# A Vanilla Rao-Blackwellization of Metropolis-Hastings Algorithms

Emilie Campos
Department of Biostatistics, UCLA

June 12, 2018

### Abstract

Duoc and Robert (2011) developed a Rao-Blackwellized solution for the accept-reject and Metropolis-Hastings algorithms that built off of the ideas of Casella and Robert (1996). Casella and Robert's solution reduced the variance of the estimators, but at a non-negligible compuation cost. This new solution, based on an independent representation, reduces the variance at a fixed computational cost.

*Keywords:* bayesian, importance sampling, variance reduction

# 1 Background

## 1.1 Gibbs Sampling Algorithm

The Gibbs sampling algorithm is sometimes called alternating conditional sampling [1]. This is an appropriate moniker as the algorithm starts from an appropriate guess and samples values from the full conditional densities in an alternating fashion. Suppose the parameter vector $\theta = (\theta_1, \theta_2, ..., \theta_d)$ with some joint target distribution. Let $\theta^{(0)} = \left(\theta_1^{(0)}, ..., \theta_d^{(0)}\right)$ be some intial values. For $s = 1, ..., S$ do:

- Sample $\theta_1^{(s)}$ from $\theta_1|\theta_2^{(s-1)}, ..., \theta_d^{(s-1)}$ (the full conditional of $\theta_1$ given all of the other $\theta_i$s)
- Sample $\theta_2^{(s)}$ from $\theta_2|\theta_1^{(s)}, \theta_3^{(s-1)}, ..., \theta_d^{(s-1)}$ (using updated $\theta_1$!)
- ...
- Sample $\theta_d^{(s)}$ from $\theta_d|\theta_1^{(s)}, ..., \theta_{d-1}^{(s)}$

then loop over again for $S$ iterations.

## 1.2 Metropolis Algorithm

The Metropolis algorithm is an adaptation of a random walk with an accept/reject rule [1].

- Starting point $\theta^{(0)}$ for which $p(\theta^{(0)}) > 0$ (a good guess), from a starting distribution $p_0(\theta)$
- For $t = 1, 2, ...$
  - Sample a proposal $\theta^*$ from "jumping"/"proposal" distribution $J_t(\theta^*|\theta^{(t-1)})$ where $J$ MUST be symmetric.
  - Calculate $r = \frac{p(\theta^*|y)}{p(\theta^{(t-1)}|y)}$
  - Set $\theta^{(t)} = \theta^*$ if a random uniform is less than $\min(r, 1)$ and $\theta^{(t-1)}$ otherwise.

## 1.3 Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm generalizes the Metropolis algorithm by allowing the proposal density to be asymmetric [1]. To correct from the asymmetry,

$$r = \frac{p(\theta^*|y)/J_t(\theta^*|\theta^{t-1})}{p(\theta^{t-1}|y)/J_t(\theta^{t-1}|\theta^*)}.$$

A good proposal distribution is one that for any $\theta$, it's easy to sample $J(\theta^*|\theta)$, it's easy to compute $r$ and each jump goes a reasonable distance (so it is not too slow to settle).

# 2 Introduction

The typical Metropolis-Hastings simulation algorithm relies on a generation of uniform random variables. The Rao-Blackwell Theorem states: If $g(X)$ is any kind of estimator of a parameter $\theta$, then the conditional expectation of $g(X)$ given $T(X)$, where $T(X)$ is a sufficient statistic, is typically a better estimator of $\theta$, and never worse. Thus, research has been done by several statstisticians to find a Rao-Blackwellization of the Monte-Carlo approximation:

$$\delta = \frac{1}{N}\sum_{t=1}^{N} h(x^{(t)}).$$

Casella and Robert found that the uniformity of the random variables added extra noise that does not provide information about the target density so they attempted to integrate out the uniforms condtional on all of the simulated proposal values. However, this strategy had a cost of $O(N^2)$. Duoc and Robert worked on a solution that "allows the variance to be reduced at a fixed computational cost" [2].

# 3 Rao-Blackwellization

## 3.1 Solution

The empirical average, using a sequence of accepted $y$'s from a Metropolis-Hastings experiment $(x^{(t)})_t$, is

$$\delta = \frac{1}{N}\sum_{t=1}^{N} h(x^{(t)}).$$

The alternative representation that Duoc and Robert will be using as a starting point is

$$\delta = \frac{1}{N}\sum_{i=1}^{M} n_i h(z_i),$$

where the $y_j$'s are the proposed Metropolis-Hastings moves, the $z_i$'s are the accepted $y_j$'s, $M$ is the number of accepted $y_j$'s up to time $N$ and $n_i$ is the number of times $z_i$ appears in the sequence $(x^{(t)})_t$.

**Lemma 1.** *The sequence $(z_i, n_i)$ is such that:*

1. *$(z_i, n_i)$ is a Markov chain;*

2. *$z_{i+1}$ and $n_i$ are independent given $z_i$;*

3. *$n_i$ is distributed as a geometric random variable with probability parameter*
$$p(z_i) := \int \alpha(z_i, y) q(y|z_i) dy; \tag{1}$$

4. *$(z_i)_i$ is a Markov chain with transition kernel $\tilde{Q}(z, dy) = \tilde{q}(y|z)dy$ and stationary distribution $\tilde{\pi}$ such that*

$$\tilde{q}(\cdot, z) \propto \alpha(z, \cdot) \text{ and } \tilde{\pi}(\cdot) \propto \pi(\cdot)p(\cdot).$$

We note the estimator is only dependent on the $z_i$'s so an optimal weight is the importance weight $1/p(z_i)$. However, this is typically not avaiable in closed form and needs to be estimated. The obvious solution is $n_i$, but there is a Rao-Blackwellized solution with smaller variance.

**Lemma 2.** *If $(y_j)_j$ is an i.i.d. sequence with distribution $q(y|z_i)$, then the quantity*

$$\hat{\xi}_i = 1 + \sum_{j=1}^{\infty} \prod_{\ell \le j} \{1 - \alpha(z_i, y_\ell\}$$

*is an unbiased estimator of $1/p(z_i)$, the variance of which, conditional on $z_i$, is lower than the conditional variance of $n_i$, $\{1 - p(z_i)\}/p^2(z_i)$.*

In this lemma, it is important to note that the sufficient statistic is the sequence $(y_j)_j$ and that $\xi$ can be rewritten as

$$\hat{\xi}_i = 1 + \sum_{j=1}^{\infty} E\left[\prod_{\ell \le j} \mathbb{I}\{u_\ell \ge \alpha(z_i, y_\ell)\}\middle|(y_t)_{t\ge 1}\right],$$

which makes it a Rao-Blackwellized solution.

It's possible for the estimator to be infinite since it depends on a ratio of probabilities. It would take an enormous number of iterations for this to happen and is not realistic. However, they created an esetimator that was based on a truncated series.

**Proposition 3.** *If $(y_j)_j$ is an i.i.d. sequence with distribution $q(y|z_i)$ and $(u_j)_j$ is an i.i.d. uniform sequence, for any $k \ge 0$, the quantity*

$$\hat{\xi}_i^k = 1 + \sum_{j=1}^{\infty} \prod_{1 \le \ell \le k \wedge j} \{1 - \alpha(z_i, y_j)\} \prod_{k+1 \le \ell \le j} \mathbb{I}\{u_\ell \ge \alpha(z_i, y_\ell)\} \tag{2}$$

*is an unbiased estimator of $1/p(z_i)$ with an almost sure finite number of terms. Moreover, for $k \ge 1$,*

$$\mathbb{V}[\hat{\xi}_i^k|z_i] = \frac{1 - p(z_i)}{p^2(z_i)} - \frac{1 - (1 - zp(z_i) + r(z_i))^k}{2p(z_i) - r(z_i)} \left(\frac{2 - p(z_i)}{p^2(z_i)}\right)(p(z_i) - r(z_i)),$$

*where $p$ is defined in (1) and $r(z_i) := \int \alpha^2(z_i, y) q(y|z_i) dy$. Therefore, we have*

$$\mathbb{V}[\hat{\xi}_i|z_i] \le \mathbb{V}[\hat{\xi}_i^k|z_i] \le \mathbb{V}[\hat{\xi}_i^0|z_i] = \mathbb{V}[n_i|z_i].$$

### 3.2 Convergence Properties

The Rao-Blackwellized solutions have important asymptotic improvements when estimating $E_\pi[h(X)]$. The following two theorems illustrate these properties.

**Theorem 4.** *Under the assumption that $\pi(p) > 0$, the following convergence properties hold:*

1. *if $h$ is in $C_\varphi$, then*
$$\delta_M^k \xrightarrow[M\to\infty]{\mathbb{P}} \pi(h);$$
2. *if, in addition, $h^2/p \in C_\varphi$ and $h \in C_\psi$, then*
$$\sqrt{M}(\delta_M^k - \pi(h)) \xrightarrow[M\to\infty]{\mathcal{L}} \mathcal{N}(0, V_k[h - \pi(h)]), \tag{3}$$
*where $V_k(h) := \pi(p) \int \pi(d_z) \mathbb{V}[\hat{\xi}_i^k|z] h^2(z) p(z) + \Gamma(h)$.*

This theorem implies that the correlation between $\xi$'s goes away eventually. This means for enough iterations, we have a pretty accurate estimation of the target density. The next theorem connects the original estimator with the previous theorem.

**Theorem 5.** *In addition to the assumptions of Theorem 4, assume that $h$ is a measurable function such that $h/p \in C_\zeta$ and $\{C_{h/p}, h^2/p^2\} \subset C_\phi$. Assume, moreover, that*

$$\sqrt{M}(\delta_M^0 - \pi(h)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V_0[h - \pi(h)]).$$

*Then, for any starting point $x$,*

$$\sqrt{M_N}\left(\frac{\sum_{t=1}^N h(x^{(t)})}{N} - \pi(h)\right) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}(0, V_0[h - \pi(h)]),$$

*where $M_N$ is defined by*

$$\sum_{i=1}^{M_N} \hat{\xi}_i^0 \leq N < \sum_{i=1}^{M_N+1} \hat{\xi}_i^0.$$

# 4 Examples

## 4.1 Standard Normal target, Normal Random Walk proposal

The first example Duoc and Robert use to illustrate the gains from the Rao-Blackwellized solution was a standard normal target density with proposal $q(y|x) = \varphi(x - y; \tau)$. The acceptance probability in this situation is the ratio of the targets. From Figure 1, we see that there is not a huge gain by using the Rao-Blackwellized estimator. This is possibly due to the randomness from both the $n_i$'s and $z_i$'s. Table 1 shows the ratio of the empirical variances of the terms $n_i h(z_i)$ and $\hat{\xi}_i h(z_i)$ for a few functions of $h$. We see that there is not much of an advantage using the $\hat{\xi}$ estimator when there is less variance and there is a huge gain when $\tau = 7$. The rejection rate is 82% when $\tau = 7$ because there is more variability in the original $n_i$'s. Table 2 then shows the additional time needed with the Rao-Blackwellized solution. The authors mention we should not look too deep into this table, as there are sometimes a few very lengths runs which can be bypassed by using the truncated estimator rather than the infinite one.

## 4.2 Standard Normal target, Cauchy proposal

This example used a standard normal density and a Cauchy $C(0, 0.25)$ proposal distribution. There was slightly superior improvement in this case using the Rao-Blackwellized solution. They point out that from Figure 2 illustrates this. We note that the shape is much steeper and comes in tighter around zero. The $\tau$'s are smaller than Table 1 in the first example and the rejection rates are much higher. Also, from Table 4, we see that there is a huge improvement from the first example in the amount of additional time needed.

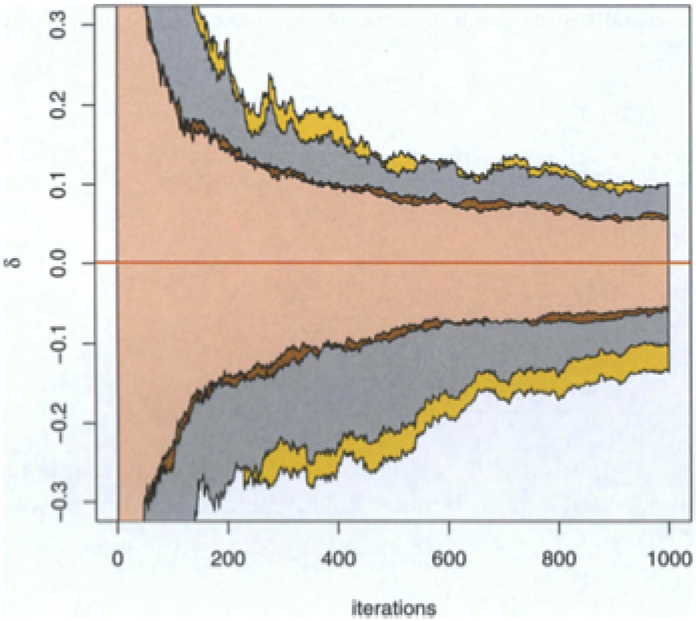

FIG. 1. *Overlay of the variations of* 250 *i.i.d. realizations of the estimates* $\delta$ (gold) *and* $\delta^{\infty}$ (grey) *of* $\mathbb{E}[X] = 0$ *for* 1000 *iterations, along with the* 90% *interquantile range for the estimates* $\delta$ (brown) *and* $\delta^{\infty}$ (pink), *in the setting of a random walk Gaussian proposal with scale* $\tau = 10$.

TABLE 1
*Ratios of the empirical variances of the components of the estimators* $\delta^{\infty}$ *and* $\delta$ *of* $\mathbb{E}[h(X)]$ *for* 100 *MCMC iterations over* $10^3$ *replications, in the setting of a random walk Gaussian proposal with scale* $\tau$, *when started with a normal simulation*

| $h(x)$ | $x$ | $x^2$ | $\mathbb{I}_{X>0}$ | $p(x)$ |
|---|---|---|---|---|
| $\tau = 0.1$ | 0.971 | 0.953 | 0.957 | 0.207 |
| $\tau = 2$ | 0.965 | 0.942 | 0.875 | 0.861 |
| $\tau = 5$ | 0.913 | 0.982 | 0.785 | 0.826 |
| $\tau = 7$ | 0.899 | 0.982 | 0.768 | 0.820 |

TABLE 2
*Evaluations of the additional computing effort due to the use of the Rao–Blackwell correction: median and mean numbers of additional iterations,* 80% *and* 90% *quantiles for the additional iterations, and ratio of the average R computing times obtained over* $10^5$ *simulations in the same setting as Table* 1

| | Median | Mean | $q_{0.8}$ | $q_{0.9}$ | Time |
|---|---|---|---|---|---|
| $\tau = 0.1$ | 1.0 | 6.49 | 5.0 | 11 | 2.33 |
| $\tau = 2$ | 0.0 | 7.06 | 4.3 | 11 | 6.5 |
| $\tau = 5$ | 0.0 | 9.02 | 4.6 | 13 | 8.4 |
| $\tau = 7$ | 0.0 | 9.47 | 4.8 | 13 | 3.5 |

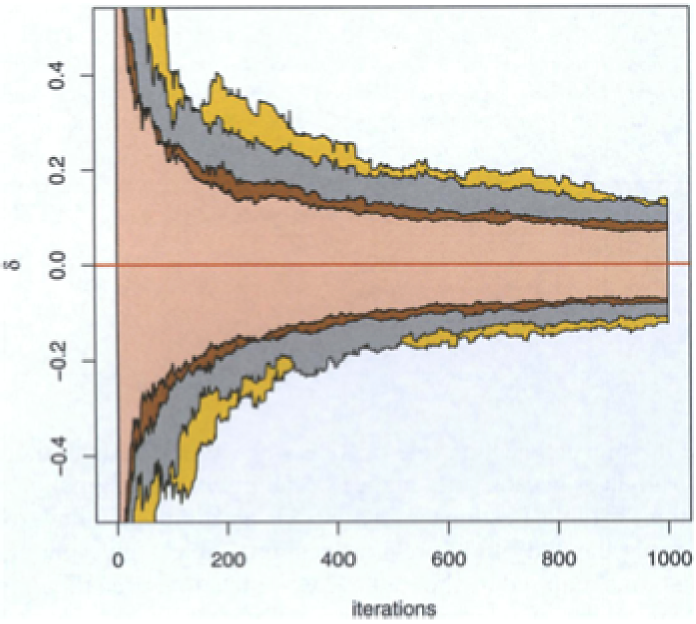

FIG. 2. *Overlay of the variations of 250 i.i.d. realizations of the estimates* $\delta$ (gold) *and* $\delta^{\infty}$ (grey) *of* $\mathbb{E}[X] = 0$ *for 1000 iterations, along with the 90% interquantile range for the estimates* $\delta$ (brown) *and* $\delta^{\infty}$ (pink), *in the setting of an independent Cauchy proposal with scale* 0.25.

TABLE 3

*Ratios of the empirical variances of the components of the estimators* $\delta^{\infty}$ *and* $\delta$ *of* $\mathbb{E}[h(X)]$ *for 100 MCMC iterations over* $10^3$ *replications, in the setting of an independent Cauchy proposal with scale* $\tau$ *started with a normal simulation*

| $h(x)$ | $x$ | $x^2$ | $\mathbb{I}_{X>0}$ | $p(x)$ |
|---|---|---|---|---|
| $\tau = 0.25$ | 0.677 | 0.630 | 0.663 | 0.599 |
| $\tau = 0.5$ | 0.790 | 0.773 | 0.716 | 0.603 |
| $\tau = 1$ | 0.937 | 0.945 | 0.889 | 0.835 |
| $\tau = 2$ | 0.781 | 0.771 | 0.694 | 0.591 |

TABLE 4

*Evaluations of the additional computing effort due to the use of the Rao–Blackwell correction: median and mean numbers of additional iterations, 80% and 90% quantiles for the additional iterations, and ratio of the average R computing times obtained over* $10^5$ *simulations in the same setting as Table* 3

| | Median | Mean | $q_{0.8}$ | $q_{0.9}$ | Time |
|---|---|---|---|---|---|
| $\tau = 0.25$ | 0.0 | 8.85 | 4.9 | 13 | 4.2 |
| $\tau = 0.50$ | 0.0 | 6.76 | 4 | 11 | 2.25 |
| $\tau = 1.0$ | 0.25 | 6.15 | 4 | 10 | 2.5 |
| $\tau = 2.0$ | 0.20 | 5.90 | 3.5 | 8.5 | 4.5 |

# References

[1] Andrew Gelman, J.B. Carlin, Hal S. Stern, and D.B Rubin. *Bayesian data analysis.* Chapman and Hall, Boca Raton, FL, third edition, 2004.

[2] Randal Douc and Christian P Robert. A Vanilla Rao-Blackwellization of Metropolis-Hastings Algorithms. *Source: The Annals of Statistics*, 39(1), 2011.