Rao-Blackwellisation for accelerating Metropolis-Hastings

Christian P. Robert

Université Paris-Dauphine, Paris & University of Warwick, Coventry

with M. Banterle, G. Casella, R. Douc, C. Grazian, & A. Lee

Outline

Given a density $f(\cdot)$ to simulate take $g(\cdot)$ density such that

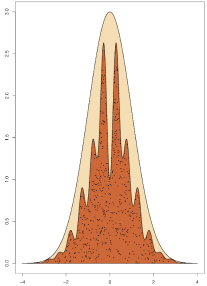
$$f(x) \leq Mg(x)$$

for $M \geq 1$
To simulate $X \sim f$, it is sufficient to generate

$$Y \sim g \quad U|Y = y \sim \mathcal{U}(0, Mg(y))$$

until

$$0 < u < f(y)$$



[Exercice 3.33, MCSM]

Raw outcome: id sequences $Y_1, Y_2, \ldots, Y_t \sim g$ and $U_1, U_2, \ldots, U_t \sim \mathcal{U}(0,1)$

Random number of accepted Y_i's

$$\mathbb{P}(N = n) = \binom{n-1}{t-1} \left({}^1/_M\right)^t \left(1 - {}^1/_M\right)^{n-t} ,$$

[Exercice 3.33, MCSM]
Raw outcome: id sequences $Y_1, Y_2, \ldots, Y_t \sim g$ and $U_1, U_2, \ldots, U_t \sim \mathcal{U}(0,1)$
Joint density of $(N, \mathbf{Y}, \mathbf{U})$

$$\begin{aligned}
&\mathbb{P}(N = n, Y_1 \le y_1, \ldots, Y_n \le y_n, U_1 \le u_1, \ldots, U_n \le u_n) \\
&= \int_{-\infty}^{y_n} g(t_n)(u_n \wedge w_n) dt_n \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_{n-1}} g(t_1) \ldots g(t_{n-1}) \\
&\qquad \times \sum_{(i_1, \cdots, i_{t-1})} \prod_{j=1}^{t-1} (w_{i_j} \wedge u_{i_j}) \prod_{j=t}^{n-1} (u_{i_j} - w_{i_j})^+ dt_1 \cdots dt_{n-1},
\end{aligned}$$

where $w_i = f(y_i)/Mg(y_i)$ and sum over all subsets of $\{1, \ldots, n-1\}$ of size $t-1$

[Exercice 3.33, MCSM]

Raw outcome: id sequences $Y_1, Y_2, \ldots, Y_t \sim g$ and $U_1, U_2, \ldots, U_t \sim \mathcal{U}(0,1)$

Marginal joint density of $(Y_i, U_i)|N = n$, $i < n$

$$
\begin{aligned}
&\mathbb{P}(N = n, Y_1 \le y, U_1 \le u_1) \\
&\quad = \binom{n-1}{t-1} \left(\frac{1}{M}\right)^{t-1} \left(1 - \frac{1}{M}\right)^{n-t-1} \\
&\quad \times \left[\frac{t-1}{n-1}(w_1 \wedge u_1)\left(1 - \frac{1}{M}\right) + \frac{n-t}{n-1}(u_1 - w_1)^+ \left(\frac{1}{M}\right)\right] \int_{-\infty}^{y} g(t_1)dt_1
\end{aligned}
$$

and marginal distribution of Y_i

$$
m(y) = {}^{t-1}/_{n-1} f(y) + {}^{n-t}/_{n-1} \frac{g(y) - \rho f(y)}{1 - \rho}
$$

$$
\mathbb{P}(U_1 \le w(y)|Y_1 = y, N = n) = \frac{g(y) w(y) M^{t-1}/_{n-1}}{m(y)}
$$

Accept-reject sample $(X_1, \ldots, X_m)$ associated with $(U_1, \ldots, U_N)$ and $(Y_1, \ldots, Y_N)$
N is stopping time for acceptance of m variables among Y_j's
Rewrite estimator of $\mathbb{E}[h]$ as

$$\frac{1}{m} \sum_{i=1}^{m} h(X_i) = \frac{1}{m} \sum_{j=1}^{N} h(Y_j) \, \mathbb{I}_{U_j \leq w_j} ,$$

with $w_j = {}^{f(Y_j)}/_{Mg(Y_j)}$

[Casella and Robert (1996)]

Rao-Blackwellisation: smaller variance produced by integrating out the U_i's,

$$\frac{1}{m}\sum_{j=1}^{N}\mathbb{E}[\mathbb{I}_{U_j\le w_j}|N,Y_1,\ldots,Y_N]\,h(Y_j)=\frac{1}{m}\sum_{i=1}^{N}\rho_i h(Y_i),$$

where $(i<n)$

$$\begin{aligned}\rho_i &= \mathbb{P}(U_i\le w_i|N=n,Y_1,\ldots,Y_n)\\ &= w_i\,\frac{\sum_{(i_1,\ldots,i_{m-2})}\prod_{j=1}^{m-2}w_{i_j}\prod_{j=m-1}^{n-2}(1-w_{i_j})}{\sum_{(i_1,\ldots,i_{m-1})}\prod_{j=1}^{m-1}w_{i_j}\prod_{j=m}^{n-1}(1-w_{i_j})},\end{aligned}$$

and $\rho_n=1$.
Numerator sum over all subsets of $\{1,\ldots,i-1,i+1,\ldots,n-1\}$ of size $m-2$, and denominator sum over all subsets of size $m-1$

[Casella and Robert (1996)]

Sample produced by Metropolis–Hastings algorithm

$$x^{(1)}, \ldots, x^{(T)}$$

based on two samples,

$$y_1, \ldots, y_T \quad \text{and} \quad u_1, \ldots, u_T$$

[Casella and Robert (1996)]

Sample produced by Metropolis–Hastings algorithm

$$x^{(1)}, \ldots, x^{(T)}$$

based on two samples,

$$y_1, \ldots, y_T \quad \text{and} \quad u_1, \ldots, u_T$$

Ergodic mean rewritten as

$$\delta^{MH} = \frac{1}{T} \sum_{t=1}^{T} h(x^{(t)}) = \frac{1}{T} \sum_{t=1}^{T} h(y_t) \sum_{i=t}^{T} \mathbb{I}_{x^{(i)}=y_t}$$

[Casella and Robert (1996)]

Sample produced by Metropolis–Hastings algorithm

$$x^{(1)}, \ldots, x^{(T)}$$

based on two samples,

$$y_1, \ldots, y_T \quad \text{and} \quad u_1, \ldots, u_T$$

Conditional expectation

$$\begin{aligned}
\delta^{RB} &= \frac{1}{T} \sum_{t=1}^{T} h(y_t)\, \mathbb{E}\left[\sum_{i=t}^{T} \mathbb{I} X^{(i)} = y_t \middle| y_1, \ldots, y_T \right] \\
&= \frac{1}{T} \sum_{t=1}^{T} h(y_t) \left(\sum_{i=t}^{T} \mathbb{P}(X^{(i)} = y_t | y_1, \ldots, y_T) \right)
\end{aligned}$$

with smaller variance

[Casella and Robert (1996)]

Take

$$\rho_{ij} = \frac{f(y_j)/q(y_j|y_i)}{f(y_i)/q(y_i|y_j)} \wedge 1 \qquad (j > i),$$

$$\overline{\rho}_{ij} = \rho_{ij} q(y_{j+1}|y_j), \quad \underline{\rho}_{ij} = (1-\rho_{ij}) q(y_{j+1}|y_i) \qquad (i < j < T),$$

$$\zeta_{jj} = 1, \quad \zeta_{jt} = \prod_{l=j+1}^{t} \underline{\rho}_{jl} \qquad (i < j < T),$$

$$\tau_0 = 1, \quad \tau_j = \sum_{t=0}^{j-1} \tau_t \zeta_{t(j-1)}\, \overline{\rho}_{tj}, \quad \tau_T = \sum_{t=0}^{T-1} \tau_t \zeta_{t(T-1)} \rho_{tT} \qquad (i < T),$$

$$\omega_T^i = 1, \quad \omega_i^j = \overline{\rho}_{ji} \omega_{i+1}^i + \underline{\rho}_{ji} \omega_{i+1}^j \qquad (0 \le j < i < T).$$

[Casella and Robert (1996)]

Theorem

The estimator δ^{RB} satisfies

$$\delta^{RB} = \frac{\sum_{i=0}^{T} \varphi_i \; h(y_i)}{\sum_{i=0}^{T-1} \tau_i \; \zeta_{i(T-1)}} \,,$$

with $(i < T)$

$$\varphi_i = \tau_i \left[\sum_{j=i}^{T-1} \zeta_{ij} \omega_{j+1}^{i} + \zeta_{i(T-1)} (1 - \rho_{iT}) \right]$$

and $\varphi_T = \tau_T$.

[Casella and Robert (1996)]

Alternative representation of Metropolis–Hastings estimator δ as

$$\delta = \frac{1}{n} \sum_{t=1}^{n} h(x^{(t)}) = \frac{1}{n} \sum_{i=1}^{M_n} \mathfrak{n}_i h(\mathfrak{z}_i)\,,$$

where

- $\mathfrak{z}_i$'s are the accepted y_j's,
- M_n is the number of accepted y_j's till time n,
- $\mathfrak{n}_i$ is the number of times $\mathfrak{z}_i$ appears in the sequence $(x^{(t)})_t$.

Define

$$\tilde{q}(\cdot|\mathfrak{z}_i) = \frac{\alpha(\mathfrak{z}_i,\cdot)\,q(\cdot|\mathfrak{z}_i)}{p(\mathfrak{z}_i)} \leq \frac{q(\cdot|\mathfrak{z}_i)}{p(\mathfrak{z}_i)}$$

where $p(\mathfrak{z}_i) = \int \alpha(\mathfrak{z}_i, y)\,q(y|\mathfrak{z}_i)dy$
To simulate from $\tilde{q}(\cdot|\mathfrak{z}_i)$

1. Propose a candidate $y \sim q(\cdot|\mathfrak{z}_i)$
2. Accept with probability

$$\tilde{q}(y|\mathfrak{z}_i) \Big/ \left(\frac{q(y|\mathfrak{z}_i)}{p(\mathfrak{z}_i)}\right) = \alpha(\mathfrak{z}_i, y)$$

   Otherwise, reject it and starts again.

- this is the transition of the HM algorithm

Define

$$\tilde{q}(\cdot|\mathfrak{z}_i) = \frac{\alpha(\mathfrak{z}_i, \cdot)\, q(\cdot|\mathfrak{z}_i)}{p(\mathfrak{z}_i)} \leq \frac{q(\cdot|\mathfrak{z}_i)}{p(\mathfrak{z}_i)}$$

where $p(\mathfrak{z}_i) = \int \alpha(\mathfrak{z}_i, y)\, q(y|\mathfrak{z}_i) dy$
The transition kernel $\tilde{q}$ admits $\tilde{\pi}$ as a stationary distribution:

$$\tilde{\pi}(x)\tilde{q}(y|x) = \underbrace{\frac{\pi(x)p(x)}{\int \pi(u)p(u)\mathrm{d}u}}_{\tilde{\pi}(x)} \underbrace{\frac{\alpha(x,y)q(y|x)}{p(x)}}_{\tilde{q}(y|x)}$$

Define

$$\tilde{q}(\cdot|\mathfrak{z}_i) = \frac{\alpha(\mathfrak{z}_i, \cdot)\, q(\cdot|\mathfrak{z}_i)}{p(\mathfrak{z}_i)} \leq \frac{q(\cdot|\mathfrak{z}_i)}{p(\mathfrak{z}_i)}$$

where $p(\mathfrak{z}_i) = \int \alpha(\mathfrak{z}_i, y)\, q(y|\mathfrak{z}_i) dy$
The transition kernel $\tilde{q}$ admits $\tilde{\pi}$ as a stationary distribution:

$$\tilde{\pi}(x)\tilde{q}(y|x) = \frac{\pi(x)\alpha(x,y)q(y|x)}{\int \pi(u)p(u)\mathrm{d}u}$$

Define

$$\tilde{q}(\cdot|\mathfrak{z}_i) = \frac{\alpha(\mathfrak{z}_i, \cdot)\, q(\cdot|\mathfrak{z}_i)}{p(\mathfrak{z}_i)} \leq \frac{q(\cdot|\mathfrak{z}_i)}{p(\mathfrak{z}_i)}$$

where $p(\mathfrak{z}_i) = \int \alpha(\mathfrak{z}_i, y)\, q(y|\mathfrak{z}_i) dy$
The transition kernel $\tilde{q}$ admits $\tilde{\pi}$ as a stationary distribution:

$$\tilde{\pi}(x)\tilde{q}(y|x) = \frac{\pi(y)\alpha(y, x)q(x|y)}{\int \pi(u)p(u)\mathrm{d}u}$$

Define

$$\tilde{q}(\cdot|\mathfrak{z}_i) = \frac{\alpha(\mathfrak{z}_i, \cdot)\, q(\cdot|\mathfrak{z}_i)}{p(\mathfrak{z}_i)} \leq \frac{q(\cdot|\mathfrak{z}_i)}{p(\mathfrak{z}_i)}$$

where $p(\mathfrak{z}_i) = \int \alpha(\mathfrak{z}_i, y)\, q(y|\mathfrak{z}_i) dy$
The transition kernel $\tilde{q}$ admits $\tilde{\pi}$ as a stationary distribution:

$$\tilde{\pi}(x)\tilde{q}(y|x) = \tilde{\pi}(y)\tilde{q}(x|y)\,,$$

Lemma (Douc & X., *AoS*, 2011)

The sequence $(\mathfrak{z}_i, \mathfrak{n}_i)$ *satisfies*

1. $(\mathfrak{z}_i, \mathfrak{n}_i)_i$ *is a Markov chain;*
2. $\mathfrak{z}_{i+1}$ *and* $\mathfrak{n}_i$ *are* independent *given* $\mathfrak{z}_i$*;*
3. $\mathfrak{n}_i$ *is distributed as a geometric random variable with probability parameter*

$$p(\mathfrak{z}_i) := \int \alpha(\mathfrak{z}_i, y)\, q(y|\mathfrak{z}_i)\, dy\,; \tag{1}$$

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

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

Lemma (Douc & X., *AoS*, 2011)

The sequence $(\mathfrak{z}_i, \mathfrak{n}_i)$ *satisfies*

1. $(\mathfrak{z}_i, \mathfrak{n}_i)_i$ *is a Markov chain;*
2. $\mathfrak{z}_{i+1}$ *and* $\mathfrak{n}_i$ *are* independent *given* $\mathfrak{z}_i$*;*
3. $\mathfrak{n}_i$ *is distributed as a geometric random variable with probability parameter*

$$p(\mathfrak{z}_i) := \int \alpha(\mathfrak{z}_i, y)\, q(y|\mathfrak{z}_i)\, dy\,; \tag{1}$$

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

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

Lemma (Douc & X., *AoS*, 2011)

The sequence $(\mathfrak{z}_i, \mathfrak{n}_i)$ *satisfies*

1. $(\mathfrak{z}_i, \mathfrak{n}_i)_i$ *is a Markov chain;*
2. $\mathfrak{z}_{i+1}$ *and* $\mathfrak{n}_i$ *are* independent *given* $\mathfrak{z}_i$*;*
3. $\mathfrak{n}_i$ *is distributed as a geometric random variable with probability parameter*

$$p(\mathfrak{z}_i) := \int \alpha(\mathfrak{z}_i, y)\, q(y|\mathfrak{z}_i)\, dy\,; \tag{1}$$

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

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

Lemma (Douc & X., *AoS*, 2011)

The sequence $(\mathfrak{z}_i, \mathfrak{n}_i)$ *satisfies*

1. $(\mathfrak{z}_i, \mathfrak{n}_i)_i$ *is a Markov chain;*
2. $\mathfrak{z}_{i+1}$ *and* $\mathfrak{n}_i$ *are* independent *given* $\mathfrak{z}_i$*;*
3. $\mathfrak{n}_i$ *is distributed as a geometric random variable with probability parameter*

$$p(\mathfrak{z}_i) := \int \alpha(\mathfrak{z}_i, y)\, q(y|\mathfrak{z}_i)\, dy\,; \tag{1}$$

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

$$\tilde{q}(\cdot|\mathfrak{z}) \propto \alpha(\mathfrak{z}, \cdot)\, q(\cdot|\mathfrak{z}) \quad \textit{and} \quad \tilde{\pi}(\cdot) \propto \pi(\cdot)p(\cdot)\,.$$

1. A natural idea:

$$\delta^* = \frac{1}{n} \sum_{i=1}^{M_n} \frac{h(\mathfrak{z}_i)}{p(\mathfrak{z}_i)} \,,$$

1. A natural idea:

$$\delta^* \simeq \frac{\sum_{i=1}^{M_n} \dfrac{h(\mathfrak{z}_i)}{p(\mathfrak{z}_i)}}{\sum_{i=1}^{M_n} \dfrac{1}{p(\mathfrak{z}_i)}} = \frac{\sum_{i=1}^{M_n} \dfrac{\pi(\mathfrak{z}_i)}{\tilde{\pi}(\mathfrak{z}_i)} h(\mathfrak{z}_i)}{\sum_{i=1}^{M_n} \dfrac{\pi(\mathfrak{z}_i)}{\tilde{\pi}(\mathfrak{z}_i)}} .$$

1. A natural idea:

$$\delta^* \simeq \frac{\sum_{i=1}^{M_n} \dfrac{h(\mathfrak{z}_i)}{p(\mathfrak{z}_i)}}{\sum_{i=1}^{M_n} \dfrac{1}{p(\mathfrak{z}_i)}} = \frac{\sum_{i=1}^{M_n} \dfrac{\pi(\mathfrak{z}_i)}{\tilde{\pi}(\mathfrak{z}_i)} h(\mathfrak{z}_i)}{\sum_{i=1}^{M_n} \dfrac{\pi(\mathfrak{z}_i)}{\tilde{\pi}(\mathfrak{z}_i)}} .$$

2. But p not available in closed form.

1. A natural idea:

$$\delta^* \simeq \frac{\sum_{i=1}^{M_n} \dfrac{h(\mathfrak{z}_i)}{p(\mathfrak{z}_i)}}{\sum_{i=1}^{M_n} \dfrac{1}{p(\mathfrak{z}_i)}} = \frac{\sum_{i=1}^{M_n} \dfrac{\pi(\mathfrak{z}_i)}{\tilde{\pi}(\mathfrak{z}_i)} h(\mathfrak{z}_i)}{\sum_{i=1}^{M_n} \dfrac{\pi(\mathfrak{z}_i)}{\tilde{\pi}(\mathfrak{z}_i)}} \,.$$

2. But p not available in closed form.
3. The geometric $\mathfrak{n}_i$ is the replacement, an obvious solution that is used in the original Metropolis–Hastings estimate since $\mathbb{E}[\mathfrak{n}_i] = 1/p(\mathfrak{z}_i)$.

The crude estimate of $1/p(\mathfrak{z}_i)$,

$$\mathfrak{n}_i = 1 + \sum_{j=1}^{\infty} \prod_{\ell \le j} \mathbb{I}\left\{u_\ell \ge \alpha(\mathfrak{z}_i, y_\ell)\right\} \,,$$

can be improved:

Lemma (Douc & X., *AoS*, 2011)

If $(y_j)_j$ is an iid sequence with distribution $q(y|\mathfrak{z}_i)$, the quantity

$$\hat{\xi}_i = 1 + \sum_{j=1}^{\infty} \prod_{\ell \le j} \left\{1 - \alpha(\mathfrak{z}_i, y_\ell)\right\}$$

is an unbiased estimator of $1/p(\mathfrak{z}_i)$ which variance, conditional on $\mathfrak{z}_i$, is lower than the conditional variance of $\mathfrak{n}_i$, $\{1 - p(\mathfrak{z}_i)\}/p^2(\mathfrak{z}_i)$.

Rao-Blackwellised, for sure?

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

1. Infinite sum but finite with at least positive probability:

$$\alpha(x^{(t)}, y_t) = \min\left\{1, \frac{\pi(y_t)}{\pi(x^{(t)})} \frac{q(x^{(t)}|y_t)}{q(y_t|x^{(t)})}\right\}$$

   For example: take a symmetric random walk as a proposal.

2. What if we wish to be sure that the sum is finite?

Finite horizon k version:

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

Rao-Blackwellised, for sure?

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

1. Infinite sum but finite with at least positive probability:

$$\alpha(x^{(t)}, y_t) = \min\left\{1, \frac{\pi(y_t)}{\pi(x^{(t)})} \frac{q(x^{(t)}|y_t)}{q(y_t|x^{(t)})}\right\}$$

   For example: take a symmetric random walk as a proposal.

2. What if we wish to be sure that the sum is finite?

Finite horizon k version:

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

Proposition (Douc & X., *AoS*, 2011)

If $(y_j)_j$ is an iid sequence with distribution $q(y|\mathfrak{z}_i)$ and $(u_j)_j$ is an iid uniform sequence, for any $k \geq 0$, the quantity

$$\hat{\xi}_i^k = 1 + \sum_{j=1}^{\infty} \prod_{1\leq\ell\leq k\wedge j} \{1 - \alpha(\mathfrak{z}_i, y_j)\} \prod_{k+1\leq\ell\leq j} \mathbb{I}\{u_\ell \geq \alpha(\mathfrak{z}_i, y_\ell)\}$$

is an unbiased estimator of $1/p(\mathfrak{z}_i)$ with an almost sure finite number of terms.

Proposition (Douc & X., *AoS*, 2011)

If $(y_j)_j$ is an iid sequence with distribution $q(y|\mathfrak{z}_i)$ and $(u_j)_j$ is an iid uniform sequence, for any $k \geq 0$, the quantity

$$\hat{\xi}_i^k = 1 + \sum_{j=1}^{\infty} \prod_{1 \leq \ell \leq k \wedge j} \{1 - \alpha(\mathfrak{z}_i, y_j)\} \prod_{k+1 \leq \ell \leq j} \mathbb{I}\{u_\ell \geq \alpha(\mathfrak{z}_i, y_\ell)\}$$

is an unbiased estimator of $1/p(\mathfrak{z}_i)$ with an almost sure finite number of terms. Moreover, for $k \geq 1$,

$$\mathbb{V}\hat{\xi}_i^k \mathfrak{z}_i = \frac{1 - p(\mathfrak{z}_i)}{p^2(\mathfrak{z}_i)} - \frac{1 - (1 - 2p(\mathfrak{z}_i) + r(\mathfrak{z}_i))^k}{2p(\mathfrak{z}_i) - r(\mathfrak{z}_i)} \left(\frac{2 - p(\mathfrak{z}_i)}{p^2(\mathfrak{z}_i)}\right) (p(\mathfrak{z}_i) - r(\mathfrak{z}_i)),$$

where $p(\mathfrak{z}_i) := \int \alpha(\mathfrak{z}_i, y)\, q(y|\mathfrak{z}_i)\, dy$. and $r(\mathfrak{z}_i) := \int \alpha^2(\mathfrak{z}_i, y)\, q(y|\mathfrak{z}_i)\, dy$.

Proposition (Douc & X., *AoS*, 2011)

If $(y_j)_j$ is an iid sequence with distribution $q(y|\mathfrak{z}_i)$ and $(u_j)_j$ is an iid uniform sequence, for any $k \geq 0$, the quantity

$$\hat{\xi}_i^k = 1 + \sum_{j=1}^{\infty} \prod_{1\leq\ell\leq k\wedge j} \{1-\alpha(\mathfrak{z}_i, y_j)\} \prod_{k+1\leq\ell\leq j} \mathbb{I}\{u_\ell \geq \alpha(\mathfrak{z}_i, y_\ell)\}$$

is an unbiased estimator of $1/p(\mathfrak{z}_i)$ with an almost sure finite number of terms. *Therefore, we have*

$$\mathbb{V}\hat{\xi}_i\mathfrak{z}_i \leq \mathbb{V}\hat{\xi}_i^k\mathfrak{z}_i \leq \mathbb{V}\hat{\xi}_i^0\mathfrak{z}_i = \mathbb{V}\mathfrak{n}_i\mathfrak{z}_i\,.$$


ICI
LE POSSIBLE
EST DEJA FAIT
L'IMPOSSIBLE
EST EN COURS
POUR LES
MIRACLES
PREVOIR 48H
DE DELAI

Non-informative inference for mixture models

Standard mixture of distributions model

$$\sum_{i=1}^{k} w_i \, f(x|\theta_i) \,, \quad \text{with} \quad \sum_{i=1}^{k} w_i = 1 \,. \tag{1}$$

[Titterington et al., 1985; Frühwirth-Schnatter (2006)]

Jeffreys' prior for mixture not available due to computational reasons : it has not been tested so far

[Jeffreys, 1939]

Warning: Jeffreys' prior improper in some settings

[Grazian & Robert, 2015]

Grazian & Robert (2015) consider genuine Jeffreys' prior for complete set of parameters in (1), deduced from Fisher's information matrix

Computation of prior density costly, relying on many integrals like

$$\int_{\mathcal{X}} \frac{\partial^2 \log \left[\sum_{i=1}^{k} w_i\, f(x|\theta_i)\right]}{\partial\theta_h \partial\theta_j} \left[\sum_{i=1}^{k} w_i\, f(x|\theta_i)\right] \mathrm{d}x$$

Integrals with no analytical expression, hence involving numerical or Monte Carlo (costly) integration

When building Metropolis-Hastings proposal over (w_i, θ_i)'s, prior ratio more expensive than likelihood and proposal ratios

Suggestion: split the acceptance rule

$$\alpha(x,y) := 1 \wedge r(x,y), \qquad r(x,y) := \frac{\pi(y|\mathcal{D})q(y,x)}{\pi(x|\mathcal{D})q(x,y)}$$

into

$$\tilde{\alpha}(x,y) := \left(1 \wedge \frac{f(\mathcal{D}|y)q(y,x)}{f(\mathcal{D}|x)q(x,y)}\right) \times \left(1 \wedge \frac{\pi(y)}{\pi(x)}\right)$$

Simulation from posterior distribution with large sample size n

- Computing time at least of order $\mathrm{O}(n)$
- solutions using likelihood decomposition

$$\prod_{i=1}^{n} \ell(\theta|x_i)$$

  and handling subsets on different processors (CPU), graphical units (GPU), or computers

  [Korattikara et al. (2013), Scott et al. (2013)]
- no consensus on method of choice, with instabilities from removing most prior input and uncalibrated approximations

  [Neiswanger et al. (2013), Wang and Dunson (2013)]

Proposed solution

"There is no problem an absence of decision cannot solve." Anonymous

Given $\alpha(x,y) := 1 \wedge r(x,y)$, factorise

$$r(x,y) = \prod_{k=1}^{d} \rho_k(x,y)$$

under constraint $\rho_k(x,y) = \rho_k(y,x)^{-1}$
Delayed Acceptance Markov kernel given by

$$\tilde{P}(x,A) := \int_A q(x,y)\tilde{\alpha}(x,y)\mathrm{d}y + \left(1 - \int_{\mathsf{X}} q(x,y)\tilde{\alpha}(x,y)\mathrm{d}y\right)\mathbf{1}_A(x)$$

where

$$\tilde{\alpha}(x,y) := \prod_{k=1}^{d} \{1 \wedge \rho_k(x,y)\}.$$

"There is no problem an absence of decision cannot solve." Anonymous

Algorithm 1 Delayed Acceptance

To sample from $\tilde{P}(x, \cdot)$:

1. Sample $y \sim Q(x, \cdot)$.
2. For $k = 1, \ldots, d$:
   - with probability $1 \wedge \rho_k(x, y)$ continue
   - otherwise stop and output x
3. Output y

Arrange terms in product so that most computationally intensive ones calculated 'at the end' hence least often

"There is no problem an absence of decision cannot solve." Anonymous

Algorithm 1 Delayed Acceptance

To sample from $\tilde{P}(x, \cdot)$:

1. Sample $y \sim Q(x, \cdot)$.
2. For $k = 1, \ldots, d$:
   - with probability $1 \wedge \rho_k(x, y)$ continue
   - otherwise stop and output x
3. Output y

Generalization of Fox and Nicholls (1997) and Christen and Fox (2005), where testing for acceptance with approximation before computing exact likelihood first suggested

More recent occurences in literature

[Golightly et al. (2014), Shestopaloff and Neal (2013)]

Potential drawbacks

- Delayed Acceptance *efficiently* reduces computing cost only when approximation $\tilde{\pi}$ is “good enough” or “flat enough”
- Probability of acceptance always smaller than in the original Metropolis–Hastings scheme
- Decomposition of original data in likelihood bits may however lead to deterioration of algorithmic properties without impacting computational efficiency...
- ...e.g., case of a term explosive in $x = 0$ and computed by itself: leaving $x = 0$ near impossible

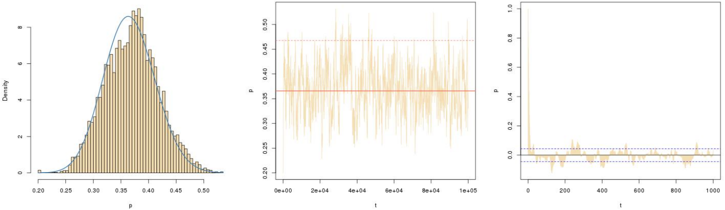


Figure : (left) Fit of delayed Metropolis–Hastings algorithm on a Beta-binomial posterior $p|x \sim Be(x + a, n + b - x)$ when $N = 100$, $x = 32$, $a = 7.5$ and $b = .5$. Binomial $\mathcal{B}(N, p)$ likelihood replaced with product of 100 Bernoulli terms. Histogram based on 10^5 iterations, with overall acceptance rate of 9%; (centre) raw sequence of p's in Markov chain; (right) autocorrelogram of the above sequence.

Delayed Acceptance intended for likelihoods or priors, but not a clear solution for "Big Data" problems

1. all product terms must be computed
2. all terms previously computed either stored for future comparison or recomputed
3. sequential approach limits parallel gains...
4. ...unless prefetching scheme added to delays

[Angelino et al. (2014), Strid (2010)]

Lemma (1)

For any Markov chain with transition kernel Π *of the form*

$$\Pi(x, A) = \int_A q(x, y)a(x, y)\mathrm{d}y + \left(1 - \int_{\mathsf{X}} q(x, y)a(x, y)\mathrm{d}y\right) \mathbf{1}_A(x),$$

and satisfying detailed balance, the function $a(\cdot)$ *satisfies (for* π*-a.e.* x, y*)*

$$\frac{a(x, y)}{a(y, x)} = r(x, y).$$

Lemma (2)

$(\tilde{X}_n)_{n\geq 1}$, the Markov chain associated with $\tilde{P}$, is a π-reversible Markov chain.

Proof.

From Lemma 1 we just need to check that

$$\begin{aligned}\frac{\tilde{\alpha}(x,y)}{\tilde{\alpha}(y,x)} &= \prod_{k=1}^{d} \frac{1 \wedge \rho_k(x,y)}{1 \wedge \rho_k(y,x)} \\ &= \prod_{k=1}^{d} \rho_k(x,y) = r(x,y),\end{aligned}$$

since $\rho_k(y,x) = \rho_k(x,y)^{-1}$ and $(1 \wedge a)/(1 \wedge a^{-1}) = a$

$\square$

The acceptance probability ordering

$$\tilde{\alpha}(x,y) = \prod_{k=1}^{d} \{1 \wedge \rho_k(x,y)\} \leq 1 \wedge \prod_{k=1}^{d} \rho_k(x,y) = 1 \wedge r(x,y) = \alpha(x,y),$$

follows from $(1 \wedge a)(1 \wedge b) \leq (1 \wedge ab)$ for $a, b \in \mathbb{R}_+$.

Remark

By construction of $\tilde{P}$,

$$\mathrm{var}(f, P) \leq \mathrm{var}(f, \tilde{P})$$

for any $f \in L^2(\mathsf{X}, \pi)$, using Peskun ordering (Peskun, 1973, Tierney, 1998), since $\tilde{\alpha}(x,y) \leq \alpha(x,y)$ for any $(x,y) \in \mathsf{X}^2$.

Condition (1)

Defining $A := \{(x, y) \in \mathsf{X}^2 : r(x, y) \geq 1\}$, there exists c such that $\inf_{(x,y)\in A} \min_{k\in\{1,\ldots,d\}} \rho_k(x, y) \geq c$.

Ensures that when

$$\alpha(x, y) = 1$$

then acceptance probability $\tilde{\alpha}(x, y)$ uniformly lower-bounded by positive constant.
Reversibility implies $\tilde{\alpha}(x, y)$ uniformly lower-bounded by a constant multiple of $\alpha(x, y)$ for all $x, y \in \mathsf{X}$.

Condition (1)

Defining $A := \{(x, y) \in \mathsf{X}^2 : r(x, y) \geq 1\}$, there exists c such that $\inf_{(x,y)\in A} \min_{k\in\{1,\dots,d\}} \rho_k(x, y) \geq c$.

Proposition (1)

Under Condition (1), Lemma 34 in Andrieu et al. (2013) implies

$$\mathrm{Gap}(\tilde{P}) \geq \varrho\, \mathrm{Gap}(P) \textit{ and}$$

$$\mathrm{var}(f, \tilde{P}) \leq (\varrho^{-1} - 1)\mathrm{var}_\pi(f) + \varrho^{-1}\mathrm{var}\,(f, P)$$

with $f \in L_0^2(\mathsf{E}, \pi)$, $\varrho = c^{d-1}$.

Proposition (1)

Under Condition (1), Lemma 34 in Andrieu et al. (2013) implies

$$\mathrm{Gap}(\tilde{P}) \geq \varrho\,\mathrm{Gap}(P) \text{ and}$$

$$\mathrm{var}(f, \tilde{P}) \leq (\varrho^{-1} - 1)\mathrm{var}_\pi(f) + \varrho^{-1}\mathrm{var}\,(f, P)$$

with $f \in L_0^2(\mathsf{E}, \pi)$, $\varrho = c^{d-1}$.

Hence if P has right spectral gap, then so does $\tilde{P}$.
Plus, quantitative bounds on asymptotic variance of MCMC estimates using $(\tilde{X}_n)_{n\geq 1}$ in relation to those using $(X_n)_{n\geq 1}$ available

Easiest use of above: modify any candidate factorisation
Given factorisation of r

$$r(x, y) = \prod_{k=1}^{d} \tilde{\rho}_k(x, y)\,,$$

satisfying the balance condition, define a sequence of functions ρ_k such that both $r(x, y) = \prod_{k=1}^{d} \rho_k(x, y)$ and Condition 1 holds.

Take $c \in (0, 1]$, define $b = c^{\frac{1}{d-1}}$ and set

$$\tilde{\rho}_k(x, y) := \min\left\{\frac{1}{b}, \max\{b, \rho_k(x, y)\}\right\}, \quad k \in \{1, \ldots, d-1\},$$

and

$$\tilde{\rho}_d(x, y) := \frac{r(x, y)}{\prod_{k=1}^{d-1} \tilde{\rho}_k(x, y)}.$$

Then:

Proposition (2)

Under this scheme, previous proposition holds with

$$\varrho = c^2 = b^{2(d-1)}$$

.

▸ optimising decomposition For any $f \in L^2(\mathsf{E}, \mu)$ define Dirichlet form associated with a μ-reversible Markov kernel $\Pi : \mathsf{E} \times \mathcal{B}(\mathsf{E})$ as

$$\mathcal{E}_\Pi(f) := \frac{1}{2} \int \mu(\mathrm{d}x)\Pi(x, \mathrm{d}y)\,[f(x) - f(y)]^2 \,.$$

The (right) spectral gap of a generic μ-reversible Markov kernel has the following variational representation

$$\mathrm{Gap}\,(\Pi) := \inf_{f \in L^2_0(\mathsf{E},\mu)} \frac{\mathcal{E}_\Pi(f)}{\langle f, f \rangle_\mu} \quad .$$

optimising decomposition

Lemma (Andrieu et al., 2013, Lemma 34)

Let Π_1 and Π_2 be μ-reversible Markov transition kernels of μ-irreducible and aperiodic Markov chains, and assume that there exists $\varrho > 0$ such that for any $f \in L_0^2(\mathsf{E}, \mu)$

$$\mathcal{E}_{\Pi_2}(f) \geq \varrho \mathcal{E}_{\Pi_1}(f) \quad ,$$

then

$$\mathrm{Gap}(\Pi_2) \geq \varrho \mathrm{Gap}(\Pi_1)$$

and

$$\mathrm{var}(f, \Pi_2) \leq (\varrho^{-1} - 1)\mathrm{var}_\mu(f) + \varrho^{-1}\mathrm{var}(f, \Pi_1) \quad f \in L_0^2(\mathsf{E}, \mu).$$

Explorative performances of random–walk MCMC strongly dependent on proposal distribution
Finding optimal scale parameter leads to efficient 'jumps' in state space and smaller...

1. expected square jump distance (ESJD)
2. overall acceptance rate (α)
3. asymptotic variance of ergodic average $\mathrm{var}(f, K)$

[Roberts et al. (1997), Sherlock and Roberts (2009)]

Provides practitioners with 'auto-tune' version of resulting random–walk MCMC algorithm

Quest for optimisation focussing on two main cases:

1. $d \to \infty$: Roberts et al. (1997) give conditions under which each marginal chain converges toward a Langevin diffusion
   Maximising speed of that diffusion implies minimisation of the ACT and also τ *free* from the functional

Remember $\mathrm{var}(f, K) = \tau_f \times \mathsf{var}_\pi(f)$ where

$$\tau_f = 1 + 2\sum_{i=1}^{\infty} \mathbb{C}\mathrm{or}(f(X_0), f(X_i))$$

Quest for optimisation focussing on two main cases:

2 finite d: Sherlock and Roberts (2009) consider unimodal elliptically symmetric targets and show proxy for ACT is Expected Square Jumping Distance (ESJD), defined as

$$\mathbb{E}\left[\|X' - X\|_\beta^2\right] = \mathbb{E}\left[\sum_{i=1}^{d} \beta_i^{-2}(X_i' - X)^2\right]$$

As $d \to \infty$, ESJD converges to the speed of the diffusion process described in Roberts et al. (1997) [close asymptotia: $d \gtrsim 5$]

"For a moment, nothing happened. Then, after a second or so, nothing continued to happen." — D. Adams, THGG

When considering efficiency for Delayed Acceptance, focus on execution time as well

$$\textbf{Eff} := \text{ESJD}\Big/\text{cost per iteration}$$

similar to Sherlock et al. (2013) for pseudo-Marginal MCMC

Set of assumptions:

(H1) Assume [for simplicity's sake] that Delayed Acceptance operates on two factors only, i.e.,

$$r(x,y) = \rho_1(x,y) \times \rho_2(x,y),$$
$$\tilde{\alpha}(x,y) = \prod_{i=1}^{2} (1 \wedge \rho_i(x,y))$$

Restriction also considers *ideal* setting where a computationally cheap approximation $\tilde{f}(\cdot)$ is available and precise enough so that

$$\rho_2(x,y) = r(x,y)/\rho_1(x,y) = \pi(y)/\pi(x) \times \tilde{f}(x)/\tilde{f}(y) = 1$$

.

(H2) Assume that target distribution satisfies (A1) and (A2) in Roberts et al. (1997), which are regularity conditions on π and its first and second derivatives, and that

$$\pi(x) = \prod_{i=1}^{n} f(x_i)$$

.

(H3) Consider only a random walk proposal

$$y = x + \sqrt{\ell^2/d}\, Z$$

where

$$Z \sim \mathcal{N}(0, I_d)$$

(H4) Assume that cost of computing $\tilde{f}(\cdot)$, c say, proportional to cost of computing $\pi(\cdot)$, C say, with $c = \delta C$.

Normalising by $C = 1$, average total cost per iteration of DA chain is

$$\delta + \mathbb{E}\left[\tilde{\alpha}\right]$$

and efficiency of proposed method under above conditions is

$$\textbf{Eff}(\delta, \ell) = \frac{ESJD}{\delta + \mathbb{E}\left[\tilde{\alpha}\right]}$$

Lemma

Under conditions **(H1)–(H4)** *on* $\pi(\cdot)$, $q(\cdot,\cdot)$ *and on* $\tilde{\alpha}(\cdot,\cdot) = (1 \wedge \rho_1(x,y))$
As $d \to \infty$

$$\mathbf{Eff}(\delta,\ell) \approx \frac{h(\ell)}{\delta + \mathbb{E}\left[\tilde{\alpha}\right]} = \frac{2\ell^2\Phi(-\ell\sqrt{I}/2)}{\delta + 2\Phi(-\ell\sqrt{I}/2)}$$

$$a(\ell) \approx \mathbb{E}\left[\tilde{\alpha}\right] = 2\Phi(-\ell\sqrt{I}/2)$$

where $I := \mathbb{E}\left[\left(\frac{(\pi(x))'}{\pi(x)}\right)^2\right]$ *as in Roberts et al. (1997).*

Proposition (3)

Under conditions of Lemma 3, optimal average acceptance rate $\alpha^*(\delta)$ is independent of I.

Proof.

Consider $\mathbf{Eff}(\delta, \ell)$ in terms of $(\delta, a(\ell))$:

$$a = g(\ell) = 2\Phi\left(-\ell\sqrt{I}/2\right) \qquad \ell = g^{-1}(a) = -\Phi^{-1}(a/2)\frac{2}{\sqrt{I}}$$

$$\mathbf{Eff}(\delta, a) = \frac{\frac{4}{I}\left[\Phi^{-1}\left(\frac{a}{2}\right)^2 a\right]}{\delta + a} = \frac{4}{I}\left\{\frac{1}{\delta + a}\left[\Phi^{-1}\left(\frac{a}{2}\right)^2 a\right]\right\}$$

□

Figure : *top panels:* $\ell^*(\delta)$ and $\alpha^*(\delta)$ as relative cost varies. For $\delta >> 1$ the optimal values converges to values computed for the standard M-H (red, dashed). *bottom panels:* close-up of interesting region for $0 < \delta < 1$.

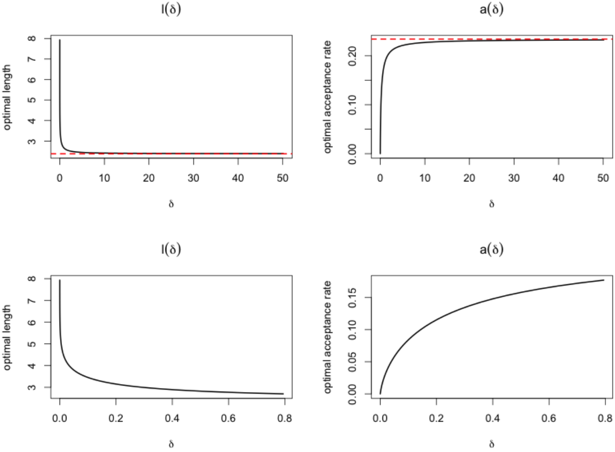
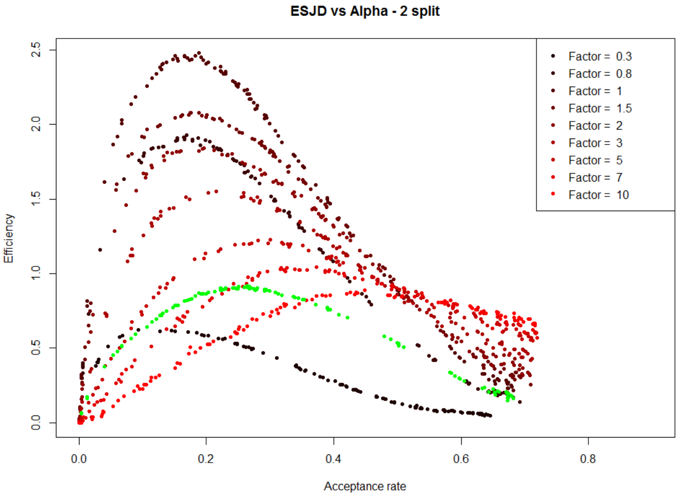


Figure : Efficiency of various DA wrt acceptance rate of the chain; colours represent scaling factor in variance of ρ_1 wrt π; $\delta = 0.3$.



Practical optimisation

If computing cost comparable for all terms in

$$(x, y) = \prod_{i=1}^{K} \xi_i(x, y)$$

- rank entries according to the success rates observed on preliminary run
- start with ratios with highest variances
- rank factors by correlation with full Metropolis–Hastings ratio

Logistic regression:

- 10^6 simulated observations with a 100-dimensional parameter space
- optimised Metropolis–Hastings with $\alpha = 0.234$
- DA optimised via empirical correlation
  - split the data into subsamples of 10 elements
  - include smallest number of subsamples to achieve 0.85 correlation
  - optimise Σ against acceptance rate

algo	**ESS** (av.)	**ESJD** (av.)
DA-MH over MH	5.47	56.18

geometric MALA:

- proposal

$$\theta' = \theta^{(i-1)} + \varepsilon^2 A^T A \nabla_\theta \log(\pi(\theta^{(i-1)}|y))/2 + \varepsilon A \upsilon$$

  with position specific A

  [Girolami and Calderhead (2011), Roberts and Stramer (2002)]
- computational bottleneck in computating 3rd derivative of π in proposal
- G-MALA variance set to $\sigma_d^2 = \frac{\ell^2}{d^{1/3}}$
- 10^2 simulated observations with a 10-dimensional parameter space
- DA optimised via acceptance rate

$$\textbf{Eff}(\delta, a) = -(2/K)^{2/3} \; \frac{a\Phi^{-1}(a/2)^{2/3}}{\delta + a(1-\delta)} \,.$$

algo	**accept**	**ESS/time (av.)**	**ESJD/time (av.)**
MALA	0.661	0.04	0.03
DA-MALA	0.09	0.35	0.31

Jeffreys for mixtures:

- numerical integration for each term in Fisher information matrix
- split between likelihood (cheap) and prior (expensive) unstable
- saving 5% of the sample for second step
- MH and DA optimised via acceptance rate
- actual averaged gain $\left(\frac{ESS_{DA}/ESS_{MH}}{time_{DA}/time_{MH}}\right)$ of 9.58

Algorithm	**ESS** (aver.)	**ESJD** (aver.)	**time** (aver.)
MH	1575.963	0.226	513.95
MH + DA	628.767	0.215	42.22