JMLR: Workshop and Conference Proceedings 1:1–6, 2018. Symposium on Advances in Approximate Bayesian Inference

Dereversibilizing Metropolis-Hastings: simple implementation of non-reversible MCMC methods

Florian Maire
MAIRE@DMS.UMONTREAL.CA
Département de mathématiques et de statistique,
Université de Montréal,
Pavillon André-Aisenstadt, Montréal, QC H3T 1J4, Canada

Editor: Editor's name

Abstract

Recent advances in the field of Markov chain Monte Carlo methods have highlighted the potential gain of using non-reversible Markov chains to reduce the variance of Monte Carlo estimators. However, designing non-reversible Markov chains that have a prescribed stationary distribution is not straightforward. As a result, most algorithms that have been proposed to simulate such Markov chains are only applicable to limited situations (*e.g.* discrete problems). This work develops an easy approach to turn any Metropolis-Hastings type algorithm into a non-reversible Markov chain. Central to this construction is to break the irreducibility of the proposal kernel by introducing a directional auxiliary variable. As a result, the marginal Markov chain is non-reversible and the random-walk behaviour is reduced: this leads to a faster convergence rate compared to the equivalent reversible Markov chain and thus to more efficient algorithms.

Keywords: Markov chain Monte Carlo methods, non-reversible Markov chains, variance reduction technics, Metropolis-Hastings algorithm

1. Introduction

In this paper, we consider the problem of sampling from a probability distribution π defined on a measurable space $(\mathsf{X},\mathcal{X})$ where $\mathsf{X}\subset\mathbb{R}^d$ $(d>0)$ and $\mathcal{X}$ is a sigma-algebra on X and, relatedly, the issue of estimating expectations under π. In Bayesian statistics, π is generally the posterior distribution of the model parameters given the observed data. We are particularly concerned with Markov chain Monte Carlo methods, see Brooks et al. (2011) for an introduction.

Most Bayesian problems tackled by MCMC methods resort to reversible Markov chains: the distribution of the Markov chain $\{X_t,\,t\in\mathbb{Z}\}$ is the same regardless the direction of the time flow, *i.e.* at stationarity, we have $\Pr\left\{\cap_{t\geq 0}(X_t\in A_t)\right\}=\Pr\left\{\cap_{t\geq 0}(X_{-t}\in A_t)\right\}$, for all $A_t\in\mathcal{X}$. So called *reversible* MCMC algorithms include the Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953) and its many variants such as the Hamiltonian Monte Carlo method (Duane et al., 1987), the Reversible Jump MCMC (Green, 1995), random-scan Gibbs sampler (Liu et al., 1995), the Metropolis Adjusted Langevin algorithm (Roberts and Tweedie, 1996), etc. Using reversible chains is essentially motivated by the fact that such Markov chains admit the distribution with respect to which they are reversible as stationary distribution. In other words, MCMC algorithms are easy to construct and analyse when their underlying Markov chain is reversible.

However, a number of references in the diffusion literature have shown that non-reversible Markov processes converge faster than the reversible ones, see for instance Hwang et al. (2005) and Duncan et al. (2016) for the Langevin diffusion case. There is a growing research interest in the statistical and machine learning communities to translate those results into discrete time settings and, to be useful, to design non-reversible MCMC algorithms. While there is a consensus on their efficiency, it is not straightforward to construct non-reversible Markov chains that have a given distribution π for stationary distribution. In fact, additional and non-trivial conditions such as a skew-detailed balance equation are often necessary for π-invariance (Turitsyn et al., 2011; Ottobre et al., 2016; Bierkens, 2016; Poncet, 2017). Even though non-reversible and rejection free methods that use Piecewise Deterministic Markov Processes (Bierkens et al., 2018) or Event Chains (Michel et al., 2014) constitute a promising alternative to traditional MCMC methods, we do not consider them in this paper. We instead propose a simple trick to *dereversibilize* the MH algorithm without any additional condition and which comes at no computational cost. Some simulation results showing some significant gains in variance reduction are outlined.

2. Turning MH into an non-reversible Markov chain

We consider a MH algorithm targeting π and using a proposal kernel Q defined on $(\mathsf{X}, \mathcal{X})$. We assume that Q is specified by a distribution $\underline{Q}$ on $(\mathsf{X}, \mathcal{X})$ such that

$$Q(x, A) = \int_{\mathbb{R}^d} \underline{Q}(\mathrm{d}\xi)\delta_{x+\xi}(A)\,, \tag{1}$$

where δ_x is the dirac probability mass at x. All random-walk MH fall under this decomposition. For completeness, we recall in Alg 1 how the MH algorithm simulates a π-invariant Markov chain using Q. By construction, the Markov chain designed at Alg. 1 is π-reversible

Algorithm 1 MH transition $X_t \to X_{t+1}$

```
Require: $X_t \in \mathsf{X}$
  Draw $X \sim Q(X_t, \cdot)$, $U \sim \mathrm{unif}(0,1)$ and set $X_{t+1} = X_t$
  Calculate the acceptance probability
      $\alpha(X_t, X) := 1 \wedge \frac{\pi(X)Q(X, X_t)}{\pi(X_t)Q(X_t, X)}$      (2)
  if $U \leq \alpha(X_t, X)$ then
    Set $X_{t+1} = X$
  end if
```

and thus admits π as invariant distribution.

2.1. Dereversibilizing MH in dimension one

For pedagogical purposes, we first propose to *dereversibilize* Alg. 1 when $d = 1$. As in most non-reversible MCMC methods, the objective is to construct a π-invariant Markov chain characterized by a momentum in one or several privileged directions. A naive way to achieve this is to truncate $\underline{Q}$ (1) so that its support is (for instance) positive. We denote by Q_+ the resulting proposal kernel. The Markov chain moves always to higher states for an arbitrary large number of iterations until it switches and uses the complementary truncation of $\underline{Q}$ to generate moves with negative increments (leading to Q_-) for another arbitrary large number of iterations, etc. One can immediately see that MH is, by construction, incompatible with this type of scheme. Taking for instance a positive increment streak, the acceptance probability $\alpha(X_t, X)$ (2) is always null since $Q_+(X, X_t) = 0$. Indeed, $X \to X_t$ is a move in the opposite direction to $X \to X_t$ and thus impossible. Remarkably, replacing $Q_+(X, X_t)$ by $Q_-(X, X_t)$ in the MH acceptance probability numerator leads to a valid algorithm (*i.e.* a π-invariant, irreducible and aperiodic Markov chain) as long as the positive/negative streak length follows a specific random dynamic which is specified hereafter. There is obviously a connection with the Zig-Zag sampler from Bierkens et al. (2016).

In the spirit of lifted approaches, an auxiliary variable $\theta \in \Theta_1 := \{-1, 1\}$ is appended to the variable of interest X and π is extended to $\bar{\pi}$ defined as $\bar{\pi}(x, \theta) := (1/2)\pi(x)\mathbb{1}_{\theta\in\Theta}$. Conceptually, the chain will move in the positive direction when $\theta = 1$ and inversely when $\theta = -1$. Starting from some $(X_0, \theta_0) \in \mathsf{X} \times \Theta_1$ and using the *same* distribution $\underline{Q}$ as in Alg. 1, a transition of the *dereversibilized* Markov chain $\{(X_t, \theta_t),\, t \in \mathbb{N}\}$ is described in Alg. 2.

Algorithm 2 Dereversiblized MH transition $(X_t, \theta_t) \to (X_{t+1}, \theta_{t+1})$

```
Require: $X_t \in \mathsf{X}$ and $\theta_t \in \Theta_1$
  Draw $\xi \sim \underline{Q}$, $U \sim \text{unif}(0, 1)$ and set $(X_{t+1}, \theta_{t+1}) = (X_t, -\theta_t)$
  Set $X = X_t + \theta_t|\xi|$
  Calculate the acceptance probability
      $\overline{\alpha}_{\theta_t}(X_t, X) = 1 \wedge \dfrac{\pi(X)\overline{Q}_{-\theta_t}(X, X_t)}{\pi(X_t)\overline{Q}_{\theta_t}(X_t, X)}\,.$   (3)
  if $U \leq \overline{\alpha}_{\theta_t}(X_t, X)$ then
    Accept the proposal: $X_{t+1} = X$ and $\theta_{t+1} = \theta_t$
  end if
```

Proposition 1 *The marginal Markov chain $\{X_t,\, k \in \mathbb{N}\}$ constructed by Alg. 2 is π-invariant and non-reversible.*

Proof We construct a non-homogeneous Markov chain $\{(\tilde{X}_t, \tilde{\theta}_t),\, t \in \mathbb{N}\}$ whose marginal $\{\tilde{X}_t,\, t \in \mathbb{N}\}$ is π-invariant and which satisfies $\{X_t,\, k \in \mathbb{N}\} = \{\tilde{X}_{2t},\, t \in \mathbb{N}\}$ where $\{X_t,\, t \in \mathbb{N}\}$ is the Markov chain specified at Alg. 2. We now detail this construction. If t is even, set $(\tilde{X}_{t+1}, \tilde{\theta}_{t+1}) = (\tilde{X}_t, -\tilde{\theta}_t)$ and if k is odd the chain moves according to the following MH transition: propose

$$(\tilde{X}, \tilde{\theta}) \sim \tilde{Q}(\tilde{X}_t, \tilde{\theta}_t;\, \cdot\,) =: Q_{-\tilde{\theta}_t}(\tilde{X}_t,\, \cdot\,) \otimes \delta_{-\tilde{\theta}_t}(\,\cdot\,)\,, \tag{4}$$

accept the proposition with probability

$$\tilde{\alpha}(\tilde{X}_t, \tilde{\theta}_t; \tilde{X}', \tilde{\theta}') := 1 \wedge \frac{\bar{\pi}(\tilde{X}, \tilde{\theta})\tilde{Q}(\tilde{X}, \tilde{\theta}; \tilde{X}_t, \tilde{\theta}_t)}{\bar{\pi}(\tilde{X}_t, \tilde{\theta}_t)\tilde{Q}(\tilde{X}_t, \tilde{\theta}_t; \tilde{X}, \tilde{\theta})} \tag{5}$$

and reject otherwise. In (4), $Q_\theta(x,\, \cdot\,)$ is the distribution of $x + \theta|\xi|$, $\xi \sim \underline{Q}$. The $\bar{\pi}$-invariance of the extended chain is inherited from either type of transition: trivial from the first one and using the $\bar{\pi}$-reversibility of the MH step for the second. The non-reversibility can be rigorously shown by contradiction. ■

2.2. Dereversibilizing MH in dimension d

Generalizing the spirit of Alg. 2 to a d-dimensional setup requires the auxiliary variable space Θ_d to contain necessarily $2d$ elements (2 for each dimension). However, there is a technical obstacle that prevents from applying the rationale of Alg. 2 to higher dimensional contexts. Indeed, a meticulous analysis of the proof of Prop. 1 shows that it can be generalized if the fictive step $(\theta, X) \to (-\theta, X)$ is replaced by $(\theta, X) \to (\phi(\theta), X)$ where ϕ is any involution on Θ_d. Hence starting from any $\theta_0 \in \Theta_d$, the Markov chain proposed at Alg. 2 would only visit $(\theta_0, \phi(\theta_0)) \subset \Theta_d$, making the Markov chain reducible.

To overcome this issue, we propose a Markov chain whose construction guarantees that θ does visit all the elements of Θ and not only $(\theta_0, \phi(\theta_0))$. Let us first define $\Theta_d = \{-1, 1\}^d$, $\mathsf{I} = \{1, 2, \ldots, d\}$ and the kernel K_i, $i \in \mathsf{I}$, that moves the i-th component of $X_t : X_{i,t} \to X_{i,t+1}$ with the dereversibilized transition (detailed at Alg. 2) that uses θ_i and set $X_{j,t+1} = X_{j,t}$ for all $j \in \mathsf{I} \backslash \{i\}$. We consider here the inhomogeneous Markov chain kernel $K_t := \{K_i : i \equiv t \pmod d\}$, by analogy to the (deterministic scan) Metropolis-within-Gibbs (MwG) sampler. It is straightforward to show that K_t is π-invariant and non-reversible, by combining Prop. 1 and the proof that MwG leaves π-invariant. Therefore, K_t can be seen as a dereversibilized version of a MwG algorithm.

3. Illustration

We present two examples showing that the simple trick introduced in this paper to deversibilize a MH or a MwG sampler, yield to a significant speed-up of the Markov chain convergence and to a variance reduction of Monte Carlo estimators. Example 1 is the distribution of a standard Gaussian r.v. with an additive exponential noise (parameter $\mu = 2$) and Example 2 is the two dimensional banana shape example (parameter $b = 0.03$) from Haario et al. (1999). Table 1 summarizes the results in terms of variance reduction. Visit http://maths.ucd.ie/~fmaire/conv.html for an animation showing the gain in terms of speed of convergence in the context of Ex. 1.

Table 1: Comparison of MH and its dereversibilized version (drvMH) – Effective Sample Size (ESS, as defined in Neal (1993)), σ^2_f is the asymptotic variance of the MC esimator of $\int f \mathrm{d}\pi$, $f_1(x) = \sum_{i=1}^d x_i$, $f_2(x) = \mathbb{1}_{x>5}$ (for ex. 1) and $f_2(x) = \mathbb{1}_{\{x_2>10\} \cup \{x_2<-10\}}$ (for ex. 2). All results were estimated from 1,000 iid Markov chains of length 1,000 for each algorithm, starting with $X_0 \sim \pi$.

		ESS	$\sigma^2_{f_1}$	$\sigma^2_{f_2}$
π_1	MH	0.12	32.1	0.48
	drvMH	0.18	19.6	0.30

		ESS	$\sigma^2_{f_1}$	$\sigma^2_{f_2}$
π_2	MH	0.11	23.3	0.099
	drvMH	0.15	9.76	0.056

Acknowledgments

I acknowledge the support of the Université de Montréal through the *fond de démarrage* programme.

References

Joris Bierkens. Non-reversible Metropolis-Hastings. *Statistics and Computing*, 26(6):1213–1228, 2016.

Joris Bierkens, Paul Fearnhead, and Gareth Roberts. The zig-zag process and super-efficient sampling for Bayesian analysis of big data. *arXiv preprint arXiv:1607.03188*, 2016.

Joris Bierkens, Alexandre Bouchard-Côté, Arnaud Doucet, Andrew B Duncan, Paul Fearnhead, Thibaut Lienart, Gareth Roberts, and Sebastian J Vollmer. Piecewise deterministic Markov processes for scalable Monte Carlo on restricted domains. *Statistics & Probability Letters*, 136:148–154, 2018.

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov chain Monte Carlo*. CRC press, 2011.

Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics letters B*, 195(2):216–222, 1987.

Andrew B Duncan, Tony Lelievre, and GA Pavliotis. Variance reduction using nonreversible Langevin samplers. *Journal of statistical physics*, 163(3):457–491, 2016.

Peter J Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

Heikki Haario, Eero Saksman, and Johanna Tamminen. Adaptive proposal distribution for random walk Metropolis algorithm. *Computational Statistics*, 14(3):375–396, 1999.

Chii-Ruey Hwang, Shu-Yin Hwang-Ma, and Shuenn-Jyi Sheu. Accelerating diffusions. *The Annals of Applied Probability*, 15(2):1433–1444, 2005.

Jun S Liu, Wing H Wong, and Augustine Kong. Covariance structure and convergence rate of the Gibbs sampler with various scans. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 157–169, 1995.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Manon Michel, Sebastian C Kapfer, and Werner Krauth. Generalized event-chain Monte Carlo: Constructing rejection-free global-balance algorithms from infinitesimal steps. *The Journal of chemical physics*, 140(5):054116, 2014.

Radford M Neal. Probabilistic inference using Markov chain Monte Carlo methods. 1993.

Michela Ottobre, Natesh S Pillai, Frank J Pinski, Andrew M Stuart, et al. A function space HMC algorithm with second order Langevin diffusion limit. *Bernoulli*, 22(1):60–106, 2016.

Romain Poncet. Generalized and hybrid Metropolis-Hastings overdamped Langevin algorithms. *arXiv preprint arXiv:1701.05833*, 2017.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

Konstantin S Turitsyn, Michael Chertkov, and Marija Vucelja. Irreversible Monte Carlo algorithms for efficient sampling. *Physica D: Nonlinear Phenomena*, 240(4-5):410–414, 2011.