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

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. A notable exception is Gustafson's Guided Walk (GW) algorithm (Gustafson, 1998). This method is designed for unidimensional problems, which is perhaps why it has received little attention both in theory and in practice. This work presents a generalization of the GW algorithm which preserves its unidimensional efficiency to multidimensional settings.

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 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).

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 measure. In other words, MCMC algorithms are easy to construct and analyse when their underlying Markov chain is reversible (spectral decomposition, etc.).

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). 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 whose limiting distribution π can be arbitrarily prescribed. In fact, additional and non-trivial conditions such as a skew-detailed balance equation are often necessary to guarantee π-invariance (Turitsyn et al., 2011; Ottobre et al., 2016; Bierkens, 2016; Poncet, 2017)[1].

2. The Guided Walk

A notable exception to the above mentioned non-reversible methods is the Guided Walk (GW) algorithm (Gustafson, 1998) whose simplicity and requirements are similar to MH. As noted by Gustafson, GW rationale makes sense only when π is unidimensional. We briefly present GW's essential idea, by emphasizing its proximity with MH, in the case $d = 1$. Let Q be a transition kernel on $(\mathsf{X}, \mathcal{X})$ such that $Q(x, \mathrm{d}x') = q_x(x' - x)\mathrm{d}x'$, where for simplicity q_x is an even function for all $x \in \mathsf{X}$. In GW, an auxiliary variable $\theta \in \Theta := \{-1, 1\}$ 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 GW chain moves in the positive direction when $\theta = 1$ and inversely when $\theta = -1$. Given some initial variables $X_0 \in \mathsf{X}$ and $\theta_0 \in \Theta$, a transition $X_t \to X_{t+1}$ of the MH and GW Markov chains are described in Alg. 1 and 2, respectively. Even though Prop. 1 is mentioned in Gustafson (1998), we provide here a rigorous proof which will help justifying the proposed Guided Walk generalization.

Algorithm 1 Random Walk (MH)

```
set $X_{t+1} = X_t$
draw $\zeta \sim q_{X_t}$, $U \sim \text{unif}(0, 1)$
set $X = X_t + \zeta$
calculate the acceptance probability
    $\alpha(X_t, X) = 1 \wedge \frac{\pi(X) q_X(X - X_t)}{\pi(X_t) q_{X_t}(X_t - X)}$
if $U \leq \alpha(X_t, X)$, set $X_{t+1} = X$
```

Algorithm 2 Guided Walk (GW)

```
set $X_{t+1} = X_t$ and $\theta_{t+1} = -\theta_t$
draw $\zeta \sim q_{X_t}$, $U \sim \text{unif}(0, 1)$
set $X = X_t + \theta_t|\zeta|$
calculate the acceptance probability
    $\alpha(X_t, X) = 1 \wedge \frac{\pi(X) q_X(X - X_t)}{\pi(X_t) q_{X_t}(X_t - X)}$
if $U \leq \alpha(X_t, X)$, set $X_{t+1} = X$ and $\theta_{t+1} = \theta_t$
```

Proposition 1 *The sequence of r.v. $\{X_t,\, t \in \mathbb{N}\}$ constructed by GW 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: if t is even, set $(\tilde{X}_{t+1}, \tilde{\theta}_{t+1}) = (\tilde{X}_t, -\tilde{\theta}_t)$ and if t is odd the chain moves according to the following MH transition: propose

$$(\tilde{X}, \tilde{\theta}) \sim \tilde{Q}(\tilde{X}_t, \tilde{\theta}_t;\, \cdot\,) := \int q_{\tilde{X}_t}(\mathrm{d}\zeta)\delta_{\tilde{X}_t - \tilde{\theta}_t|\zeta|}(\,\cdot\,) \otimes \delta_{-\tilde{\theta}_t}(\,\cdot\,)\,, \tag{1}$$

accept the proposition with probability $\alpha(\tilde{X}_t, \tilde{X}')$ defined above and reject otherwise. The $\bar{\pi}$-invariance of the extended chain is inherited from either type of transition: trivial from the

[1] Even though non-reversible and rejection free methods based on Piecewise Deterministic Markov Processes (Bierkens et al., 2018) constitute a promising alternative to traditional MCMC methods, they represent a significant departure from MH and we do not consider them in this paper.

first one and using the $\bar{\pi}$-reversibility of the MH step for the second. Indeed, assuming that π admits a density wrt to a common dominating measure $\mathrm{d}x$, it can be shown that the Radon-Nikodym derivative of $\mathrm{d}\bar{\pi}(x,\theta)\tilde{Q}(x,\theta;x',\theta')$ wrt $\mathrm{d}\theta\mathrm{d}\theta'$ is simply $2\pi(x)q_x(x-x')\mathbb{1}_{\theta,\theta'\in\{-1,1\}^2}$, leading to the same acceptance probability as MH. The non-reversibility can be rigorously shown by contradiction. ■

As observed by Gustafson (1998), it is possible to extend GW to higher-dimensional setting by applying a series of GW unidimensional updates (in a Gibbs fashion) but this approach loses the essence of the Guided Walk as the chain evolves in a series of orthogonal directions, as opposed to in straight line. This work presents a generalization of the GW algorithm preserving the guided aspect of the dynamic even in multivariate settings.

3. Generalized Guided Walk

In contrast to the case $d = 1$, general sampling problems in $d > 1$ offer an infinite number of propagation directions. In the spirit of the original GW, a guided walk in dimension $d > 1$ should thus move according to specific directions until a proposed candidate is rejected and the momentum flipped. More precisely, let $\mathcal{H}$ be a subspace of X, $\rho_{\mathcal{H}}$ a projection of X onto $\mathcal{H}$ and $\boldsymbol{e}\in\mathcal{H}$ be a *directional vector*. The GW is generalized by proposing a perturbation to the current state along $\mathcal{H}$, in either of the half-spaces directed by $\boldsymbol{e}$, according to the momentum variable θ. The proof of Prop. 1 transposes well in this setting and we show that for any couple $(\mathcal{H},\boldsymbol{e})$ the Markov kernel whose proposal follows this dynamic and summarized at Alg. 3 is, marginally, π-invariant. However, if $\dim(\mathcal{H}) < d$ the Markov chain is reducible. To guarantee that π is indeed the limiting distribution of the chain, it is necessary to consider a collection of kernels $P_1,\ldots,P_r$ each of which moves in a subspace $\mathcal{H}_i$ according to its directional vector $\boldsymbol{e}_i\in\mathcal{H}_i$ and momentum $\theta_i\in\{-1,1\}$, such that there exists a subset of $(\boldsymbol{e}_1,\ldots,\boldsymbol{e}_r)$ spanning X.

Naturally, the proposed method is particularly useful when the probability mass of π is concentrated around lower dimensional subspaces, in which case the direction of propagation is very specific provided that $\mathcal{H}_1,\mathcal{H}_2,\ldots$ are meaningfully chosen. This situation arises in Bayesian inverse problems (Knapik et al., 2011) or in models in demography (Raftery and Bao, 2010). We assume that the collection of subspaces, taken linear as a first approximation, exist. Those can for instance be taken outputs of a principal component analysis on samples of π obtained using a simple MH algorithm. Other learning approaches are possible (Adaptive MCMC, online PCA, etc.) but the theoretical aspect of such constructions has to be carefully addressed and are left for future works.

Algorithm 3 Generalized Guided Walk (GGW)

```
set $X_{t+1} = X_t$ and $\theta_{t+1} = \theta_t$
draw $U \sim \mathrm{unif}(0,1)$, $i \in \{1,\ldots,r\}$ unif. at random
conditionally on $i$, draw $\zeta \sim q^i_{X_t}$ where $q^i_{X_t}(\mathrm{d}\zeta) \propto q_{X_t}(\mathrm{d}\zeta)\mathbb{1}_{\rho_{\mathcal{H}_i}(\zeta)^{\mathrm{T}}\boldsymbol{e}_i>0}$
set $X = X_t + \theta^i_t\rho_{\mathcal{H}_i}(\zeta)$ and $\theta^i_{t+1} = -\theta^i_t$
calculate the acceptance probability
    $$\alpha_i(X_t,X) = 1 \wedge \frac{\pi(X)q^i_X(X-X_t)}{\pi(X_t)q^i_{X_t}(X-X_t)}$$
if $U \le \alpha_i(X_t,X)$, set $X_{t+1} = X$ and $\theta^i_{t+1} = \theta^i_t$
```

4. Illustrations

We consider three examples: the two-dimensional banana shape example from Haario et al. (1999) with twisting parameter $b = 0.03$, a mixture of four centered two-dimensional Gaussian distributions each stretched in a particular direction given by vectors $\{u_i, v_i\}_{i=1}^4$ (covariance structure $\Gamma_i = [u_i\, v_i]\text{diag}(1, 1/1000)[u_i\, v_i]^{\text{t}}$) and a three-dimensional distributions featuring one and two dimensional subspaces, namely the distribution of (X_1, X_2, X_3) specified as follows. With probability 0.01, set $Z_1 = Z_2 = 0$ and draw $Z_3 \sim \text{unif}(-1, 0)$ and with probability 0.99, draw $Z_1 \sim \text{unif}(-1, 1)$, set $Z_2 = |Z_1|$, draw $Z_3 \sim \text{unif}(0, 1)$ and with equal probability, set $Z_1 = Z_1$, $Z_1 = Z_1 + 2$ or $Z_1 = Z_1 - 2$. Finally set $X_i = Z_i + 0.05\zeta$, $\zeta \sim \mathcal{N}(0, .1)$.

We compare GGW with an equivalent MH algorithm, namely the exact same algorithm as Alg. 3 but where $\theta_t^i \sim \text{ber}(1/2)$ is drawn afresh at each iteration. Remarkably, this simple difference confers a significant speed-up in convergence to the GGW Markov chain compared to MH and yield a variance reduction of Monte Carlo estimators when using GGW compared to MH. Proposal distributions are defined as $q_x = \mathcal{N}(0, \sigma^2)$ where σ is set so MH achieves a 40% acceptance probability.

Table 1: Comparison of MH and GGW in stationary regime, $X_0 \sim \pi$. Effective Sample Size (ESS, Neal (1993)), σ_f^2 is the asymptotic variance of the MC estimator of $\int f \mathrm{d}\pi$, $f_1(x) = x_1 + \cdots + x_d$, $f_2(x) = \mathbb{1}_{|x_2|>10}$ and $f_3(x) = \mathbb{1}_{\|x\|>2}$. All results were estimated from 1,000 i.i.d Markov chains of length 1,000 for each algorithm.

		ESS	$\sigma_{f_1}^2$	$\sigma_{f_2}^2$		ESS	$\sigma_{f_1}^2$	$\sigma_{f_2}^2$
Ex. 1	MH	0.02	23.3	5.32	Ex. 2	0.01	42.9	13.88
	GGW	0.06	9.76	2.51		0.05	11.4	3.06

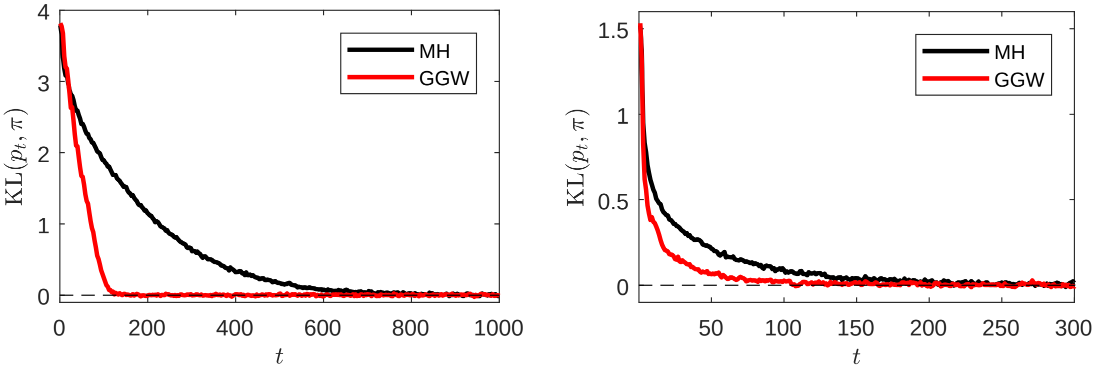

Figure 1: Convergence to stationarity of MH and GGW in Examples 2 (right) and 3 (left), measured in terms of Kullback-Leibler divergence between the chain law and π. Initial laws were respectively set as $\mathcal{N}([3, 3], 0.1\,\text{Id})$ and $\mathcal{N}([0, 0, 0], 0.1\,\text{Id})$. KL divergences were estimated using the nearest neighbor entropy estimator proposed in Chauveau and Vandekerkhove (2014).

Acknowledgments

I acknowledge the support of the Université de Montréal.

References

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

Didier Chauveau and Pierre Vandekerkhove. The nearest neighbor entropy estimate: an adequate tool for adaptive MCMC evaluation. 2014.

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.

Paul Gustafson. A guided walk Metropolis algorithm. *Statistics and computing*, 8(4):357–364, 1998.

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.

Bartek T Knapik, Aad W van der Vaart, J Harry van Zanten, et al. Bayesian inverse problems with gaussian priors. *The Annals of Statistics*, 39(5):2626–2657, 2011.

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.

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.

Adrian E Raftery and Le Bao. Estimating and projecting trends in HIV/AIDS generalized epidemics using incremental mixture importance sampling. *Biometrics*, 66(4):1162–1173, 2010.

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.