Control Variates
for Reversible MCMC Samplers

Petros Dellaportas

Athens U of Economics and Business

Joint work with Ioannis Kontoyiannis (AUEB)

UCL

April 2010
Control variates in simple (i.i.d.) Monte Carlo

**Goal**: Compute the expected value of some function $F$ evaluated on i.i.d. samples $X_1, X_2, \ldots$

**Idea**: Variance of the standard ergodic averages $\frac{1}{n} \sum_{i=1}^{n} F(X_i)$ can be reduced by exploiting available zero-mean statistics

**Modified estimators**: If there is one or more functions $U_1, U_2, \ldots, U_k$ – the control variates – for which it is known that $E[U(X_i)] = 0$, then subtracting any linear combination

$$\frac{1}{n} \sum_{i=1}^{n} \left[ F(X_i) - \theta_1 U_1(X_i) - \theta_2 U_2(X_i) - \cdots - \theta_k U_k(X_i) \right]$$

does not change the asymptotic mean

**Practice**: For the optimal choice of $\{\theta_j\}$, the variance is no larger than before and often much smaller. The optimal $\{\theta_j^*\}$ are usually estimated adaptively, based on the same samples
Control Variates for Markov chains

- Extension of the above methodology to estimators based on MCMC samples is limited


Two fundamental difficulties:

⇝ \{U_j\}? hard to find (nontrivial and useful) functions with known expectation wrt the stationary distribution of the chain

⇝ \{\theta_j\}? even in cases where control variates are available, no effective way to obtain useful estimates for the optimal coefficients \{\theta_j^*\}

Reason: This is a fundamentally difficult problem, because the MCMC variance of ergodic averages is intrinsically an infinite-dimensional object. It cannot be written in closed form as a function of the transition kernel and the stationary distribution.
What we do

**Starting point:** Among others, Henderson (1997) 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.

**Choice of \( G \):** Our black-box method minimizes the variance for normal posteriors, we provide rules of thumb for any MCMC.

**Estimating \( \{\theta_j\} \):** We use control variates of this form conjunction with a new, efficiently implementable and provably optimal estimator for the coefficients \( \{\theta^*_j\} \) for reversible chains.

- Our estimator for \( \{\theta^*_j\} \) is adaptive, in the sense that is based on the same MCMC output.
- Unlike the case of independent sampling where control variates need to be found in an *ad hoc* manner depending on the specific problem at hand, *here the control variates (as well as the estimates of the corresponding optimal coefficients) come for free!*
The setting

- \( \{X_n\} \) is a discrete-time Markov chain with initial state \( X_0 = x \), and transition kernel \( P \):

\[
P(x, A) := \Pr\{X_{k+1} \in A \mid X_k = x\}, \quad \text{all } x, A
\]

**Typical application:** Construct an easy-to-simulate Markov chain \( \{X_n\} \) which has a target distribution \( \pi \) as its unique invariant measure.

**Ergodicity:** If we write \( PF(x) := E[F(X_1) \mid X_0 = x] \), then for appropriate \( F \)'s:

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

Moreover, \( \hat{F}(x) = \sum_{n=0}^{\infty} \left[ P^n F(x) - \pi(F) \right] \) where \( \hat{F} \) satisfies the **Poisson equation for** \( F \):

\[
P \hat{F} - \hat{F} = -F + \pi(F)
\]
The setting, continued

**Ergodic averages:** Estimate $\pi(F)$ by $\mu_n(F) := \frac{1}{n} \sum_{i=0}^{n-1} F(X_i)$

**Ergodic theorem:** $\mu_n(F) \to \pi(F)$, a.s., as $n \to \infty$, for appropriate $F$'s

**Central limit theorem:**

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

where $\sigma_F^2$, the asymptotic variance of $F$, is given by

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

**Asymptotic variance:** An alternative and more useful representation is in terms of the solution $\hat{F}$ to Poisson’s equation:

$$\sigma_F^2 = \pi\left(\hat{F}^2 - (P\hat{F})^2\right)$$
Construction of control variates for Markov chains

Suppose the chain \( \{X_n\} \) takes values in some space \( S \), typically \( S \subset \mathbb{R}^d \)

Construction of \( U \): Given any \( G : S \rightarrow \mathbb{R} \) with \( \pi(|G|) < \infty \), if we let

\[
U(x) := G(x) - PG(x) = G(x) - E[G(X_1)|X_0 = x]
\]

then \( \pi(U) := E_{\pi}[U(X)] = 0 \)

Modified Estimators: Given such a function \( U \) with \( \pi(U) = 0 \) and \( \theta \in \mathbb{R} \), define

\[
F_{\theta} = F - \theta U
\]

\[
\mu_n(F_{\theta}) = \mu_n(F) - \theta \mu_n(U)
\]

Goals: Search for particular choices for: (i) \( G \) (with corresponding \( U = G - PG \));

(ii) \( \theta \), so that the asymptotic variance \( \sigma^2_{F_{\theta}} \) of the modified estimators is significantly smaller than the variance \( \sigma^2_F \) of the standard ergodic averages \( \mu_n(F) \)
First suppose we have complete freedom in the choice of $G$. Set $\theta = 1$ without loss of generality.

We wish to make the asymptotic variance of 

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

as small as possible. But, in view of the Poisson equation

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

the choice $G = \hat{F}$ yields

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

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

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

With a choice $G$ that (we hope) approximates $\hat{F}$, we form the modified estimators $\mu_n(F_\theta)$ with respect to the function $F_\theta = F - \theta U = F - \theta G + \theta PG$

**Next task: Choose $\theta$:** Minimize the resulting variance

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

From the definitions, $\hat{U} = G$ and $\hat{F}_\theta = \hat{F} - \theta G$. Therefore,

$$\sigma_\theta^2 = \pi \left( (\hat{F} - \theta G)^2 \right) - \pi \left( (P\hat{F} - \theta PG)^2 \right)$$

Expanding the above quadratic in $\theta$, the optimal value is

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

Hard to estimate $\theta^*$ – it depends on $\hat{F}$
Interpretation of $\theta^*$

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

$$\sigma^2_\theta = \sigma^2_F + \theta^2 \sigma^2_U - 2\theta \sum_{n=-\infty}^{\infty} \text{Cov}_\pi (F(X_0), U(X_n)),$$

so that $\theta^*$ can also be expressed as

$$\theta^* = \frac{1}{\sigma^2_U} \sum_{n=-\infty}^{\infty} \text{Cov}_\pi (F(X_0), U(X_n))$$

leading to the optimal asymptotic variance

$$\sigma^2_{\theta^*} = \sigma^2_F - \frac{1}{\sigma^2_U} \left[ \sum_{n=-\infty}^{\infty} \text{Cov}_\pi (F(X_0), U(X_n)) \right]^2$$

This leads to our second rule of thumb for selecting control variates:

**Choose a control variate $U = G - PG$ so that $U$ and $F$ are highly correlated**
A different representation of \( \theta^* \)

\[
\theta^* = \frac{\pi(\hat{F}G - (P\hat{F})(PG))}{\pi(G^2 - (PG)^2)}.
\]

Since \( \hat{U} = G \), the denominator is simply \( \sigma^2_U \), and the fact that \( \sigma^2_U \) is always nonnegative suggests that there should be a way to rewrite the expression \( \pi(G^2 - (PG)^2) \) in the denominator of \( \theta^* \) in a way which makes this nonnegativity obvious. Indeed:

**Proposition.**

\[
\sigma^2_U = \pi(G^2 - (PG)^2) = E_\pi \left[ (G(X_1) - PG(X_0))^2 \right]
\]

and

\[
\theta^* = \frac{\pi(\hat{F}G - (P\hat{F})(PG))}{E_\pi \left[ (G(X_1) - PG(X_0))^2 \right]}.
\]
A suboptimal estimate for $\theta^*$

Returning momentarily to the case where $\{X_n\}$ are i.i.d. with distribution $\pi$, we have $\hat{F} = F$ and

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

which can be adaptively estimated by

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

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

$$
\mu_{n, \text{iid}}(F) := \mu_n(F_{\hat{\theta}_{n, \text{iid}}}) = \mu_n\left(F - \frac{\mu_n(FU)}{\mu_n(U^2)} U\right) = \mu_n(F) - \frac{\mu_n(U)\mu_n(FU)}{\mu_n(U^2)}
$$
Optimal empirical estimates

**Theorem.** If the chain \( \{X_n\} \) is *reversible*, then the optimal coefficient \( \theta^* \) for the control variate \( U = G - PG \) can be expressed as

\[
\theta^* = \theta_{\text{rev}}^* := \frac{\pi((F - \pi(F))(G + PG))}{E_\pi\left[\left(G(X_1) - PG(X_0)\right)^2\right]}
\]

Therefore, we can estimate:

\[
\theta^* \text{ as } \hat{\theta}_{n,\text{rev}} = \frac{\mu_n(F(G + PG)) - \mu_n(F)\mu_n(G + PG)}{\frac{1}{n} \sum_{i=0}^{n-1} (G(X_i) - PG(X_{i-1}))^2}
\]

\[
\pi(F) \text{ as } \mu_{n,\text{rev}}(F) := \mu_n(F_{\hat{\theta}_{n,\text{rev}}}) = \mu_n(F - \hat{\theta}_{n,\text{rev}} U)
\]

**Key:** Expressions do *not* involve the solution \( \hat{F} \) to Poisson’s equation

**Also:** Generalization to a multiple control variates \( U_j = G_j - PG_j \) is straightforward
Proof

Let $\Delta = P - I$ denote the generator of a discrete time Markov chain $\{X_n\}$ with transition kernel $P$.

Reversibility $\iff$ $\Delta$ is a self-adjoint linear operator on the space $L_2(\pi)$:

$$\pi(F \Delta G) = \pi(\Delta F G), \quad \text{for any two functions } F, G \in L_2(\pi)$$

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

$$\pi(\hat{F}G - (P\hat{F})(PG)) = \pi(\hat{F}G - (\hat{F} - \bar{F})(PG))$$

$$= \pi(\bar{F} PG - \hat{F} \Delta G)$$

$$= \pi(\bar{F} PG - \Delta \hat{F} G)$$

$$= \pi(\bar{F} PG + \bar{F} G)$$

$$= \pi(\bar{F}(G + PG))$$
Example: Gaussian-Gamma posterior

- $x = (x_1, x_2, \ldots, x_N) \sim \text{i.i.d. } N(\mu, \gamma^{-1})$

- Priors $\mu \sim N(0, 1)$ and $\gamma \sim \text{Gamma}(2, 1)$

- $N = 10, x = (-23, 27, 12, 17, -8, 2, -18, 17, 7, -33)$, so the sample mean is zero

- Simulate from the posterior on $(\mu, \gamma)$ using a random-scan Gibbs sampler

- We wish to estimate posterior mean of $\mu$ so we set $F(\mu, \gamma) = \mu$

- Compare the performance of the standard empirical averages $\mu_n(F)$ with the adaptive estimator $\mu_{n,\text{rev}}(F)$, based on the control variate $U = G - PG$ with $G(\mu, \gamma) = F(\mu, \gamma) = \mu$

- $\pi(F)$ is not computable in closed form, but the posterior marginal density of $\mu$ is proportional to the product of a Student’s $t$ density with mean zero (because the sample mean of $x$ is zero) and the prior $N(0, 1)$ density. Therefore, the resulting density is symmetric around zero, which implies that the posterior mean of $\mu$ is actually zero]
Example: Gaussian-Gamma posterior

Variance reduction factors

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Simulation steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 1000$</td>
<td>$n = 5000$</td>
</tr>
<tr>
<td>$n = 10000$</td>
<td>$n = 50000$</td>
</tr>
<tr>
<td>$\mu_{n,\text{rev}}(F)$</td>
<td>9403</td>
</tr>
<tr>
<td></td>
<td>341095</td>
</tr>
<tr>
<td></td>
<td>419766</td>
</tr>
<tr>
<td></td>
<td>20453186</td>
</tr>
</tbody>
</table>
Example: Discrete variables

Consider a bivariate density with a discrete variable $z$ and a continuous variable $p$, where

$$z|p \sim \text{Bern}(p) \text{ and } p \sim \text{Beta}(\alpha, \beta)$$

The random-scan Gibbs sampler draws randomly from either $z|p \sim \text{Bern}(p)$ or from

$$p|z \sim \text{Beta}(\alpha + z, \beta + 1 - z)$$

We wish to estimate the mean of $z$, so we set $F(z, p) = z$

Compare the performance of the standard ergodic averages $\mu_n(F)$ with the adaptive estimator $\mu_{n,\text{rev}}(F)$ based on the control variate $U = G - PG$ with $G(z, p) = z + p$

Random-scan Gibbs sampler, with $\alpha = 2$, $\beta = 1$, starting values $z_0 = 0$, $p_0 = 1/2$

[The true value of $\pi(F)$ is $\alpha/(\alpha + \beta) = 2/3$]
Example: Discrete variables

Variance reduction factors

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Simulation steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_{n,\text{rev}}(F)$</td>
<td>$n = 1000$</td>
</tr>
<tr>
<td></td>
<td>247.4</td>
</tr>
</tbody>
</table>
Example: Bivariate normal posterior

Let \((X, Y) \sim \pi(x, y)\) be a zero mean, bivariate Normal distribution, with \(\text{Var}(X) = 1, \text{Var}(Y) = \tau^2\), and \(E(XY) = \rho \tau\).

To estimate the expected value of \(X\) under \(\pi\) we sample from \(\pi\) using a random-scan Gibbs sampler and set \(F(x, y) = x\).

Use two control variates \(U_1, U_2\) defined in terms of \(G_1(x, y) = x\) and \(G_2(x, y) = y\).

Parameter values: \(\rho = 0.99, \tau^2 = 10\), with initial values \(x_0 = y_0 = 10\).

Also compare with estimates based on a single control variate \(U\) with \(G(x, y) = x + y\).
Example: Bivariate normal posterior

Variance reduction factors

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Simulation steps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n = 1000$</td>
</tr>
<tr>
<td>$\mu_n,rev(F)$</td>
<td>2.89</td>
</tr>
<tr>
<td>$\mu_n,rev,2(F)$</td>
<td>4.13</td>
</tr>
</tbody>
</table>
Theorem for Normal posteriors

**Theorem**  Let \( \{X_n\} \) denote the Markov chain constructed from the random-scan Gibbs sampler used to simulate from an arbitrary multivariate Normal distribution \( \pi \sim N(\mu, \Sigma) \) in \( \mathbb{R}^k \). If

\[
F(x) = x^{(1)} \quad \text{for each} \quad x = (x^{(1)}, x^{(2)}, \ldots, x^{(k)})^t \in \mathbb{R}^k,
\]

then the solution \( \hat{F} \) of the Poisson equation for \( F \) can be expressed as linear combination of the basis functions \( G_j(x) := x^{(j)}, \ x \in \mathbb{R}^k, \)

\[
\hat{F} = \sum_{j=1}^{k} \theta_j G_j.
\]

**Black-box G functions for any MCMC:** Use basis functions!
More complex examples

1. **Binary probit**

   *Data:* SAT data from Johnson & Albert (1999)

   *Model:* Diffuse prior on the parameters $\beta$, truncated Normal latent variables

   *Sampling from the posterior:* Random-scan Gibbs sampler

   *Control variates:* Vector of six $G$'s: $(\beta, \beta^2, (x^t x)^{-1} x^t Z)$

   *Variance reduction factors:* In estimating the posterior mean of $\beta_1$, around 50-80

2. **Two-threshold AR model:**

   *Data:* U.S. 3-month treasury bill rates 1962 – 1999

   *Model:* Self-exciting, two-threshold AR model, independent improper conjugate priors

   *Sampling:* 6-dim’al integration, and a discrete Metropolis-Hastings algorithm over $(c_1, c_2)$

   *Control variates:* Indicator functions of the three most likely models $(c_1, c_2)$

   *Variance reduction factors:* In estimating the posterior prob of MAP model, around 30-120
Example 3: A log-linear model

**Data:** 2 × 3 × 4 table of Knuiman and Speed (1988): 491 subjects classified according to
hypertension (yes, no), obesity (low, average, high) and alcohol consumption (0, 1-2, 3-5, or
6+ drinks/day)

**“Best” (main effects) model:** \( y_i \sim \text{Poisson}(\mu_i), \quad \log(\mu_i) = x_i^T \beta, \quad i = 1, 2 \ldots, 24 \)

**Prior:** Flat improper prior on \( \beta \in \mathbb{R}^7 \)

**Sampling:** Standard Bayesian inference via MCMC performed either by a Gibbs sampler (full
conditional densities are log-concave) or by a multivariate random walk Metropolis-Hastings
sampler
A log-linear model

**Sampling:** Here we use a simple random-scan Gibbs sampler, noting that a sample from the full conditional density of each $\beta_j$ can be obtained directly as the logarithm of a

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

**Estimation:** To estimate the posterior means of the $\beta_j$, set $F_j(\beta) = \beta_j$ for each $j = 1, 2, \ldots, 7$ and use the same seven control variates $U_1, U_2, \ldots, U_7$ for each $F_j$, where each $U_\ell = G_\ell - PG_\ell$ is defined in terms of $G_\ell(\beta) = \exp(\beta_\ell)$

**Computing PG:** The computation of $PG_\ell$ is straightforward, since the mean of $\exp(\beta_j)$ under the full conditional density of $\beta_j$ is

$$\frac{\sum_i y_i x_{ij}}{\sum_i x_{ij} = 1 \exp \left( \sum_{\ell \neq j} \beta_{\ell} x_{i\ell} \right)}$$
A log-linear model

The variance reduction factors obtained by our estimator $\mu_{n,\text{rev}}(F)$ for different parameters $\beta_j$ are in the range $3.55–5.57$, $38.2–57.69$, $66.20–135.51$, $57.16–170.34$ and $85.41–179.11$, after $n = 1000$, $10000$, $50000$, $100000$ and $200000$ simulation steps, respectively.
**Example 4: Gaussian mixtures**

- Still numerous unresolved issues in inference for finite mixtures. Such models are often ill-posed or non-identifiable. Difficulties reflect important problems in prior specifications and label switching.

- Improper priors are hard to use, and proper mixing over all (many!) posterior modes may require enforcing label-switching moves through Metropolis steps.

- We begin with \( N = 500 \) data points \( x = (x_1, x_2, \ldots, x_N) \) generated from the mixture

\[
\frac{7}{10} N(0, \frac{1}{4}) + \frac{3}{10} (0.1, 9)
\]

- Assume the means, variances and mixing proportions are all unknown. Usual conjugate prior setting with non-informative priors based on Richardson and Green (1997).

- Impose *a priori* restriction \( \mu_1 < \mu_2 \)

- To facilitate sampling from the posterior, introduce latent indicator variables \( Z_1, Z_2, \ldots, Z_N \)

- Problem: Estimate the two means \( \mu_1, \mu_2 \)
Gaussian mixtures: Sampling

- Standard random-scan Gibbs sampler that selects one of the four parameter blocks $(\mu_1, \mu_2)$, $(\sigma_1, \sigma_2)$, $Z$ or $p$, each with probability $1/4$

- Preferable to first obtain draws from the unconstrained posterior distribution and then to impose the identifiability (ordering) constraint at the post-processing stage

- The data $x$ have been generated so that the two means are very close, which results in frequent label switching throughout the MCMC run and in near-identical (unordered) marginal densities of $\mu_1$ and $\mu_2$

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

\[ F(\mu_1, \mu_2, \sigma_1, \sigma_2, Z, p) := \mu^o_1 = \min\{\mu_1, \mu_2\} \]

To reduce the variance of \( \mu_n(F) \) we use a bivariate control variate \( U = G - PG \), where \( G = (G_1, G_2) = (\mu^o_1, \sigma^o_1) \)

\( PG_1(\mu_1, \mu_2, \sigma_1, \sigma_2, Z, p) \) is the one-step expected value of \( \min\{\mu_1, \mu_2\} \)

\[
\frac{3}{4}\mu^o_1 + \frac{\nu_1}{4}\Phi\left(\frac{\nu_2 - \nu_1}{\sqrt{\tau^2_1 + \tau^2_2}}\right) + \frac{\nu_2}{4}\Phi\left(\frac{\nu_1 - \nu_2}{\sqrt{\tau^2_1 + \tau^2_2}}\right) - \frac{1}{4}\sqrt{\tau^2_1 + \tau^2_2}\Phi\left(\frac{\nu_2 - \nu_1}{\sqrt{\tau^2_1 + \tau^2_2}}\right)
\]

where \( \nu_j \) and \( \tau^2_j \) are the means and variances of \( \mu_j \), respectively, for \( j = 1, 2 \), under the corresponding full conditional densities.
Gaussian mixtures: \( PG_2 \)

First calculate the probability \( p(\text{order}) \) that \( \mu_1 < \mu_2 \):

\[
p(\text{order}) = \frac{\Phi\left( E(\mu_2 | \cdot) - E(\mu_1 | \cdot) \right)}{\sqrt{E(\sigma_1^2 | \cdot) + E(\sigma_2^2 | \cdot)}},
\]

where all four expectations above are taken under the corresponding full conditional densities, and, since the full conditional of each \( \sigma_j^{-2} \) is a Gamma density, the expectations of \( \sigma_1, \sigma_2, \sigma_1^2, \) and \( \sigma_2^2 \), are all available in closed form. Therefore, \( p(\text{order}) \) can be computed explicitly, and, \( PG_2 \) is:

\[
\frac{\sigma_1^2}{2} + \frac{1}{4} \left[ \mathbb{I}_{\{\mu_1 < \mu_2\}} E(\sigma_1 | \cdot) + \mathbb{I}_{\{\mu_1 > \mu_2\}} E(\sigma_2 | \cdot) \right] + \frac{1}{4} \left[ p(\text{order}) \sigma_1 + (1 - p(\text{order})) \sigma_2 \right]
\]

where all expectations are taken under the corresponding full conditional densities.
Gaussian mixtures: Variance reduction

With this choice for $G_1$, $G_2$ and corresponding control variates $U_1$, $U_2$, the variance reduction factors obtained by $\mu_{n,\text{rev}}(F)$ are 16.17, 25.36, 38.99, 44.5 and 36.16, after $n = 1000, 10000, 50000, 100000$ and 200000 simulation steps, respectively.
Discussion: Applicability

1. The methodology presented applies to any reversible MCMC sampler, as long as it is possible to compute the one-step expectation of some function $G$ of the parameters, in closed form.

2. These estimators can be used in a “black-box” fashion to various state-of-the-art samplers used in Bayesian inference via MCMC:

   ~→ all standard hierarchical models
   ~→ all random-walk Metropolis-Hastings samplers with a discrete proposal
   ~→ all conjugate Gibbs samplers
   ~→ many hybrid, Metropolis-within-Gibbs samplers
Theorem: “Under minimal assumptions, it all works”

Suppose \( \{X_n\} \) is \( \psi \)-irreducible, aperiodic, reversible and satisfies the Lyapunov drift condition (V3),

\[
P V \leq V - W + b \mathbb{I}_C.
\]

If \( F, G \in L^W_\infty \) and they are non-degenerate, then:

(i) **[Ergodicity]** The chain is positive Harris recurrent, it has a unique invariant measure \( \pi \), and it converges in distribution to \( \pi \) in a strong sense.

(ii) **[LLN]** The ergodic averages \( \mu_n(F) \), as well as the adaptive averages \( \mu_n,\text{rev}(F) \), both converge to \( \pi(F) \) a.s., as \( n \to \infty \).

(iii) **[Poisson Equation]** There is an essentially unique solution \( \hat{F} \in L^{V+1}_\infty \) to the Poisson eqn.

(iv) **[CLT for \( \mu_n(F) \)]** The normalized ergodic averages \( \sqrt{n}[\mu_n(F) - \pi(F)] \) converge in distribution to \( N(0, \sigma_F^2) \).

(v) **[CLT for \( \mu_{n,\text{rev}}(F) \)]** The normalized adaptive averages \( \sqrt{n}[\mu_{n,\text{rev}}(F) - \pi(F)] \) converge in distribution to \( N(0, \sigma_{F\theta*}^2) \), where the variance \( \sigma_{F\theta*}^2 \) is minimal among all estimators based on the control variate \( U = G - PG \).
Final remarks

- Bias and MSE issues

- Continuous Metropolis-Hastings algorithms are currently under investigation - here $PG$ is hard to be calculated analytically but simple Monte Carlo is very effective and $G$ can be any function. Extensive examples for simple random walk Metropolis algorithms show that variance reduction of a factor of 100 is readily achievable.