu_n(F_{\tilde{\theta}_{n,M}}) = \mu_n(F) - \tilde{\theta}_{n,M}\, \mu_n(U). \tag{34}$$

This methodology has been used extensively in the literature, including, among others, by Andradòttir *et al.* (1993), Mira *et al.* (2003), Stein *et al.* (2004), Meyn (2007) and Hammer and Tjelmeland (2008). Our main point here is to show that it is strictly suboptimal, and in certain cases severely so. For this, we next give a precise expression for the amount by which the asymptotic variance of the batch means estimators $\tilde{\mu}_{n,M}(F)$ is *larger* than the (theoretically minimal) variance $\sigma^2_{\theta^*}$ of $\mu_{n,K}(F)$. Proposition 3 is proved at the end of this section.

Proposition 3. The sequences of estimators $\{\tilde{\mu}_{n,M}(F)\}$ and $\{\mu_{n,K}(F)\}$ are both asymptotically normal: as $n \to \infty$,

$$\{\tilde{\mu}_{n,M}(F) - \pi(F)\}\sqrt{n} \xrightarrow{\mathcal{D}} N(0, \tau^2_M),$$
$$\{\mu_{n,K}(F) - \pi(F)\}\sqrt{n} \xrightarrow{\mathcal{D}} N(0, \sigma^2_{\theta^*}),$$

where '$\to^{\mathcal{D}}$' denotes convergence in distribution. Moreover, the difference between the variance of the batch means estimators $\tilde{\mu}_{n,M}(F)$ and that of the modified estimators $\mu_{n,K}(F)$ is

$$\tau^2_M - \sigma^2_{\theta^*} = \frac{1}{\sigma^2_U}\left\{\sum_{|j|\geqslant M+1} E_\pi[F(X_0)\,U(X_j)]\right\}^2 \geqslant 0. \tag{35}$$

It is evident from equation (35) that the variance τ^2_M of the batch means estimators $\tilde{\mu}_{n,M}(F)$ will often be significantly larger than the minimal variance $\sigma^2_{\theta^*}$ that is achieved by $\mu_{n,K}(F)$, especially so if either

(a) the MCMC samples are highly correlated (as is often the case with samplers that tend to make small local moves), so that the terms of the series $\Sigma_j\, E_\pi[F(X_0)\,U(X_j)]$ decay slowly with $|j|$, or
(b) $|F|$ tends to take on large values; for example, note that the difference in expression (35) can be made arbitrarily large by multiplying F by a big constant.

But these are exactly the two most common situations that call for the use of a variance reduction technique such as control variates.

Indeed, in numerous simulation experiments (some simple cases of which are reported in the unpublished notes of Dellaportas and Kontoyiannis (2009)) we observed that, compared with $\mu_{n,K}(F)$, the batch means estimators $\tilde{\mu}_{n,M}(F)$ require significantly more computation (especially for large M) and they are typically much less effective. Also, we are unaware of any reasonably justified (non-*ad-hoc*) guidelines for the choice of the parameter M, which is critical for any potentially useful application of $\tilde{\mu}_{n,M}(F)$.

Using the obvious extension of the above construction to the case when more than a single control variate is used, we re-examined the three examples of the basic methodology that were presented in Section 5. Table 4 shows the results obtained in example 2 by the batch means estimators $\tilde{\mu}_{n,M}(F)$ for various choices of M, together with the earlier results obtained by $\mu_{n,K}(F)$. The sampling algorithm and all relevant parameters are as in example 2 in Section 5. For brevity, we display results only for the problem of estimating what is probably the statistically most significant parameter in this study, namely the mean slope β_c, which corresponds to taking $F\{(\phi_i), \mu_c, \Sigma_c, \sigma^2_c\} = \beta_c$.

6.2.1. Proof of proposition 3

The asymptotic normality statements are established in corollaries 1 and 2 of Section 7. To simplify the notation we decompose the infinite series in equation (31) into the sum $S_M + T_M$, where S_M is the sum of the terms corresponding to $-M \leqslant j \leqslant M$ and T_M is the double-sided tail

Table 4. Estimated factors by which the variance of $\mu_n(F)$ is larger than the variances of $\tilde{\mu}_{n,M}(F)$ and $\mu_{n,K}(F)$†

Estimator	*Variance reduction factors for example 2 for the following simulation steps:*			
	$n=1000$	$n=10000$	$n=50000$	$n=200000$
$\tilde{\mu}_{n,M}(F), M=0$	1.00	1.00	1.00	1.00
$\tilde{\mu}_{n,M}(F), M=1$	0.94	1.00	1.00	1.00
$\tilde{\mu}_{n,M}(F), M=5$	0.24	1.00	1.00	0.99
$\tilde{\mu}_{n,M}(F), M=10$	0.09	1.00	1.00	0.99
$\tilde{\mu}_{n,M}(F), M=20$	0.45	0.99	1.00	0.98
$\tilde{\mu}_{n,M}(F), M=100$	10^{-4}	0.84	0.95	0.96
$\mu_{n,K}(F)$	*3.05*	*19.96*	*39.22*	*36.04*

†All estimators are applied to MCMC data sampled from the posterior of the hierarchical linear model that was described in example 2, and the parameter being estimated is β_c so that here $F\{(\phi_i), \mu_c, \Sigma_c, \sigma_c^2\} = \beta_c$. Results are shown after $n = 1000, 10000, 50000, 200000$ simulation steps, with the batch means parameter $M = 0, 1, 5, 10, 20, 100$. The variance reduction factors are computed from $T = 100$ independent repetitions of the same experiment.

series corresponding to the same sum over all $|j| \geqslant M+1$. The variances of the two estimators are given by

$$\tau_M^2 = \sigma_F^2 + \tilde{\theta}_M^2 \sigma_U^2 - 2\tilde{\theta}_M(S_M + T_M),$$

$$\sigma_{\theta^*}^2 = \sigma_F^2 - \frac{1}{\sigma_U^2}(S_M + T_M)^2;$$

see equation (39) in corollary 2 and equation (18) respectively. Taking the difference between the two and substituting the value of $\tilde{\theta}_M = S_M/\sigma_U^2$ from expression (33) gives

$$\tau_M^2 - \sigma_{\theta^*}^2 = \frac{1}{\sigma_U^2}\{S_M^2 - 2S_M(S_M + T_M) + (S_M + T_M)^2\} = \frac{1}{\sigma_U^2} T_M^2,$$

as claimed in expression (35). □

In view of corollary 1 in Section 7, a simple examination of this proof shows that the result of proposition 3 also holds with the estimators $\mu_{n,\Gamma}(F)$ introduced in Section 6.1 in place of $\mu_{n,K}(F)$.

7. Theory

In this section we give precise conditions under which the asymptotics that were developed in Sections 2, 3 and 6.2 are rigorously justified. The results together with their detailed assumptions are stated below and the proofs are contained in Appendix A.

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 $\mathbf{X}$ equipped with a σ-algebra $\mathcal{B}$. The distribution of $\{X_n\}$ is described by its initial state $X_0 = x \in \mathbf{X}$ and its transition kernel $P(x, \mathrm{d}y)$ as in expression (5). The kernel P, as well as any of its powers P^n, acts linearly on functions $F : \mathbf{X} \to \mathbb{R}$ via $PF(x) = E[F(X_1)|X_0 = x]$.

The first assumption on the chain $\{X_n\}$ is that it is ψ *irreducible* and *aperiodic*. This means that there is a σ-finite measure ψ on $(\mathbf{X}, \mathcal{B})$ 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.

The second, and stronger, assumption, is an ergodicity condition (see Meyn and Tweedie (2009)): we assume that there are functions $V : \mathbf{X} \to [0, \infty)$ and $W : \mathbf{X} \to [1, \infty)$, a 'small' set $C \in \mathcal{B}$ and a finite constant $b > 0$ such that the Lyapunov drift condition holds:

$$PV - V \leqslant -W + b\mathbb{I}_C. \tag{36}$$

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

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

Under condition (36), we are assured that the chain is positive recurrent and that it has a unique invariant (probability) measure π. Our final assumption on the chain is that the Lyapunov function V in condition (36) satisfies $\pi(V^2) < \infty$.

These assumptions are summarized as follows.

> The chain $\{X_n\}$ is ψ irreducible and aperiodic, with unique invariant measure π, and there are functions $V : \mathbf{X} \to [0, \infty)$, $W : \mathbf{X} \to [1, \infty)$, a small set $C \in \mathcal{B}$ and a finite constant $b > 0$, such that condition (36) holds and $\pi(V^2) < \infty$. (37)

Although these conditions may seem somewhat involved, their verification is often straightforward; see Meyn and Tweedie (2009) and Robert and Casella (2004), as well as the numerous examples that were developed in Roberts and Tweedie (1996), Hobert and Geyer (1998), Jarner and Hansen (2000), Fort *et al.* (2003) and Roberts and Rosenthal (2004). It is often possible to avoid having to verify condition (36) directly, by appealing to the property of geometric ergodicity, which is essentially equivalent to the requirement that condition (36) holds with W being a multiple of the Lyapunov function V. For large classes of MCMC samplers, geometric ergodicity has been established in these references, 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 assumptions in result (37).

Apart from conditions on the Markov chain $\{X_n\}$, the asymptotic results that were stated earlier also require some assumptions on the function $F : \mathbf{X} \to \mathbb{R}$ whose mean under π is to be estimated, and on the (possibly vector-valued) function $G : \mathbf{X} \to \mathbb{R}^k$ which is used for the construction of the control variates $U = G - PG$. These assumptions are most conveniently stated within the weighted L_∞ framework of Meyn and Tweedie (2009). Given an arbitrary function $W : \mathbf{X} \to [1, \infty)$, the weighted L_∞ space L_∞^W is the Banach space

$$L_\infty^W := \left\{ \text{functions } F : \mathbf{X} \to \mathbb{R} \text{ subject to } \|F\|_W := \sup_{x \in \mathbf{X}} \left\{ \frac{|F(x)|}{W(x)} \right\} < \infty \right\}.$$

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

Theorem 2. Suppose that the chain $\{X_n\}$ satisfies conditions (36), and let $\{\theta_n\}$ be any sequence of random vectors in $\mathbb{R}^k$ such that θ_n converge to some constant $\theta \in \mathbb{R}^k$ almost surely as $n \to \infty$. Then:

(a) (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 $x \in \mathbf{X}$ and $A \in \mathcal{B}$,

$$P^n(x, A) \to \pi(A), \qquad \text{as } n \to \infty$$

(in fact, there is a finite constant B such that

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

uniformly over all initial states $x \in \mathbf{X}$ and all functions F such that $|F| \leqslant W$);

(b) (law of large numbers) 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 modified averages $\mu_n(F_{\theta_n})$ both converge to $\pi(F)$ almost surely as $n \to \infty$;

(c) (Poisson equation) if $F \in L_\infty^W$, then there is 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;

(d) (central limit theorem 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 non-zero, then the normalized ergodic averages $\{\mu_n(F) - \pi(F)\}\sqrt{n}$ converge in distribution to $N(0, \sigma_F^2)$, as $n \to \infty$;

(e) (central limit theorem 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 non-zero, then the normalized modified averages $\{\mu_n(F_{\theta_n}) - \pi(F)\}\sqrt{n}$ converge in distribution to $N(0, \sigma_{F_\theta}^2)$, as $n \to \infty$.

Suppose that the chain $\{X_n\}$ satisfies conditions (36) above, and that the functions F and $G = (G_1, G_2, \ldots, G_k)^{\mathrm{T}}$ are in L_∞^W. Theorem 2 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 coefficient vector $\theta = \theta^*$ which minimizes the limiting variance $\sigma_{F_\theta}^2$ of the modified averages, and the asymptotic behaviour of the estimators $\hat{\theta}_{n,\Gamma}$ and $\hat{\theta}_{n,\mathrm{K}}$ for θ^*.

As in Section 2.2, 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 2, there is a solution $\hat{F}$ to the Poisson equation for F. The simple computation that was outlined in Section 2.2 (and justified in the proof of theorem 3) leading to equation (14) shows that the variance $\sigma_{F_\theta}^2$ is minimized by the choice

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

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

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

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

$$(\Gamma_n(G))_{ij} = \mu_n(G_i G_j) - \mu_n\{(PG_i)(PG_j)\},$$

and

$$(K_n(G))_{ij} = \frac{1}{n-1}\sum_{t=1}^{n-1}\{G_i(X_t) - PG_i(X_{t-1})\}\{G_j(X_t) - PG_j(X_{t-1})\}.$$

Theorem 3. Suppose that the chain $\{X_n\}$ is reversible and satisfies conditions (36). If the functions F and G are both in L^W_∞ and the matrix $\Gamma(G)$ is non-singular, then both of the estimators for θ^* are almost surely consistent:

$$\hat{\theta}_{n,\Gamma} \to \theta^* \qquad \text{almost surely, as } n \to \infty;$$
$$\hat{\theta}_{n,K} \to \theta^* \qquad \text{almost surely, as } n \to \infty.$$

Recall the definitions of the two estimators $\mu_{n,\Gamma}(F)$ and $\mu_{n,K}(F)$ in equations (30) and (27), from Sections 3 and 6.1 respectively. 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 (36). If the functions F and G are both in L^W_∞ and the matrix $\Gamma(G)$ is non-singular, then the modified estimators $\mu_{n,\Gamma}(F)$ and $\mu_{n,K}(F)$ for $\pi(F)$ satisfy the following conditions:

(a) (law of large numbers) the modified estimators $\mu_{n,\Gamma}(F)$ and $\mu_{n,K}(F)$ both converge to $\pi(F)$ almost surely, as $n \to \infty$;
(b) (central limit theorem) if $\sigma^2_{\theta^*} := \pi\{\hat{F}^2_{\theta^*} - (P\hat{F}_{\theta^*})^2\}$ is non-zero, then the normalized modified averages $\{\mu_{n,\Gamma}(F) - \pi(F)\}\sqrt{n}$ and $\{\mu_{n,K}(F) - \pi(F)\}\sqrt{n}$ converge in distribution to $N(0, \sigma^2_{\theta^*})$, as $n \to \infty$, where the variance $\sigma^2_{\theta^*}$ is minimal among all estimators based on the control variate $U = G - PG$, in that $\sigma^2_{\theta^*} = \min_{\theta \in \mathbb{R}^k}(\sigma^2_\theta)$.

Finally we turn to the batch means estimators of Section 6.2. Recall the definitions of the estimators $\tilde{\theta}_{n,M}$ and $\tilde{\mu}_{n,M}(F)$ in equations (32) and (34) respectively. Our next result shows that $\tilde{\theta}_{n,M}$ converges to $\tilde{\theta}_M$ defined in expression (33) and gives an almost sure law of large numbers result and a corresponding central limit theorem for the estimators $\tilde{\mu}_{n,M}(F)$. Its proof follows along the same line as the proofs of the corresponding statements in theorem 3 and corollary 1. (In the one-dimensional setting of corollary 2 the assumption that $\Gamma(G)$ is non-singular reduces to assuming that $\sigma^2_U = \pi\{G^2 - (PG)^2\}$ is non-zero.)

Corollary 2. Under the assumptions of theorem 3, for any fixed $M \geqslant 0$, as $n \to \infty$ we have the following results:

(a) (law of large numbers) $\tilde{\theta}_{n,M} \to \tilde{\theta}_M$ almost surely and $\tilde{\mu}_{n,M}(F) \to \pi(F)$ almost surely;
(b) (central limit theorem) $\{\tilde{\mu}_{n,M}(F) - \pi(F)\}\sqrt{n} \to^{\mathcal{D}} N(0, \tau^2_M)$, where the variance τ^2_M is given by

$$\tau^2_M = \sigma^2_F + \tilde{\theta}^2_M \sigma^2_U - 2\tilde{\theta}_M \sum_{n=-\infty}^{\infty} \mathrm{cov}_\pi\{F(X_0), U(X_n)\}. \tag{39}$$

Some additional results on the long-term behaviour of estimators similar to those considered above can be found in Meyn (2006) and Meyn (2007), chapter 11, and finer asymptotics (including large deviations bounds and Edgeworth expansions) can be derived under stronger assumptions from the results in Kontoyiannis and Meyn (2003, 2005).

8. Concluding remarks

8.1. Applicability

One of the strengths of the present approach to the use of control variates in MCMC estimation is that, unlike in the classical case of independent sampling where control variates need to be identified in an *ad hoc* fashion for each specific application, this methodology is immediately applicable to a wide range of MCMC estimation problems. The most natural class of

such problems consists of all Bayesian inference studies where samples from the posterior are generated by a conditionally conjugate random-scan Gibbs sampler. Recall that conditionally conjugate Gibbs sampling is the key ingredient in, among others, Bayesian inference for dynamic linear models, e.g. Reis *et al.* (2006), applications of the slice sampler with auxiliary variables, e.g. Damien *et al.* (1999), Dirichlet processes, e.g. MacEachern and Muller (1998), and spatial regression models, e.g. Gamerman *et al.* (2003).

More generally, the present methodology applies to any MCMC setting satisfying the following two requirements: that the chain be reversible and that the conditional expectations $PG(x) = E[G(X_{n+1})|X_n = x]$ are explicitly computable for some simple functions G. There is a large collection of samplers with these properties, including certain versions of hybrid Metropolis-within-Gibbs algorithms (as in example 3), certain Metropolis–Hastings samplers on discrete states spaces and Markovian models of stochastic networks (as in Meyn (2007)). To ensure that these two requirements are satisfied, most of the experiments that were reported in Sections 5 and 6 were performed by using the *random-scan* version of Gibbs sampling. This choice is not *a priori* restrictive since the convergence properties of random-scan algorithms are generally comparable with (and sometimes superior to) those of systematic scan samplers; see, for example, Diaconis and Ram (2000) and Roberts and Sahu (1997).

We also observe that, as the present methodology is easily implemented as a post-processing algorithm and does not interfere in the actual sampling process, any implementation technique that facilitates or accelerates the MCMC convergence (such as blocking schemes, transformations and other reversible chains) can be used, as long as reversibility is maintained. Moreover, we note that the present work addresses mainly the problem of reducing the estimation error, in cases when the MCMC sampler is designed so that it explores the entire effective support of the target distribution π.

8.2. Further extensions

Probably the most interesting class of samplers to be considered next is that of general Metropolis–Hastings algorithms. When the target distribution is discrete or, more generally, when the proposal distribution is discrete and the number of possible moves is not prohibitively large, then the present methodology can be used as illustrated in example 6 of Dellaportas and Kontoyiannis (2010). But, in the case of general, typically continuous or multi-dimensional proposals, there is a basic obstacle: the presence of the accept–reject step makes it impossible to compute the required conditional expectation $PG(x)$ in closed form, for any G. If we consider the extended chain $\{(X_n, Y_n)\}$ that includes the values of the proposed moves Y_n (as done, for example, by Hammer and Tjelmeland (2008) and Delmas and Jourdain (2009)), then the computation of PG is straightforward for any $G(x, y)$ that depends only on x; but the chain $\{(X_n, Y_n)\}$ is no longer reversible and there are no clear candidates for good basis functions G. A possibly more promising point of view is to consider the computation of PG an issue of numerical integration, and to try to estimate the required values $PG(X_n)$ on the basis of importance sampling or any one of the numerous standard numerical integration techniques.

In a different direction, an interesting and potentially useful point would be to examine the effect of the use of control variates in the estimation *bias*. Although the variance of the standard ergodic averages $\mu_n(F)$ is a 'steady state' object, in that it characterizes their long-term behaviour and depends neither on the initial condition $X_0 = x$ nor on the transient behaviour of the chain, the bias depends heavily on the initial condition and it vanishes asymptotically. Preliminary computations (see the unpublished notes of Dellaportas and Kontoyiannis (2009)) indicate that the bias of $\mu_n(F)$ decays to 0 approximately like $\hat{F}(x)/n$, and that, using a single

control variate $U = G - PG$ based on a function $G \approx \hat{F}$ can significantly reduce the bias. It would be interesting in future work to compute the coefficient vector θ^b which minimizes the bias of $\mu_n(F_\theta)$ for a given collection of basis functions $\{G_j\}$, to study ways in which θ^b can be estimated empirically and to examine the effects that the use of θ^b in conjunction with the control variates $U = G - PG$ would have on the variance of the resulting estimator for $\pi(F)$.

A final point which may merit further attention is the potential problem of including too many control variates in the modified estimator $\mu_{n,K}(F)$. This issue has been studied extensively in the classical context of estimation based on independent identically distributed samples; see, for example, Lavenberg and Welch (1981), Law and Kelton (1982), Nelson (2004) and Glasserman (2004), pages 200–202. Since the optimal coefficient vector θ^* is not known *a priori*, using many control variates may in fact increase the variance of the modified estimators $\mu_{n,K}(F)$ relative to $\mu_n(F)$, and care must be taken to ensure that the most effective subset of all available control variates is chosen. Common sense suggests that the values of all the estimated parameters in the vector $\hat{\theta}_{n,K}$ should be examined, and the control variates corresponding to coefficients that are near 0 should be discarded. Since the MCMC output consists of simulated data from a known distribution, it may be possible to do this in a systematic fashion by using a classical hypothesis testing procedure.

Acknowledgements

We are grateful to Sean Meyn and Zoi Tsourti for many interesting conversations, and to Persi Diaconis and Christian Robert for insightful comments on an earlier version of this paper.

I. Kontoyiannis was supported in part by a Marie Curie International Outgoing Fellowship, PIOF-GA-2009-235837.

Appendix A: Proofs of theorems 2 and 3 and corollaries 1 and 2

A.1. Proof of theorem 2

Since any small set is petite (Meyn and Tweedie (2009), section 5.5.2), the f-norm ergodic theorem of Meyn and Tweedie (2009) implies that $\{X_n\}$ is positive recurrent with a unique invariant measure π such that condition (38) holds, and Meyn and Tweedie (2009), theorem 11.3.4, proves the Harris property, giving part (a).

From Meyn and Tweedie (2009), theorem 14.0.1, we have that, under conditions (36), $\pi(W) < \infty$. Since F is in L^W_∞, $\pi(|F|)$ is finite and, since $G \in L^W_\infty$, Jensen's inequality guarantees that $\pi(|U|)$ is finite. The invariance of π then implies that $\pi(U) = 0$; therefore, Meyn and Tweedie (2009), theorem 17.0.1, shows that $\mu_n(F) \to \pi(F)$ and $\mu_n(U) \to 0$ almost surely as $n \to \infty$ and, since $\theta_n \to \theta$ by assumption, $\mu_n(F_\theta)$ also converges to $\pi(F)$ almost surely, proving part (b).

The existence of a solution $\hat{F}$ to the Poisson equation in part (c) follows from Meyn and Tweedie (2009), theorem 17.4.2, and its uniqueness from Meyn and Tweedie (2009), theorem 17.4.1. The central limit theorem in part (d) is a consequence of Meyn and Tweedie (2009), theorem 17.4.4.

Finally, since $F, G \in L^W_\infty$, the functions U and F_θ are in L^W_∞ also, so $\hat{U}_j$ and $\hat{F}_\theta$ exist for each $j = 1, 2, \ldots, k$. As in part (d), the scaled averages $\{\mu_n(F_\theta) - \pi(F)\}\sqrt{n}$ and $\mu_n(U_j)\sqrt{n}$ converge in distribution to $N(0, \sigma^2_\theta)$ and $N(0, \sigma^2_{U_j})$ respectively, for each j, where the variances σ^2_θ and $\sigma^2_{U_j}$ are as in part (c). Writing $\theta = (\theta_1, \theta_2, \ldots, \theta_k)^{\mathrm{T}}$ and $\theta_n = (\theta_{n,1}, \theta_{n,2}, \ldots, \theta_{n,k})^{\mathrm{T}}$, we can express

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

Each of the terms in the second sum on the right-hand side above converges to 0 in probability, since $\mu_n(U_j)\sqrt{n}$ converges to a normal distribution and $\theta_{n,j} - \theta_j \to 0$ almost surely. Therefore, the sum converges to 0 in probability, and the central limit theorem in part (e) follows from part (d). □

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

A.2. Proof of theorem 3

We begin by justifying the computations in Sections 2.2 and 3. Define $\sigma^2_\theta = \pi\{\hat{F}^2_\theta - (P\hat{F}_\theta)^2\}$, where $\hat{F}$ exists by theorem 2. 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 σ^2_θ yields expression (13). To see that all the functions in expression (13) are indeed integrable recall that $\hat{F} \in L^{V+1}_\infty$ and note that, since V is non-negative, condition (36) implies that $1 \leqslant W \leqslant V + b\mathbb{1}_C$; hence $\pi(W^2)$ is finite since $\pi(V^2)$ is finite by assumption. Therefore, since $G \in L^W_\infty$, $\hat{F}$ 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 $\pi(P\hat{F}\langle\theta, PG\rangle) < \infty$. And, for the same reasons, all the functions appearing in the computations leading to the results of propositions 1 and 2 are also integrable.

The expression for the optimal θ^* in expression (14) is simply the solution for the minimum of the quadratic in expression (13). 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 that were established in theorem 2. First note that, since $G \in L^W_\infty$ 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 (2009), theorem 17.0.1, implies that $\Gamma_n(G) \to \Gamma(G)$ almost surely. Similarly, the functions F, G, PG, FG and FPG are all π integrable, so the same ergodic theorem implies that $\hat{\theta}_{n,\Gamma}$ indeed converges to θ^* almost surely, as $n \to \infty$.

To establish the corresponding result for $\hat{\theta}_{n,K}$, it suffices to show that $K_n(G) \to K(G)$ almost surely, and for that we consider the bivariate chain $Y_n = (X_n, X_{n+1})$ on the state space $\mathbf{X} \times \mathbf{X}$. Since $\{X_n\}$ is ψ irreducible and aperiodic, $\{Y_n\}$ is $\psi^{(2)}$ irreducible and aperiodic with respect to the bivariate measure $\psi^{(2)}(\mathrm{d}x, \mathrm{d}x') := \psi(\mathrm{d}x)P(x, \mathrm{d}y)$. Given functions W and V a small set C and a constant b so that condition (36) holds, it is immediate that condition (36) also holds for $\{Y_n\}$ with respect to the functions $V^{(2)}(x, x') = V(x')$ and $W^{(2)}(x, x') = W(x')$, the small set $\mathbf{X} \times C$ and the same b. The unique invariant measure of $\{Y_n\}$ is then $\pi^{(2)}(\mathrm{d}x, \mathrm{d}x') := \pi(\mathrm{d}x)P(x, \mathrm{d}y)$, and $\pi^{(2)}\{(V^{(2)})^2\}$ is finite. Therefore, assumptions (36) hold for $\{Y_n\}$ and for each pair $1 \leqslant i, j \leqslant k$ we can invoke the ergodic theorem Meyn and Tweedie (2009), 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, $K_n(G) \to K(G)$ almost surely.

A.3. Proof of corollary 1

The ergodic theorems in part (a) of corollary 1 are immediate consequences of theorem 2, part (b), combined with theorem 3. The computation in Section 2.2 which shows that θ^* in expression (14) indeed minimizes σ^2_θ (which is justified in the proof of theorem 3) shows that $\sigma^2_{\theta^*} = \min_{\theta \in \mathbb{R}^k}(\sigma^2_\theta)$. Finally, the assumption that $\Gamma(G)$ is non-singular combined with proposition 1 imply that all the variances $\sigma^2_{U_j}$ must be non-zero. Therefore, theorem 3 combined with the central limit theorems in parts (d) and (e) of theorem 2 prove part (b) of corollary 1.

A.4. Proof of corollary 2

The almost sure convergence statements in part (a) of corollary 2 follow the ergodic theorem, as in the proofs of theorems 2 and 3. The almost sure convergence of the denominator of expression (32), $\mu_n\{G^2 - (PG)^2\} \to \pi\{G^2 - (PG)^2\}$, is a special case (corresponding to $k = 1$) of the almost sure convergence of $\Gamma_n(G)$ to $\Gamma(G)$ that was proved in theorem 3. Considering the $(2M + 1)$-variate chain instead of the bivariate chain as in the proof of theorem 3, we can apply the ergodic theorem with the same integrability assumptions, to obtain that the sum in the numerator of expression (32) converges almost surely to $\Sigma_{|j| \leqslant M} E_\pi[F(X_0)\, U(X_j)]$, proving that $\tilde{\theta}_{n,M} \to \tilde{\theta}_M$ almost surely, as $n \to \infty$. For the modified averages, note that $\tilde{\mu}_{n,M}(F)$ is simply $\mu_n(F_{\tilde{\theta}_{n,M}})$. Then the law of large numbers and central limit theorem results for $\tilde{\mu}_{n,M}(F)$ follow from parts (b) and (e) of theorem 2 respectively. Finally, the limiting variance τ^2_M equals $\sigma^2_{\tilde{\theta}_M}$, which, using the representation in equation (16), can be expressed as claimed in expression (39).

References

Andradòttir, S., Heyman, D. and Ott, T. (1993) Variance reduction through smoothing and control variates for Markov Chain simulations. *ACM Trans. Modlng Comput. Simuln*, **3**, 167–189.

Assaraf, R. and Caffarel, M. (1999) Zero-variance principle for monte carlo algorithms. *Phys. Rev. Lett.*, **83**, 4682–4685.

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

Barone, P. and Frigessi, A. (1989) Improving stochastic relaxation for Gaussian random fields. *Probab. Engng Inform. Sci.*, **4**, 369–389.

Dalla Valle, L. and Leisen, F. (2010) A new multinomial model and a zero variance estimation. *Communs Statist. Simuln Computn*, **39**, 846–859.

Damien, P., Wakefield, J. and Walker, S. (1999) Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *J. R. Statist. Soc.* B, **61**, 331–344.

Dellaportas, P. and Kontoyiannis, I. (2009) Notes on using control variates for estimation with reversible MCMC samplers. Athens University of Economics and Business, Athens. (Available from `http://arxiv.org/abs/0907.4160`.)

Dellaportas, P. and Kontoyiannis, I. (2010) Control variates for reversible MCMC samplers. *Manuscript*. Athens University of Economics and Business, Athens. (Available from `http://arxiv.org/abs/1008.1355`.)

Dellaportas, P. and Roberts, G. (2003) An introduction to MCMC. In *Spatial Statistics and Computational Methods* (ed. J. Møller), pp. 1–42. New York: Springer.

Delmas, J.-F. and Jourdain, B. (2009) Does waste recycling really improve the multi-proposal Metropolis-Hastings algorithm?: an analysis based on control variates. *J. Appl. Probab.*, **46**, 938–959.

Diaconis, P. and Ram, A. (2000) Analysis of systematic scan Metropolis algorithms using Iwahori-Hecke algebra techniques. *Mich. Math. J.*, **48**, 157–190.

Fan, Y., Brooks, S. and Gelman, A. (2006) Output assessment for Monte Carlo simulations via the score statistic. *J. Computnl Graph. Statist.*, **15**, 178–206.

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

Gamerman, D., Moreirab, A. R. and Rue, H. (2003) Space-varying regression models: specifications and simulation. *Computnl Statist. Data Anal.*, **42**, 513–533.

Gelfand, A. E., Hills, S. E., Racine-Poon, A. and Smith, A. F. M. (1990) Illustration of Bayesian inference in normal data models using Gibbs sampling. *J. Am. Statist. Ass.*, **85**, 972–985.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.

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

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

Glynn, P. and Szechtman, R. (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 Han, X. (1992) Metropolis methods, Gaussian proposals, and antithetic variables. *Lect. Notes Statist.*, **74**, 142–164.

Hammer, H. and Tjelmeland, H. (2008) Control variates for the Metropolis-Hastings algorithm. *Scand. J. Statist.*, **35**, 400–414.

Henderson, S. (1997) Variance reduction via an approximating Markov process. *PhD Thesis*. Department of Operations Research, Stanford University, Stanford.

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

Henderson, S., Meyn, S. and Tadić, V. (2003) Performance evaluation and policy selection in multiclass networks. *Discr. Event Dyn. Syst.*, **13**, 149–189.

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

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

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

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

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

Kontoyiannis, I. and Meyn, S. (2005) Large deviation asymptotics and the spectral theory of multiplicatively regular Markov processes. *Electron. J. Probab.*, **10**, 61–123.

Lavenberg, S. and Welch, P. (1981) A perspective on the use of control variables to increase the efficiency of Monte Carlo simulations. *Mangmnt Sci.*, **27**, 322–335.

Law, A. and Kelton, W. (1982) *Simulation Modeling and Analysis*. New York: McGraw-Hill.
Liu, J. (2001) *Monte Carlo Strategies in Scientific Computing*. New York: Springer.
MacEachern, S. N. and Muller, P. (1998) Estimating mixture of Dirichlet process models. *J. Computnl Graph. Statist.*, **7**, 223–238.
Mengersen, K. L., Robert, C. P. and Guihenneuc-Jouyaux, C. (1999) MCMC convergence diagnostics: a review. In *Bayesian Statistics 6* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 415–440. Oxford: Oxford University Press.
Meyn, S. (2006) Large deviation asymptotics and control variates for simulating large functions. *Ann. Appl. Probab.*, **16**, 310–339.
Meyn, S. (2007) *Control Techniques for Complex Networks*. Cambridge: Cambridge University Press.
Meyn, S. P. and Tweedie, R. L. (2009) *Markov Chains and Stochastic Stability*, 2nd edn. London: Cambridge University Press.
Mira, A., Solgi, R. and Imparato, D. (2010) Zero variance Markov chain Monte Carlo for Bayesian estimators. *Technical Report*, `arXiv.org.1012.2983`.
Mira, A., Tenconi, P. and Bressanini, D. (2003) Variance reduction for MCMC. *Technical Report 2003/29*. Universitá degli Studi dell' Insubria, Insubria.
Müller, P. (1993) Alternatives to the Gibbs sampling scheme. *Technical Report*. Institute of Statistics and Decision Sciences, Duke University, Durham.
Nelson, B. (2004) Stochastic simulation research in management science. *Mangmnt Sci.*, **50**, 855–868.
Nummelin, E. (1984) *General Irreducible Markov Chains and Nonnegative Operators*. Cambridge: Cambridge University Press.
Philippe, A. and Robert, C. (2001) Riemann sums for MCMC estimation and convergence monitoring. *Statist. Comput.*, **11**, 103–115.
Reis, E. A., Salazar, E. and Gamerman, D. (2006) Comparison of sampling schemes for dynamic linear models. *Int. Statist. Rev.*, **74**, 203–214.
Robert, C. and Casella, G. (2004) *Monte Carlo Statistical Methods*, 2nd edn. New York: Springer.
Roberts, G. and Rosenthal, J. (1998) Markov-chain Monte Carlo: some practical implications of theoretical results (with discussion). *Can. J. Statist.*, **26**, 5–31.
Roberts, G. and Rosenthal, J. (2004) General state space Markov chains and MCMC algorithms. *Probab. Surv.*, **1**, 20–71.
Roberts, G. and Rosenthal, J. (2009) Examples of adaptive MCMC. *J. Computnl Graph. Statist.*, **18**, 349–367.
Roberts, G. O. and Sahu, S. K. (1997) Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *J. R. Statist. Soc.* B, **59**, 291–317.
Roberts, G. and Tweedie, R. (1996) Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, **83**, 95–110.
Stein, C., Diaconis, P., Holmes, S. and Reinert, G. (2004) Use of exchangeable pairs in the analysis of simulations. In *Stein's Method: Expository Lectures and Applications*, pp. 1–26. Beachwood: Institute of Mathematical Statistics.
Tierney, L. (1994) Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.*, **22**, 1701–1762.