A Vanilla Rao-Blackwellization of Metropolis-Hastings Algorithms

Randal Duoc and Christian P. Robert

Emilie Campos

University of California, Los Angeles

June 12, 2018

Gibbs Sampling Algorithm

Suppose $\theta_1, ..., \theta_d$ have 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 θ_1 given all of the other θ_is)
- Sample $\theta_2^{(s)}$ from $\theta_2 | \theta_1^{(s)}, \theta_3^{(s-1)}, ..., \theta_d^{(s-1)}$ (using updated θ_1!)
- . . .
- Sample $\theta_d^{(s)}$ from $\theta_d | \theta_1^{(s)}, ..., \theta_{d-1}^{(s)}$

then loop over again for S iterations.

Metropolis

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

The Algorithm

- 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 θ^* 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.

Metropolis - Hastings

Generalizes the Metropolis algorithm, J_t need not be symmetric

- Correct for asymmetry with

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

- A good jumping distribution J_t is one that
  - for any θ, it's easy to sample $J(\theta^*|\theta)$
  - easy to compute r
  - each jump goes a reasonable distance (otherwise slow to settle)
  - does not reject jumps frequently

Rao-Blackwellization of Metropolis-Hastings

Theorem (Rao-Blackwell)

If $g(X)$ is any kind of estimator of a parameter θ, then the conditional expectation of $g(X)$ given $T(X)$, where $T(X)$ is a sufficient statistic, is typically a better estimator of θ, and never worse.

Casella and Robert (1996)

- Metropolis-Hastings relies on the generation of uniform variables, which are extraneous noise
- Casella and Robert [2] tried integrating out the random uniforms conditional on the simulated ys
- This had a nonnegligible cost of $O(N^2)$

Solution

Duoc and Robert reproduce the Rao-Blackwellization process from Casella and Robert but by means of an independent representation that allows the variance to be reduced at a fixed computational cost [3].

Solution

- The outcome of a Metropolis - Hastings experiment – $(x^{(t)})_t$, the accepted y's – is used in Monte Carlo approximation as

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

- Alternative estimator:

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

where

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

Solution

Lemma

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

Solution

Only the accepted y_j's are involved in the Metropolis-Hastings estimator δ so an optimal weight is $1/p(z_i)$, but this is typically not available in closed form and needs to be estimated. The estimator proposed earlier, n_i, is the obvious solution but others exist with smaller variance

Proposal

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

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, $\{1 - p(z_i)\}/p^2(z_i)$.

Solution

It's possible for $\hat{\xi}_i$ to be infinite since $\alpha(z_i, y_i)$ involves a ratio of probability densities and can therefore take on the value 1 with positive probability. However, this would take forever. Thus an intermediate estimator:

Proposal 2.0

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

is an unbiased estimator of $1/p(z_i)$ with an almost sure finite number of terms.

Solution

Moreover, for $k \geq 1$,

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

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

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

Convergence Properties

The estimator of $E_\pi[h(X)]$ is now for any $M > 0$,

$$\delta_M^k = \frac{\sum_{i=1}^M \hat{\xi}_i^k h(z_i)}{\sum_{i=1}^M \hat{\xi}_i^k}$$

For any positive function φ, denote C_φ as the set of functions bounded by φ up to a constant. Assume the reference important sampling estimator is sufficiently well behaved.

Convergence Properties

Theorem

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

1. *if h is in C_φ, 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 $V_k(h) := \pi(p) \int \pi(d_z)\mathbb{V}[\hat{\xi}_i^k|z]h^2(z)p(z) + \Gamma(h)$.

Asymptotically, the correlation between the ξ_i's vanishes

Convergence Properties

Theorem

In addition to the assumptions of the previous theorem, 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

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

Example - Random Walk Proposal

Target: $N(0,1)$
Proposal: $q(y|x) = \varphi(x-y;\tau)$

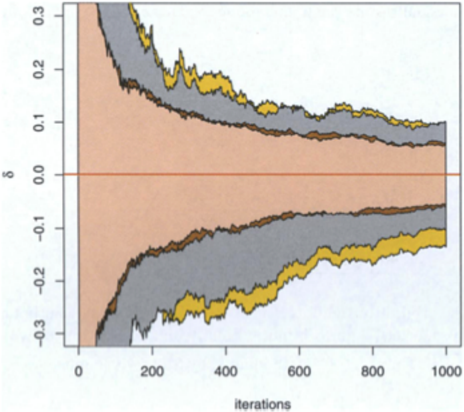

FIG. 1. *Overlay of the variations of 250 i.i.d. realizations of the estimates* δ (gold) *and* δ^{∞} (grey) *of* $\mathbb{E}[X]=0$ *for 1000 iterations, along with the 90% interquantile range for the estimates* δ (brown) *and* δ^{∞} (pink), *in the setting of a random walk Gaussian proposal with scale*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

Example - Random Walk Proposal

TABLE 1

Ratios of the empirical variances of the components of the estimators δ^{∞} and δ of $\mathbb{E}[h(X)]$ for 100 MCMC iterations over 10^3 replications, in the setting of a random walk Gaussian proposal with scale τ, 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

Example - Random Walk Proposal

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

Example - Cauchy proposal

Target: $N(0,1)$
Proposal: Cauchy $C(0, 0.25)$

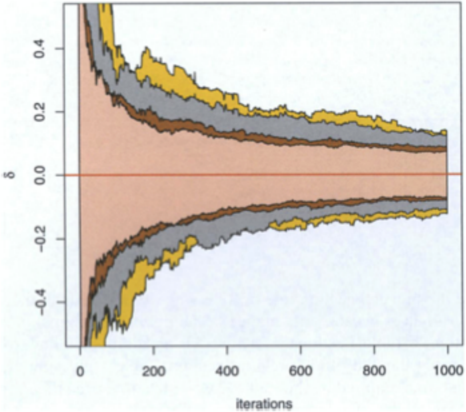

FIG. 2. *Overlay of the variations of 250 i.i.d. realizations of the estimates* δ (gold) *and* δ^{∞} (grey) *of* $\mathbb{E}[X] = 0$ *for 1000 iterations, along with the 90% interquantile range for the estimates* δ (brown) *and* δ^{∞} (pink), *in the setting of an independent Cauchy proposal with scale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

Example - Cauchy proposal

TABLE 3

Ratios of the empirical variances of the components of the estimators δ^{∞} and δ of $\mathbb{E}[h(X)]$ for 100 MCMC iterations over 10^3 replications, in the setting of an independent Cauchy proposal with scale τ 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

Example - Cauchy proposal

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] A. Gelman, J. Carlin, H. S. Stern, et al. *Bayesian data analysis*. Third. Boca Raton, FL: Chapman and Hall, 2004, p. 668. ISBN: 9781439840955. DOI: 10.1002/wcs.72. eprint: arXiv:1011.1669v3.

[2] G. Casella and C. P. Robert. "Rao-Blackwellisation of sampling schemes". In: *Biometrika* 83.1 (1996), pp. 81-94. ISSN: 0006-3444. DOI: 10.1093/biomet/83.1.81.

[3] R. Douc and C. P. Robert. "A Vanilla Rao-Blackwellization of Metropolis-Hastings Algorithms". In: *Source: The Annals of Statistics* 39.1 (2011). DOI: 10.1214/10-AOS838.