Review of Unbiased Markov chain Monte Carlo with couplings

Matthieu Dinot
Ecole Polytechnique
Palaiseau, France
matthieu.dinot@polytechnique.edu

Yvann Le Fay
ENSAE
Palaiseau, France
yvann.lefay@ensae.fr

Abstract

Markov chain Monte Carlo (MCMC) methods provide asymptotic exact estimators of expectations under a target distribution π. It typically proceeds by, first, designing a Markov chain X_t whose stationary distribution is π, and, second, computing an empirical counterpart to the expectation using the X_t's. However, such estimators present a bias in the non-asymptotic regime, this ultimately hinders the parallelization of the MCMC technique as one would capture a finite-time bias. For this reason, there is a growing interest in designing unbiased estimators in the non-asymptotic regime. Jacob et al. [2020] propose an unbiased estimator H with finite variance which can be computed in finite time using a coupled Markov chain $\{(X_t, Y_t)\}$. In this report, we review the construction of Jacob et al. [2020] for the coupled Metropolis-Hasting (MH) algorithm. In particular, we numerically assess the statistical and computational properties of the unbiased estimator, by doing so, we reproduce experimental results from Jacob et al. [2020]. As a minor contribution and following the discussion by Gerber and Lee. [2020], we empirically study the impact on the variance of the cost of using a modification of the maximal-coupling algorithm.

1 Introduction

Markov chain Monte Carlo (MCMC) methods constitute a powerful paradigm to numerically estimate intractable expectations by leveraging Markov chains. The setup is the following, let $h : \mathcal{X} \to \mathbb{R}$ be some well-behaved function, and let π be the target distribution we are interested in integrating h against. In the context of Bayesian inference, h is typically a posterior-based quantity and π is the posterior distribution given some observations. Let P be the transition kernel on $\mathcal{X}$ that leaves π invariant. We define the following Markov chain, let $X_0 \sim \pi_0$ where π_0 is an initial distribution, and $X_{t+1} \sim P(X_t, \cdot)$. Then, under some mild assumptions, the asymptotic normal convergence of the MCMC estimator (1) can be proved [Robert and Casella, 2004, Th. 6.64],

$$\frac{1}{N}\sum_{t=1}^{N} h(X_t) \to \mathcal{N}(\mathbb{E}_\pi[h(X)], \sigma_h^2/N), \tag{1}$$

for some $\sigma_h \geq 0$. However, the MCMC estimator (1) exhibits a bias after any finite number of iterations. To correct for this non-asymptotic bias, Jacob et al. [2020] design a coupled chain $\{(X_t, Y_t)\}$ evolving according to a coupled transition kernel $\bar{P}$ on $\mathcal{X} \times \mathcal{X}$, such that the marginal chains $\{X_t\}$ and $\{Y_t\}$ admit P as a transition kernel, and is such that the two chains will meet in almost surely finite time $\tau = \inf\{t \geq L, X_t = Y_{t-L}\}$. Decomposing $\mathbb{E}[h(X)]$ as the expectation of a telescopic

sum and using the fact that the marginal chains have the same distribution, we obtain

$$\mathbb{E}[h(X)] = \underbrace{\mathbb{E}[h(X_k) + \sum_{j=1}^{\infty} h(X_{k+jL}) - h(Y_{k+(j-1)L})]}_{H_{k,L}}, \tag{2}$$

for any k. Let $0 \leq k \leq m$, the estimator $H_{k:m,L}$ is obtained by averaging over the $m - k + 1$ estimates $H_{n,L}$ for $n \in [k, m]$. After some algebraic manipulations, the estimate is given by the finite-size summation,

$$H_{k:m,L}(X, Y) = \frac{1}{m-k+1} \sum_{l=k}^{m} h(X_l) + \sum_{l=k+L}^{\tau-1} v_l(k, m, L)(h(X_l) - H(Y_{l-L})), \tag{3}$$

where $v_l(k, m, L) = (\lfloor (l-k)/L \rfloor - \lceil \max(L, l-m)/L \rceil + 1)/(m-k+1)$ [Atchadé and Jacob, 2023]. Under some technical assumptions on the initial and target distributions, the test function h, as well as the coupled transition kernel $\bar{P}$, H can be shown to be an asymptotically normal estimator of $\mathbb{E}_\pi[h(X)]$ in the length of the Markov chain [Jacob et al., 2020][Th 1.2.2].

2 The cost-variance trade-off

The unbiasedness of (3) comes at the price of an increased variance of the estimator compared to the classic MCMC estimator. Jacob et al. [2020] derive an approximate upper bound on $\mathbb{V}[H_{k:m,L}]$ which says that in the long-running time, i.e., k and $m - k$ are both large, we essentially recovers the variance of the classic MCMC estimator. However, running for a long time the chain means an increased computational cost. Thus, as a meaningful performance measure, Jacob et al. [2020] consider the inefficiency statistic given by a product between the average cost of the procedure, which is essentially $O(\mathbb{E}[\tau] + m)$, and the variance $\mathbb{V}[H_{k:m,L}]$. The time-averaged estimator (3) is an MCMC estimate with a correction that term that vanishes once the chains have coupled. Choosing the burning period k to be large compared to τ, i.e., such that $\mathbb{P}[\tau > k] \approx 0$, will ensure that with high probability the estimate agrees with the MCMC estimate, thus leading to the optimal variance. Since k quantifies the amount of wasted states of the chain, one is interested in designing the coupling procedure to obtain early coupling of the chain. However, it is not clear how to achieve that. Further theoretical analysis of the properties of τ and the correction term is expected to be useful for understanding how to practically choose k and m given some desired computational budget and statistical performance.

3 Experiments

Throughout this section, we numerically assess the statistical properties of the unbiased estimator $H_{k:m,L}$. In particular, we reproduce the experiment from Jacob et al. [2020], i) of the scaling properties of the meeting time with respect to the dimension, and ii) of the impact of k, m and the lag L on the variance of $H_{k:m,L}$. The JAX code to reproduce the experiments is available at `https://github.com/ylefay/unbiased_mcmc_with_couplings`. All three experiments are run over 1000 independent replications.

3.1 The impact of the dimension on the meeting time

As a proof of concept, we reproduce the experience of Section 4.2 of Jacob et al. [2020], of the impact on the meeting time of the dimension d.

The target distribution is $\mathcal{N}(0, V)$, where V is sampled from the Inverse-Wishart distribution with d degrees of freedom and scale parameter I_d. We consider two cases, the initial sample being from i) the target distribution and ii) from the offset multivariate normal $\mathcal{N}(\mathbb{1}_d, I_d)$. The proposal for the MH kernel is $\mathcal{N}(0, V/d)$. As expected, the meeting times scale poorly with the dimension, see Figure 1.

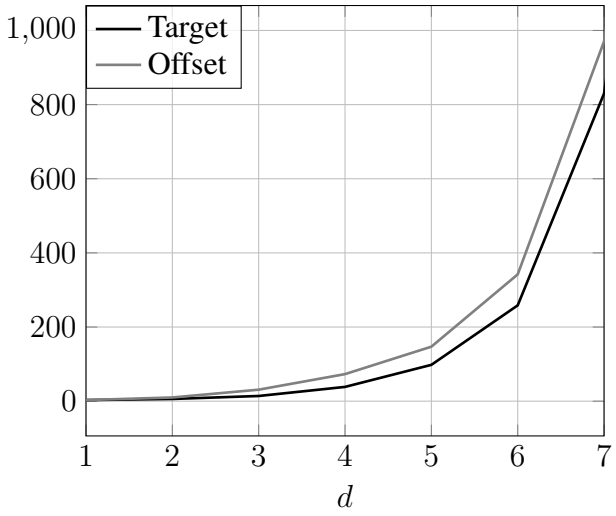

Figure 1: Median of the meeting time τ for both initial distribution cases.

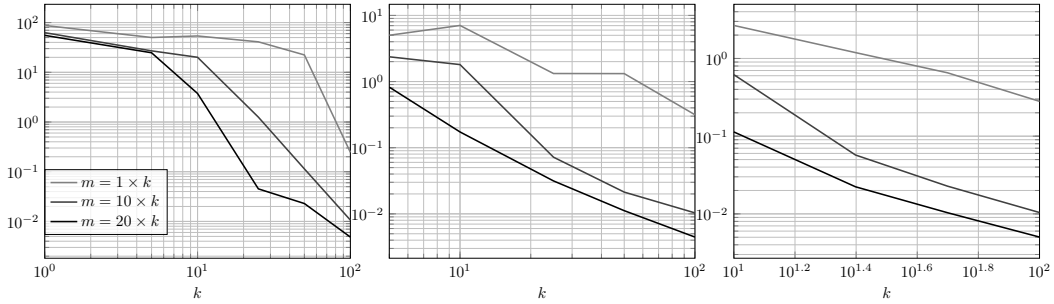

Figure 2: Empirical variance of the time-average estimate $H_{k:m,L}$ with respect to k for $1 \leq k \leq 100$. In log-log scale. From the left to the right, $L = 1, 5, 10$.

3.2 The impact of k, m and L on the mean cost and the variance of $H_{k:m,L}$

In this section, we reproduce the results of Section 5.1 with a minor contribution by including the analysis of the lag parameter L on the mean cost and variance of $H_{k:m,L}$. We consider a mixture of two normal distributions $0.5\mathcal{N}(-4, 1) + 0.5\mathcal{N}(4, 1)$ as a target distribution, and we draw the sample from the right mode of the target distribution. The proposal for the MH kernel is a normal with variance 3^2. The test function is $h(x) = \mathbb{1}(x \geq 3)$. As expected, the variance decreases with k and $m - k$, see Figure 2.

3.3 Controlling the variance of the cost

For the sake of simplicity and without loss of generalisation, we set $L = 1$. Following the discussion by Gerber and Lee. [2020] on a modified version of the maximal coupling Algorithm, we let $\phi : \mathcal{X} \to [0, 1]$ be a function such that $\phi(x) \leq q(x)/p(x)$ for each $x \in \mathcal{X}$. It can be shown that indeed (X, Y) given by Algorithm 1 has the right marginals p and q respectively. Let us denote by N the number of sampling procedures from q in Algorithm 1, which we call the cost, then $N = (1 - B)G$, where $B \mid X \sim \text{Bernoulli}(\phi(X))$ and $G \sim \text{geometric}(1 - p(\phi))$ are independent and where $p(\phi) = \int \phi(x)p(x)\mathrm{d}x$. Thus, the average cost is $\mathbb{E}[N] = 1$ and its variance is $2p(\phi)/(1 - p(\phi))$. Furthermore, the coupling probability is the probability of the first if-condition to be true, i.e., $P(X = Y) = p(\phi)$, which, by the maximal coupling inequality [Torgny, 1992], is maximised when $\phi = \min(q/p, 1)$.

We are interested in controlling the variance of the total cost $\tau + \sum_{i=1}^{\tau} N_i + \max(1, m + 1 - \tau)$, where the N_i's are the cost of running the i-th iteration of Algorithm 1. This is essentially the same

Sample $X \sim p$, $W \sim \mathcal{U}[0,1]$
if $W \leq \phi(x)$ **then**
  $(X, Y) \leftarrow (X, X)$
else
  repeat
    Sample $Y \sim q$, $W \sim \mathcal{U}[0,1]$
  until $W > \phi(Y)p(Y)/q(Y)$
end
output : Sample from a coupling Γ_ϕ between p and q
return (X, Y)

Algorithm 1: Modified Thorisson's Algorithm

as controlling $\mathbb{V}\left(\sum_{i=1}^{\tau} N_i\right)$. We have by the total variance formula,

$$\mathbb{V}\left(\sum_{i=1}^{\tau} N_i\right) = \mathbb{V}\left(\mathbb{E}\left(\sum_{i=1}^{\tau} N_i \mid \tau, X_{\tau-1}\right)\right) + \mathbb{E}\left(\mathbb{V}\left(\sum_{i=1}^{\tau} N_i \mid \tau, X_{\tau-1}\right)\right). \tag{4}$$

Since all the $N_i \mid \tau, X_{\tau-1}$ have same expectancy $\mathbb{E}[N] = 1$, the first term is equal to $\mathbb{V}[\tau]$. Let us tackle the second term, using Cauchy-Schwarz inequality, we have

$$\mathbb{V}\left(\sum_{i=1}^{\tau} N_i \mid \tau, X_{\tau-1}\right) \leq \tau(\mathbb{V}[N_\tau \mid \tau, X_{\tau-1}] + \sum_{i=1}^{\tau-1} \mathbb{V}[N_i]). \tag{5}$$

Let $\eta \in [0,1]$ and let us assume that ϕ_t is given at each time step $1 \leq t \leq \tau - 1$ by $\phi_t(x) = \min(q(Y_{t-1}, x)/p(X_t, x), \eta)$. Using the equality on the variance of N_τ, we have $\mathbb{V}[N_\tau \mid \tau, X_{\tau-1}] = 2\int_{\mathcal{X}} \phi_{\tau-1}(x)p(X_{\tau-1}, x)\mathrm{d}x/(1 - \int_{\mathcal{X}} \phi_{\tau-1}(x)p(X_{\tau-1}, x)\mathrm{d}x) \leq \frac{2\eta}{1-\eta}$, thus, the previous inequality becomes

$$\mathbb{V}\left(\sum_{i=1}^{\tau} N_i \mid \tau, X_{\tau-1}\right) \leq \tau\left(\frac{2\eta}{1-\eta} + \sum_{i=1}^{\tau-1} \mathbb{V}[N_i]\right) \leq \tau^2 \frac{2\eta}{1-\eta}. \tag{6}$$

Finally, plugging all the previous results into (4), the variance of the total cost is bounded by

$$\mathbb{V}\left(\sum_{i=1}^{\tau} N_i\right) \leq \mathbb{V}[\tau] + \frac{2\eta}{1-\eta}\mathbb{E}[\tau^2], \tag{7}$$

where the distribution of τ depends upon η. As an additional heuristic analysis, suppose we have access to the target distribution π. We obtain that the probability of coupling is $\mathbb{P}[X \neq Y] = \eta$, thus, $\tau \sim$ geometric$(1-\eta)$. In that case, we obtain $\mathbb{V}\left(\sum_{i=1}^{\tau} N_i\right) \lesssim \frac{3\eta+1}{(\eta-1)\ln(1-\eta)^2}$. This corresponds to η worsening the sampling procedure in the best case when both proposals sample from π. We run the simulation with the initial and the target distributions defined by $\mathcal{N}(0, \sigma^2)$ where $\sigma^2 \sim$ inverse-gamma$(1/2, 1/2)$, which is the same setting as 3.1 for $d = 1$. See Figure 3 for the bound (7) with respect to η for both, the example and the perfect Monte Carlo instance.

The variance of the total cost is indeed of interest when parallelising the computation of estimate. To illustrate this, we implement the coupled MH algorithm in JAX and we run the experiment N times. Each experiment includes passing sequentially through a `while_loop` due to the sampling procedures given by Algorithm 1. A `vmap` creates a `while_loop` that does not terminate untill all mapped loops terminate. Thus, at each time step, the total cost of running the needed sampling procedures depends upon the slowest sampling procedure which scales with the variance $\mathbb{V}[N]$.

4 Discussion

In this report, we reviewed the coupled MCMC estimate derived in Jacob et al. [2020] in the MH instance where the Thorisson's Algorithm is used as a coupling procedure. We numerically assessed the statistical properties of the unbiased estimate $H_{k:m,L}$ on some toy examples. As a contribution, we have studied the impact of the choice of ϕ_η in the Thorisson's modified algorithm on the variance

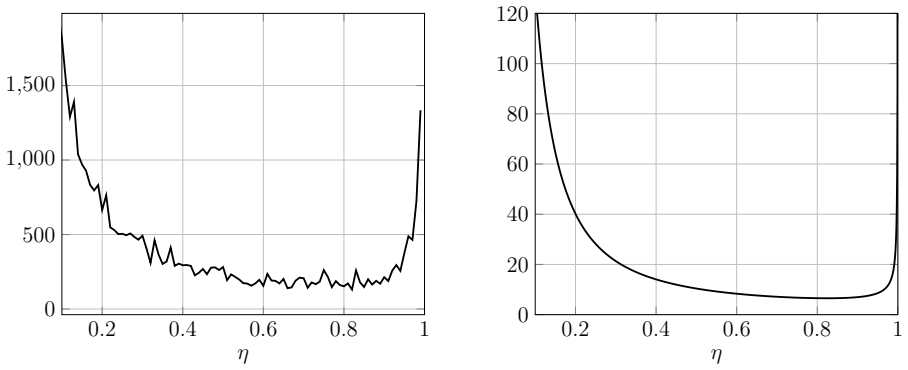

Figure 3: Bound on the variance of the total cost. Left: realisation of bound (6). Right: theoretical bound (7) for a perfect Monte Carlo chain coupling.

of the total cost. In particular, in the case where we are interested in controlling the cost, we've empirically shown that there might exist some optimal η that minimizes the variance of the total cost. We do believe that the theoretical study of the impact of η on the computational performance of the procedure is of interest for future work.

Acknowledgments

The authors are grateful to Adrien Corenflos for the valuable discussion on the runtime variance in the perfect Monte Carlo instance.

References

Yves F. Atchadé and Pierre E. Jacob. Unbiased Markov Chain Monte Carlo: what, why and how, 2023. URL `https://math.bu.edu/people/atchade/umcmc_rev.pdf`.

Mathieu Gerber and Anthony Lee. Discussion on the paper by Jacob, O'Leary, and Atchad´e. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(3):584–585, 05 2020. ISSN 1369-7412. doi: 10.1111/rssb.12336. URL `https://doi.org/10.1111/rssb.12336`.

Pierre E. Jacob, John O'Leary, and Yves F. Atchadé. Unbiased Markov Chain Monte Carlo Methods with Couplings. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(3):543–600, 05 2020. ISSN 1369-7412. doi: 10.1111/rssb.12336. URL `https://doi.org/10.1111/rssb.12336`.

C.P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Verlag, 2004.

Lindvall Torgny. *Lectures on the coupling method*. New York: Dover Publications, 1992.