# Control variates and Rao-Blackwellization for deterministic sweep Markov chains

Stephen Berg[1], Jun Zhu[2], and Murray Clayton[2]

[1]Penn State University Statistics, [2]University of Wisconsin-Madison Statistics

## Summary


- A type of Rao-Blackwellization provably leads to variance reduction for deterministic sweep Gibbs sampling, for any number of components $K \geq 2$
- Further gains, theoretically and empirically, using a control variate approach
- For 2-component data augmentation Gibbs sampling, control variates are theoretically and empirically superior to the common Rao-Blackwellization approach of conditioning on the auxiliary random variable


## Setup

**Integral approximations:**

- $\pi$ is a probability measure on $(X, \mathscr{X})$
- want to compute expectations wrt $\pi$
- but not tractable: use MCMC

**Default approach:** to estimate

$$\mu = \int \pi(dx) g(x),$$

use Markov chain $X_0, X_1, X_2, \ldots$ and empirical average

$$\hat{\mu}_M^{emp} = M^{-1} \sum_{t=0}^{M-1} g(X_t)$$

## Asymptotics

Under mild conditions, we have

**CLT:** $M^{-1/2} \sum_{t=0}^{M-1} \{g(X_t) - \mu\} \xrightarrow{d} N(0, \underset{\substack{\uparrow \\ \text{Asymptotic variance}}}{\Sigma})$

**Goal:** variance reduction (reduce $\Sigma$ in Markov chain CLT)

## Deterministic sweep samplers

Cycle in fixed order through $K$ kernels

$$\Pi_k, \quad k = 1, \ldots, K$$

Example with $K = 3$:

$$X_0 \xrightarrow{\Pi_1(X_0,\cdot)} X_1 \xrightarrow{\Pi_2(X_1,\cdot)} X_2 \xrightarrow{\Pi_3(X_2,\cdot)} X_3 \xrightarrow{\Pi_1(X_3,\cdot)} X_4 \xrightarrow{\Pi_2(X_4,\cdot)} \ldots$$

**Typical use case**

- easier to find Markov kernel to update a component of a vector state $x \in X$ than to update the entire state at once
- commonly, Gibbs sampling or Metropolis-within-Gibbs

## Variance reduction via conditioning/Rao-Blackwellization

$$\hat{\mu}_M^{RB} = M^{-1} \sum_{t=0}^{M-1} \Pi_t g(X_t)$$

$\Pi_t$: transition kernel at step $t$

$$\Pi_t g(X_t) = \int \Pi_t(X_t, dx) g(x)$$

## Variance reduction via control variates

Use

$$M^{-1} \sum_{t=0}^{M-1} \{g(X_t) - W_t\}$$

where $W_t$ are mean 0 R.V.'s

We consider control variates with form

$$W_t = C^\top \{f(X_t) - \Pi_t f(X_t)\},$$

- $C \in \mathbb{R}^{p \times d}$ is a weight matrix
- $f : X \to \mathbb{R}^p$ is an arbitrary function

**Control variate estimator:**

$$\hat{\mu}_M^{CV} = \hat{\mu}_M^{emp} - \underbrace{M^{-1} C^\top \sum_{t=0}^{M-1} \{f(X_t) - \Pi_t f(X_t)\}}_{\text{control variate}}$$

- from $\pi$ stationarity of $\Pi_k$,

  $$E_\pi\{f(X) - \Pi_k f(X)\} = 0, \quad k = 1, \ldots, K$$

  so $f(x) - \Pi_k f(x)$ can be used as a control variate

## Control variate asymptotic variance

$$\text{(empirical)} \quad M^{1/2}(\hat{\mu}_M^{emp} - \mu) \xrightarrow{d} N(0, \Sigma^{emp})$$

$$\text{(control variate)} \quad M^{1/2}(\hat{\mu}_M^{CV} - \mu) \xrightarrow{d} N(0, \Sigma_C)$$

where

$$\Sigma_C = \Sigma^{emp} + C^\top U C - V^\top C - C^\top V$$

$$U = K^{-1} \sum_{k=1}^{K} \int \pi(dx) \{ff^T - (\Pi_k f)(\Pi_k f^T)\}$$

$$V = K^{-1} \sum_{k=1}^{K} \int \pi(dx) \{f \hat{g}_{\sigma(k)}^T - (\Pi_k f)(\Pi_k \hat{g}_{\sigma(k)})^T\}$$

**Optimal weight:** $\Sigma_C$ minimized at $\tilde{C} = U^{-1} V$

## Simplifications for Gibbs sampling:

For deterministic sweep Gibbs sampling, $V$ simplifies to

$$V = \int \pi(dx) f(x) \{g(x) - \mu\}^\top$$

- for Gibbs sampling, the optimal weight depends only on lag-0 and lag-1 autocovariances
  - these are easy to estimate based on the MCMC run

## Deterministic sweep Gibbs sampling result ($K \geq 2$)

**Asymptotic variance ordering**

$$\Sigma_{\tilde{C}} \leq \Sigma^{RB} \leq \Sigma^{emp}$$

where

- $\Sigma_{\tilde{C}}$ asymptotic variance of the control variate estimator $\hat{\mu}_M^{CV}$ with optimal weight $\tilde{C}$
- $\Sigma^{RB}$ asymptotic variance of the conditioning estimator $\hat{\mu}_M^{RB}$
- $\Sigma^{emp}$ asymptotic variance of the empirical average $\hat{\mu}^{emp}$

## Data augmentation Gibbs sampling

Setup:

- $K = 2$
- want $X \sim \pi$
- use $Z = (X, Y) \sim \tilde{\pi}$ where $\tilde{\pi}$ is a joint distribution with correct marginals:

  $$\tilde{\pi}\{(X, Y) \in A \times \Omega\} = \pi(X \in A)$$

- $X$ is variable of interest; function of interest $\tilde{g}(X, Y) = g(X)$ only depends on $X$
- $Y$ is auxiliary variable
- $Z = (X, Y)$ is augmented/joint state
- Gibbs kernels:

$$\Pi_1 h(z) = E_{\tilde{\pi}}\{h(Z)|Y\}$$
$$\Pi_2 h(z) = E_{\tilde{\pi}}\{h(Z)|X\}$$

**Another Rao-Blackwellization estimator:**

$$\hat{\mu}_M^{DA} = M^{-1} \sum_{t=0}^{M-1} \Pi_1 \tilde{g}(Z_t)$$

- $\hat{\mu}_M^{DA} \neq \hat{\mu}_M^{RB}$
  - since $\hat{\mu}_M^{DA}$ only averages conditional expectations wrt auxiliary variable $Y$

**Asymptotic variance comparison:**

$$\Sigma_{\tilde{C}} \leq \Sigma^{DA} \leq \Sigma^{RB} \leq \Sigma^{emp}$$

- Control variates outperform conditioning in this setting

## Simulation study

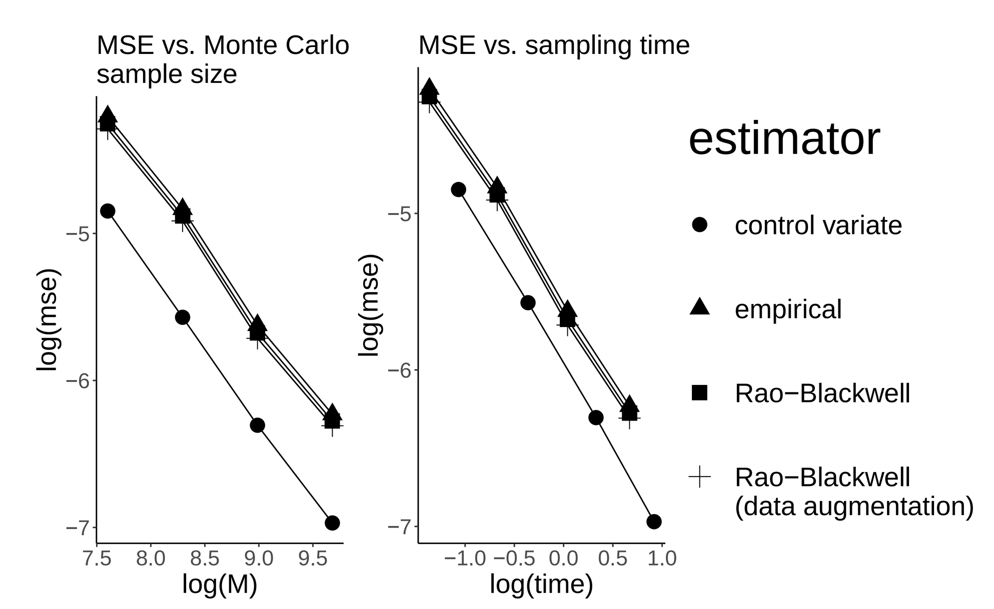


Figure: Mean squared error for estimating the posterior mean of $\beta$, versus number of Monte Carlo samples (left) and computing time (right).

**Bayesian probit regression example**

- Glass dataset from the UCI dataset repository
- $n = 214$ observations and $p = 10$ features (including an intercept column)
- predict Type $= 1$ vs. Type $\neq 1$ (originally 7 types)
- Normal prior $\beta \sim N(0, \tau^{-1} I_{p \times p})$
- Observations $Y_i | \beta \overset{ind}{\sim} Bernoulli(\Phi(x_i^\top \beta))$ (probit link)

**Sampling scheme:**

- DA Gibbs sampler of Albert and Chib [1993]
- Estimate posterior mean of $\beta$

## References

Poster based on Berg et al. [2019]

J. H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.*, 88, 1993.

S. Berg, J. Zhu, and M. K. Clayton. Control variates and Rao-Blackwellization for deterministic sweep Markov chains. *arXiv*, art. arXiv:1912.06926, Dec 2019.