Chapter 1

Unbiased Markov Chain Monte Carlo: what, why and how

Yves F. Atchadé, Pierre E. Jacob

1.1 Introduction

This chapter presents techniques to remove the *initialization* or *burn-in* bias from MCMC estimates, with consequences on parallel computing, convergence diagnostics and performance assessment. First we define the bias under consideration, then we present the benefits of removing that bias with *unbiased MCMC* methods, and how to do it.

1.1.1 Initialization bias in MCMC

The object of interest is a probability measure π on a space $(\mathbb{X}, \mathcal{X})$. An MCMC algorithm generates a chain $(X_t)_{t\geq 0}$ via a π-invariant Markov transition kernel P, starting from an initial distribution π_0 which is not equal to π. The marginal distribution of X_t at time t is

denoted by π_t. MCMC algorithms are often provably ergodic in the sense

$$|\pi_t - \pi| \leq b(t) \to 0 \text{ as } t \to \infty, \tag{1.1.1}$$

for some decreasing function b, e.g. in total variation (TV) distance. Results of the form (1.1.1) abound in the literature, but the function $b(t)$ typically features unspecified quantities, and thus cannot actually be evaluated for a given iteration t. There are exceptions, e.g. Theorem 11 of Example 2 in Rosenthal (1995), or the references in Section 3.5 in Roberts and Rosenthal (2004), where bounds are made fully explicit for non-trivial MCMC algorithms, using both analytical and numerical computation. Efforts have been long made to design numerical recipes that would provide explicit upper bounds as automatically as possible (e.g. Cowles and Rosenthal, 1998; Johnson, 1996, 1998) and unbiased MCMC methods contribute to that effort (more on this in Section 1.3.1).

The initialization bias comes from the marginal distribution π_t, at any time t, being different from π. We introduce a test function h in $L^p(\pi) = \{f : \pi(|f|^p) < \infty\}$ where $\pi(f) := \int f(x)\pi(\mathrm{d}x)$. The MCMC estimator of $\pi(h)$ is the ergodic average $t^{-1}\sum_{s=0}^{t-1} h(X_s)$, possibly after discarding an initial portion of the trajectory. The initialization bias is defined as $\mathbb{E}[t^{-1}\sum_{s=0}^{t-1} h(X_s)] - \pi(h)$. It is only zero if π_0 is precisely π; most often it is considered unknown. The bias vanishes as $t \to \infty$ and becomes a negligible part of the mean squared error, which is dominated by the variance $v(P, h)$ in the Central Limit Theorem (CLT):

$$\sqrt{t}\left(\frac{1}{t}\sum_{s=0}^{t-1} h(X_s) - \pi(h)\right) \xrightarrow{d} \text{Normal}(0, v(P, h)), \qquad \text{as } t \to \infty. \tag{1.1.2}$$

There may be other sources of bias in MCMC, such as the use of pseudo-random rather than random numbers, limited floating point precision, as well as various deliberate approximations that can accelerate computation. This chapter focuses on initialization bias.

Despite its asymptotic disappearance, the initialization bias poses practical issues. The bias is an obstacle to the parallelization of MCMC computation (Rosenthal, 2000). Indeed, users can generate short MCMC runs independently in parallel, but the bias prevents the consistent estimation of $\pi(h)$ by averages over the independent runs. The bias can be reduced

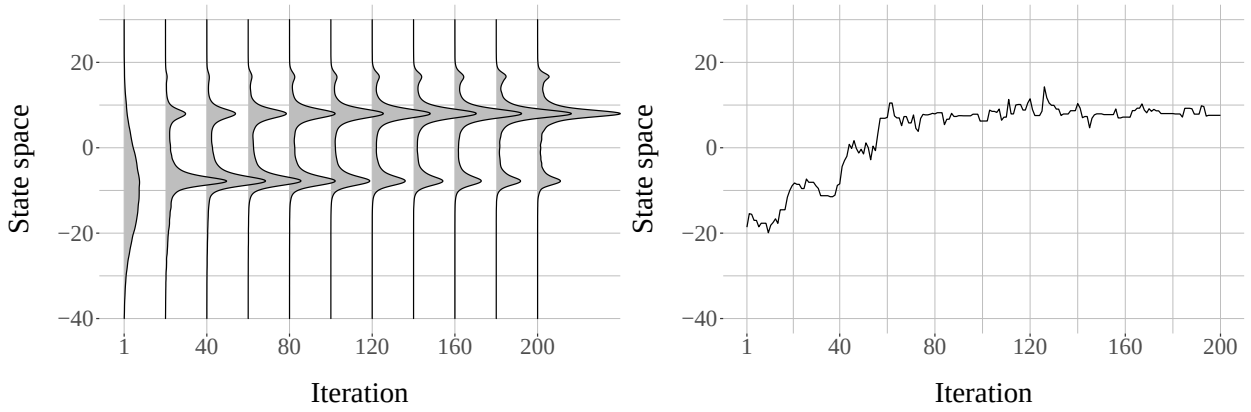

Figure 1.1: Convergence of the marginal distribution π_t to π (left) and a realized trajectory (right), here generated by a Metropolis–Rosenbluth–Teller–Hastings (MRTH) algorithm with Normal random walk proposals. The target distribution is described in Example 3.1 of Robert (1995). This setting is used for all illustrations of this chapter.

by discarding a larger initial portion of each parallel run, but the choice of the length to discard is both difficult and critical. This is in part because a sufficient length for the bias to be small is vastly different from one application to the next; numbers of iterations reported in the literature span many orders of magnitude, e.g. Metropolis et al. (1953) perform a few dozen sweeps whereas McCartan and Imai (2020) mention a run of a trillion ($10^{12}\ldots!$) MCMC iterations for the task of sampling redistricting plans.

1.1.2 The promise of unbiased MCMC

The term *unbiased MCMC* (Jacob et al., 2020b) refers to the removal of the initialization bias. The key requirement of these methods is that the user can generate certain *successful couplings* of Markov chains (see Section 1.1.3). This requirement is weaker than that of Coupling From The Past (Propp and Wilson, 1996), and thus unbiased MCMC is more widely applicable; but it does not provide perfect samples from π. Instead, these methods generate unbiased approximations of π in the form of signed empirical measures:

$$\hat{\pi} = \sum_{n=1}^{N} \omega_n \delta_{Z_n}, \tag{1.1.3}$$

where N is a random integer, $(Z_n)_{n=1}^N$ are atoms on the state space $\mathbb{X}$, and $(\omega_n)_{n=1}^N$ are real-valued weights (see Section 1.2.4 for the precise construction). The lack-of-bias property of $\hat{\pi}$ means that, for a class of functions h, $\hat{\pi}(h) := \sum_{n=1}^N \omega_n h(Z_n)$ has expectation exactly equal to $\pi(h)$. The significance of the lack of bias is that users can generate independent copies of $\hat{\pi}(h)$ in parallel and average over the copies to obtain consistent estimators of $\pi(h)$ as well as confidence intervals, provided that the estimators have a finite variance. The lack of bias is thus clearly appealing as a means toward a parallel-friendly consistent Monte Carlo scheme; there are other appeals discussed in Section 1.3.

With unbiased signed measures, the question of initialization bias seems to be resolved. This convenience comes at a price: both the computing time and the variance can be prohibitively high. To quantify this price, we can compute the *inefficiency*, a key descriptor of asymptotic performance for unbiased estimators (Glynn and Whitt, 1992), defined as the expected computing time multiplied by the variance. Both terms can be estimated from independent runs, and the inefficiency can be compared with the asymptotic variance in the CLT for standard MCMC averages (1.1.2). The first unbiased estimators constructed from coupled chains, proposed in Glynn and Rhee (2014) and applied to MCMC settings in Agapiou et al. (2018), were found to be either competitive or not relative to regular MCMC, depending on the chain and on the coupling. The simple enhancements proposed in Jacob et al. (2020a,b); Vanetti and Doucet (2020) (see Section 1.2.3) led to unbiased MCMC estimators that are both provably and practically competitive with regular MCMC estimators, whenever they are applicable.

1.1.3 Successful couplings of Markov chains

Unbiased MCMC belongs to a family of algorithms that require couplings of Markov chains, as in Coupling From The Past (CFTP, Propp and Wilson, 1996), circularly-coupled MCMC (Neal, 1999), and certain convergence diagnostics (Johnson, 1996, 1998). A *coupling* of two distributions p and q on $\mathbb{X}$ refers to a joint distribution on $\mathbb{X} \times \mathbb{X}$, with prescribed marginals p and q. For Markov chains, a coupling refers to a joint process (X_t, Y_t) such that (X_t) and (Y_t) are individually identical to prescribed Markov chains. For simplicity, we focus on

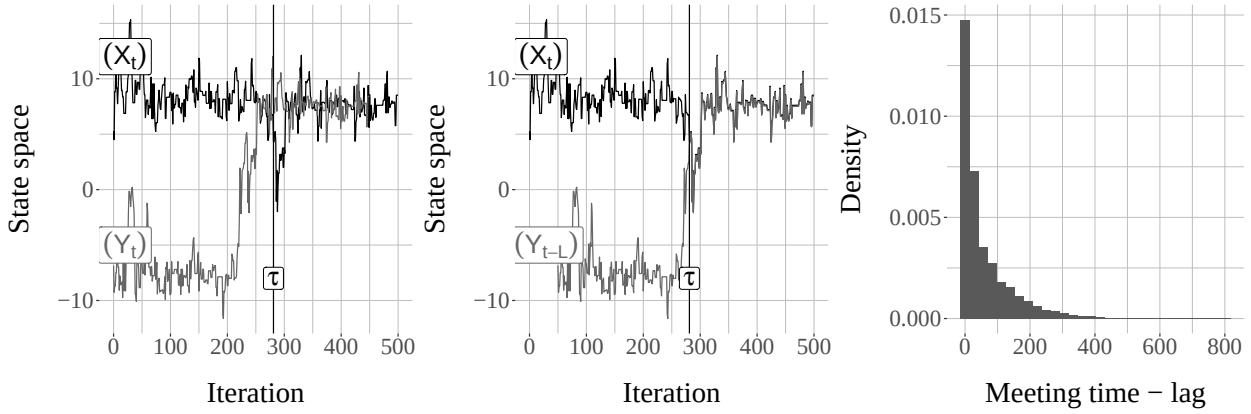

Figure 1.2: Successful coupling of Markov chains, meeting at time τ with a lag L so that $X_t = Y_{t-L}$ for all $t \geq \tau$. Left: trajectory of (X_t, Y_t). Middle: trajectory of (X_t, Y_{t-L}). Right: histogram of 10^4 independent copies of $\tau - L$.

Markovian couplings, where (X_t, Y_t) is itself a Markov chain, with initial distribution $\bar{\pi}_0$ on $\mathbb{X} \times \mathbb{X}$ and Markov transition $\bar{P}$.

Unbiased MCMC requires draws of (X_t, Y_t) such that both chains (X_t) and (Y_t) are copies of the same Markov process, with initial distribution π_0 and transition P. Thus, $\bar{\pi}_0$ should be a coupling of π_0 with itself, and $\bar{P}$ a coupling of P with itself. Furthermore, for each trajectory of (X_t, Y_t) we require the existence of a finite *meeting time* τ such that $X_t = Y_{t-L}$ for all $t \geq \tau$, where L is a user-chosen *lag* parameter. Coupling resulting in finite meeting times can be called *successful*, following Pitman (1976). The construction, summarized in Algorithm 1, does not assume anything about the initialization π_0, does not require any form of monotonicity in the coupling, and does not require the existence and identification of a pair of extremal states, as in most practical instances of CFTP. An illustration of a successful coupling is shown in Figure 1.2, where $L = 50$ and $\ell = 500$.

Algorithm 1 Successful coupling of Markov chains with lag L and length ℓ. Initial distribution: π_0, transition kernel: P, coupled transition kernel: $\bar{P}$. The meeting time is $\tau = \inf\{t \geq L : X_t = Y_{t-L}\}$.

1. Sample (X_0, Y_0) from $\bar{\pi}_0$, .
2. If $L \geq 1$, for $t = 1, \ldots, L$, sample X_t from $P(X_{t-1}, \cdot)$.
3. For $t \geq L$, sample (X_{t+1}, Y_{t-L+1}) from $\bar{P}((X_t, Y_{t-L}), \cdot)$ until $X_{t+1} = Y_{t-L+1}$ and $t+1 \geq \ell$.

The key assumption required for unbiased MCMC is about the tails of the meeting times. We formulate the following assumption as in Douc et al. (2023), and it is equivalent to the tails $\mathbb{P}(\tau > t)$ decaying at the polynomial rate $t^{-\kappa}$ as $t \to \infty$. We emphasize that unbiased MCMC methods do not require the user to specify a value for κ.

Assumption 1. *There exists $\kappa \geq 1$ such that, if the two chains (X_t) and (Y_t) started independently from π and evolved according to the coupled transition $\bar{P}$, the meeting time $\tau = \inf\{t \geq 1 : X_t = Y_t\}$ would have κ finite moments: $\mathbb{E}[\tau^\kappa] < \infty$.*

Assumption 1 can be verified for example if the transition P satisfies a Lyapunov drift condition with function V, combined with a non-zero probability of meeting over one step when the chains are simultaneously in a level set of V (Section 3.2 in Jacob et al., 2020b). Explicit Lyapunov functions have been elicited for many MCMC algorithms, such as random walk Metropolis–Rosenbluth–Teller–Hastings, abbreviated MRTH (see Roberts and Tweedie, 1996), Langevin Monte Carlo (Durmus and Moulines, 2022), Hamiltonian Monte Carlo (Durmus et al., 2017) and many examples of Gibbs samplers. The reasoning extends to polynomial drift conditions (Section 1.4 in Middleton et al., 2020). When applicable the above approach is sufficient to establish that Assumption 1 holds for some or all $\kappa \geq 1$, without providing insight on how the meeting time behaves as a function of salient features of the problem.

As a concrete example, a simple random walk MRTH algorithm is implemented in `R` in Figure 1.3. It employs a coupling of Normal proposal distributions presented in Section 2.3 of Bou-Rabee et al. (2020), and in Section 1.4.1 below. This coupling was employed to generate all figures in this chapter. Assumption 1 can then be verified for all $\kappa > 1$ via Proposition 4 in Jacob et al. (2020b) using the geometric drift function $V(x) = \pi(x)^{-1/2}$ under the conditions of Theorem 3.2 in Roberts and Tweedie (1996). Section 1.4 provides more discussion on the design of successful couplings of MCMC algorithms.

```r
# Metropolis-Rosenbluth-Teller-Hastings transition with Normal proposals
mrth = function(x, U, sigma)
{
  # proposal = current location + Normal(0,sigma^2)
  xprop = x + sigma * rnorm(length(x))
  # log Uniform to accept/reject proposals
  logu = log(runif(1))
  # return state according to decision to accept or not
  return(if (logu < (U(x) - U(xprop))) xprop else x)
}
# coupling of MRTH transition with Normal proposals
coupledmrth = function(x, y, U, sigma)
{
  # draw proposals using maximal coupling of Bou-Rabee, Eberle and Zimmer (AAP 2020)
  xstd = rnorm(length(x)) # standard Normal variables
  z = (x - y) / sigma # length(sigma) could be 1 or length(x)
  e = z / sqrt(sum(z^2)) # normalise
  logu = log(runif(1))
  sameprop = (logu < sum(dnorm(xstd + z, log = TRUE) - dnorm(xstd, log = TRUE)))
  ystd = if (sameprop) xstd + z else xstd - 2 * sum(e * xstd) * e
  # xprop is marginally Normal(x,sigma^2)
  xprop = x + sigma * xstd
  # yprop is marginally Normal(y,sigma^2)
  yprop = y + sigma * ystd
  # log Uniform to accept/reject proposals
  logu = log(runif(1))
  # decision to accept or not
  xaccept = (logu < (U(x) - U(xprop)))
  yaccept = (logu < (U(y) - U(yprop)))
  # return state according to decision
  return(list(nextx = if (xaccept) xprop else x,
              nexty = if (yaccept) yprop else y,
              nextxequalsnexty = sameprop && xaccept && yaccept))
}
```

Figure 1.3: `R` code for MRTH algorithm with Normal random walk proposals, and a coupling of it. This defines the transition P and coupled transition $\bar{P}$ required by Algorithm 1. Inputs: current states `x` and `y`, a potential function `U` corresponding to $x \mapsto -\log \pi(x)$, proposal standard deviation `sigma` (a scalar or a vector of the same length as `x` and `y`).

1.2 Unbiased MCMC

This section presents unbiased MCMC estimators, assuming that successful couplings can be implemented. We start with classical bias removal techniques in Section 1.2.1, re-derive a simple unbiased MCMC estimator via the Poisson equation in Section 1.2.2, and present more efficient versions in Sections 1.2.3-1.2.4. In Section 1.2.5 we comment on performance and cost, and propose tuning strategies.

1.2.1 Bias removal with a telescope

Randomized telescoping sums. Consider a quantity of interest expressed as the limit as $k \to \infty$ of a deterministic sequence $(b_k)_{k\geq 0}$ that can be computed recursively. We can write the limit as a telescoping series $\sum_{k=0}^{\infty}(b_k - b_{k-1})$, where $b_{-1} = 0$. How can we estimate a series $\sum_{k=0}^{\infty} a_k$ without bias? The following reasoning dates back to at least Glynn (1983); Rychlik (1990). Let ξ be a random variable on $\{0, 1, 2 \ldots\}$ with $p_k := \mathbb{P}(\xi = k)$ for all $k \geq 0$; ξ is the *random truncation variable*. Then sample ξ and compute: $G = a_\xi / p_\xi$. If the expectation of G is finite then it is equal to $\sum_{k=0}^{\infty} a_k$, and its expected cost is $\mathbb{E}[\xi] = \sum_{k\geq 0} k p_k$. The cost is smaller if (p_k) decay faster. However, the variance of G involves $\mathbb{E}[G^2] = \sum_{k\geq 0} a_k^2 / p_k$ which is smaller if (p_k) decay slower. An alternative is to sample ξ and then compute $H = a_0 + \sum_{k=1}^{\xi} a_k / \mathbb{P}(\xi \geq k)$. The estimator H also has expectation $\sum_{k=0}^{\infty} a_k$, its cost is similar to that of G, but its variance is finite under weaker conditions than that of G. The estimators G and H are termed *single term* and *coupled sum* in Rhee and Glynn (2015); Vihola (2018).

Bias removal in MCMC. Direct use of the above strategy to remove the bias of MCMC averages, where $b_k = k^{-1} \sum_{s=0}^{k-1} h(X_s)$, is considered in McLeish (2011). An immediate difficulty is that ergodic averages converge at the (slow) Monte Carlo rate, resulting in unbiased estimators that tend to have either a large cost or a large variance. The convergence of marginal distributions $\pi_t \to \pi$ e.g. in total variation is comparably faster. Using contractive couplings of Markov chains, Glynn and Rhee (2014) propose a debiasing strategy, described below, that benefits from the fast convergence of marginal distributions. Agapiou et al. (2018) explore that strategy in MCMC settings and highlight the practical difficulties associated with the specification of the truncation variable. Jacob et al. (2020a) find that the conditional particle filter, which is an MCMC algorithm developed specifically for latent variable estimation in continuous state space models (Andrieu et al., 2010), could be easily coupled such that a pair of chains would meet, and that this removes the need for truncation variables in the construction of Glynn and Rhee (2014). Jacob et al. (2020b) find that many MCMC algorithms can be coupled successfully, building on works such as Johnson (1998).

A first unbiased MCMC estimator. The idea of Glynn and Rhee (2014) in the

context of successful couplings goes as follows. Write $\pi(h)$ as a telescopic sum, for all $k \geq 0$, for any choice of lag L,

$$\pi(h) = \lim_{t\to\infty} \pi_t(h) = \pi_k(h) + \sum_{j=1}^{\infty} \pi_{k+jL}(h) - \pi_{k+(j-1)L}(h). \tag{1.2.1}$$

Since for all $t \geq 0$, X_t and Y_t have the same distribution π_t, we can write $\pi_{k+jL}(h) = \mathbb{E}[h(X_{k+jL})]$ and $\pi_{k+(j-1)L}(h) = \mathbb{E}[h(Y_{k+(j-1)L})]$. A bold swap of expectation and limit suggests that the random variable defined as

$$H_k := h(X_k) + \sum_{j=1}^{\infty} (h(X_{k+jL}) - h(Y_{k+(j-1)L})), \tag{1.2.2}$$

is an unbiased estimator of $\pi(h)$. For instance, if $|h|$ is bounded by 1, then $\sum_{j=1}^{\infty} |h(X_{k+jL}) - h(Y_{k+(j-1)L})| \leq 2\max(0, (\tau - k)/L)$. Therefore, if Assumption 1 holds with $\kappa = 1$, that is if $\mathbb{E}[\tau] < \infty$, then by Fubini's theorem H_k indeed satisfies $\mathbb{E}[H_k] = \pi(h)$. Higher moments of H_k can also be controlled as we discuss below.

The infinite sum can be computed in finite time since the lagged chains meet at a finite time τ: the differences $h(X_{k+jL}) - h(Y_{k+(j-1)L})$ are equal to zero for all k, j, L such that $k + jL \geq \tau$. The estimator (1.2.2) can be computed without specifying truncation probabilities thanks to the stopping criterion offered by the meeting time.

1.2.2 Alternative construction via the Poisson equation

Poisson equation and bias. Douc et al. (2023) provide an alternative derivation of H_k in (1.2.2) via the Poisson equation. Write $Pf(x) = \int P(x, dx')f(x')$ for a function $f : \mathbb{X} \to \mathbb{R}$. A function g in $L^1(\pi)$ is a solution of the Poisson equation associated with h and P if

$$g(x) - Pg(x) = h(x) - \pi(h) \quad \forall x \in \mathbb{X}. \tag{1.2.3}$$

For example, the function

$$g_\star : x \mapsto \sum_{t=0}^{\infty} P^t \left\{h - \pi(h)\right\}(x), \tag{1.2.4}$$

is a solution to (1.2.3); see Chapter 21 of Douc et al. (2018). In fact, all solutions are equal to $g_\star$ up to an additive constant. It is known that (1.2.4) is related to MCMC bias. Indeed, consider the ergodic average $t^{-1}\sum_{s=0}^{t-1} h(X_s)$ when the chain starts from a fixed $x_0 \in \mathbb{X}$. The bias is $\mathbb{E}_{x_0}[t^{-1}\sum_{s=0}^{t-1} h(X_s)] - \pi(h)$, and if we multiply by t and consider the limit $t \to \infty$,

$$\lim_{t\to\infty} t \times \{\mathbb{E}_{x_0}[t^{-1}\sum_{s=0}^{t-1} h(X_s)] - \pi(h)\} = \lim_{t\to\infty} \sum_{s=0}^{t-1} \mathbb{E}_{x_0}[h(X_s) - \pi(h)] = g_\star(x_0), \tag{1.2.5}$$

with $g_\star$ as in (1.2.4) (Kontoyiannis and Dellaportas, 2009).

Estimation of g and unbiased MCMC. It turns out that we can estimate solutions of the Poisson equation using successful chains, which then leads to unbiased estimators of $\pi(h)$. Consider the function $x \mapsto g(x, y) := g_\star(x) - g_\star(y)$, equal to $g_\star$ up to the constant $g_\star(y)$, for any fixed $y \in \mathbb{X}$, and thus solution of the Poisson equation. It can be written $g(x, y) = \sum_{t\geq 0}\{P^t h(x) - P^t h(y)\}$. Consider successful chains with no lag ($L = 0$), where X_0 is set to x and Y_0 is set to y. Then $h(X_t)$ and $h(Y_t)$ have expectation equal to $P^t h(x)$ and $P^t h(y)$ for all $t \geq 0$, but for t larger than $\inf\{t \geq 1 : X_t = Y_t\}$ we have $h(X_t) - h(Y_t) = 0$. This suggests the following, implementable estimator of $g(x, y)$:

$$G(x, y) := \sum_{t=0}^{\tau-1} \{h(X_t) - h(Y_t)\}, \tag{1.2.6}$$

where $X_0 = x$, and $Y_0 = y$, $\tau = \inf\{t \geq 1 : X_t = Y_t\}$. With the ability to estimate solutions of the Poisson equation we might envision the estimation of $\pi(h)$ via the re-arranged equation: $\pi(h) = h(x) + Pg(x) - g(x)$. Setting $x \in \mathbb{X}$ arbitrarily, sample $X_1 \sim P(x, \cdot)$ (performing one step of MCMC), and sample $G(X_1, x)$ (running two chains, initialized at X_1 and x, until they meet). Then $\mathbb{E}_x[G(X_1, x)] = Pg_\star(x) - g_\star(x)$, therefore $h(x) + G_x(X_1)$ is an unbiased estimator of $\pi(h)$. It is in fact exactly H_k in (1.2.2) with $k = 0$, $L = 1$ and $\pi_0 = \delta_x$.

1.2.3 Unbiased MCMC estimators

Improved efficiency by averaging. Simple modifications of (1.2.2) can go a long way to improve its efficiency, as noted in Jacob et al. (2020a) and Jacob et al. (2020b, and its discussion). Consider a run of Algorithm 1 with lag $L \geq 1$ (Vanetti and Doucet, 2020) and length $\ell \geq 0$, from which we can construct unbiased estimators $H_k, \ldots, H_\ell$ as in (1.2.2) for a range of integers $k, \ldots, \ell$ where $0 \leq k \leq \ell$. Since these estimators $(H_t)_{t=k}^{\ell}$ are unbiased, their average is unbiased as well. After some algebraic manipulations, the average estimator reads

$$H_{k:\ell} = \underbrace{\frac{1}{\ell - k + 1} \sum_{t=k}^{\ell} h(X_t)}_{\text{MCMC}} + \underbrace{\sum_{t=k+L}^{\tau-1} v_t(k, \ell, L) \left\{ h(X_t) - h(Y_{t-L}) \right\}}_{\text{bias cancellation}}. \tag{1.2.7}$$

The first term on the right-hand side is the regular MCMC ergodic average, computed from the trajectory $(X_k, \ldots, X_\ell)$. The second term performs bias cancellation from weighted differences between the chains. The weight $v_t(k, \ell, L)$ is defined as the number of appearances of the difference $h(X_t) - h(Y_{t-L})$ in the bias cancellation terms of $H_k, \ldots, H_\ell$, divided by $\ell - k + 1$, and can be computed as

$$v_t(k, \ell, L) = \frac{\lfloor (t-k)/L \rfloor - \lceil \max(L, t-\ell)/L \rceil + 1}{\ell - k + 1}. \tag{1.2.8}$$

Both the number of terms in the bias cancellation and their weights can be reduced by increasing the tuning parameters k, ℓ, L. Section 1.2.5 provides guidance on tuning unbiased MCMC such that its efficiency becomes comparable to that of regular MCMC.

Variance reduction. The estimator in (1.2.7) is presented with a generic lag L following the observation of Vanetti and Doucet (2020) that increasing L can lead to significant variance reduction. Control variates for $H_{k:\ell}$ are proposed in Craiu and Meng (2022). They observe that $\mathbb{E}[h(X_t) - h(Y_t)] = 0$ for all $t \geq 0$, thus $\sum_{t \geq 0} \eta_t \{h(X_t) - h(Y_t)\}$ can be added to $H_{k:\ell}$, for any real sequence (η_t), without modifying its expectation. Even imperfect optimization over the sequence (η_t) can lead to a worthwhile reduction in variance. Other coupling-based variance reduction strategies have been proposed for MCMC (Pinto and Neal, 2001), and could be considered also for unbiased MCMC.

Finite moments. The following result is taken from Douc et al. (2023).

Theorem 1.2.1. *Under Assumption 1 with $\kappa > 1$, i.e. $\mathbb{E}[\tau^\kappa] < \infty$, let $h \in L^m(\pi)$ for some $m > \kappa/(\kappa - 1)$. Assume that π_0 is such that $d\pi_0/d\pi \leq M$ with $M < \infty$. Then for any k, ℓ, L, the estimator $H_{k:\ell}$ in (1.2.7) satisfies $\mathbb{E}[|H_{k:\ell}|^p] < \infty$ for $p \geq 1$ such that $\frac{1}{p} > \frac{1}{m} + \frac{1}{\kappa}$.*

Finiteness of the variance of $H_{k:\ell}$ is sufficient to validate the following classical construction of confidence intervals: generate C independent copies of $H_{k:\ell}$, compute their average $\hat{\mu}$ and their standard deviation $\hat{\sigma}$, and an asymptotically (as $C \to \infty$) valid confidence interval for $\pi(h)$ is given by $[\hat{\mu} + q_{\alpha/2}\hat{\sigma}/\sqrt{C}, \hat{\mu} + q_{1-\alpha/2}\hat{\sigma}/\sqrt{C}]$, where q_s is the s-th quantile of the standard Normal distribution. According to Theorem 1.2.1 finiteness of second moments ($p = 2$) results from a mild condition on π_0, and the assumption that $\kappa > 2$, for all $h \in L^m(\pi)$ such that $m > 2\kappa/(\kappa - 2)$. For geometrically ergodic Markov chains where κ can be taken arbitrary large, the condition becomes arbitrarily close to $h \in L^2(\pi)$. The assumption that $d\pi_0/d\pi \leq M$ is satisfied for example if π_0 is supported entirely in a bounded set included in the support of π. A similar result holds if the initial distribution π_0 is a Dirac mass.

1.2.4 Unbiased signed measure

Replacing function evaluations by Dirac masses. The empirical measure

$$\hat{\pi}(\mathrm{d}x) = \underbrace{\frac{1}{\ell - k + 1} \sum_{t=k}^{\ell} \delta_{X_t}(\mathrm{d}x)}_{\text{MCMC}} + \underbrace{\sum_{t=k+L}^{\tau-1} v_t(k, \ell, L) \left\{ \delta_{X_t} - \delta_{Y_{t-L}} \right\} (\mathrm{d}x)}_{\text{bias cancellation}}, \tag{1.2.9}$$

is an unbiased approximation of π, where v_t is defined in (1.2.8). This is of the form $\sum_{n=1}^{N} \omega_n \delta_{Z_n}$ as in (1.1.3), with $N = \max(0, \tau - (k+L)) + (\ell - k + 1)$, Z_n are states from either (X_t) or (Y_t) and ω_n are either $(\ell - k + 1)^{-1}$, or of the form $\pm v_n(k, \ell, L)$; in particular the weights can be negative. Figure 1.4 represents the unbiased MCMC approximation, made of MCMC and bias cancellation components. This was obtained by kernel density estimation from the weighted samples constituting the different elements in (1.2.9).

Sub-sampling and negative weights. We can sub-sample from the empirical measure

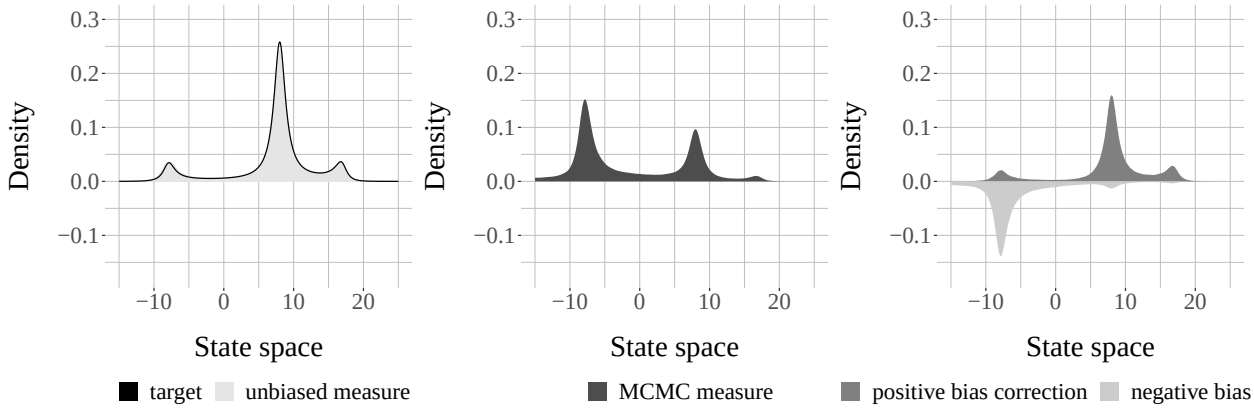

Figure 1.4: Unbiased MCMC (left) = MCMC (middle) + bias cancellation (right). The target density is the black curve on the left-most plot. On the right, the bias cancellation is made of a positive measure (darker grey) added to a negative measure (lighter grey).

in (1.1.3). For example, we can draw an index I uniformly in $\{1, \ldots, N\}$ and return the sample Z_I with weight $N\omega_I$. Then for a class of functions h, $N\omega_I h(Z_I)$ will have expectation equal to $\pi(h)$ (Douc et al., 2023). We can also sample the index I non-uniformly, with probabilities $\xi_1, \ldots, \xi_N$ that depend on the atoms in (1.1.3), and the selected atom Z_I is then weighted by $\xi_I^{-1}\omega_I$, and we may repeat this selection multiple times to obtain a weighted sub-sample from (1.1.3) with a desired size. Yet this does not produce a perfect sample due to the weights being possibly negative. We can arbitrarily decrease the proportion of negative weights in (1.1.3) by increasing the value of k, but we cannot make it zero. As a result, unbiased MCMC estimators $H_{k:\ell}$ can take values outside the range of the function h, e.g. we may obtain negative estimates of positive quantities. There may not be any general solution to this problem: according to Lemma 2.1 in Jacob and Thiery (2015) there is no algorithm that takes unbiased estimators of a nonnegative quantity as input (and nothing else), and returns nonnegative unbiased estimators of that same quantity.

1.2.5 Efficiency, cost and tuning

Asymptotic equivalence with MCMC. Theorem 1.2.1 validates unbiased MCMC for the estimation of $\pi(h)$ but does not help for the comparison of its performance with standard MCMC, or choosing the three tuning parameters k, ℓ, L. The guiding principle in the

tuning of k, ℓ, L is that a judicious choice will make unbiased MCMC competitive with standard MCMC in terms of cost and variance. Proposition 3 in Jacob et al. (2020b) provides conditions under which the increase of either k or $\ell - k$ results in variance reduction, and in particular the variance of $H_{k:\ell}$ is shown to be asymptotically equivalent to the variance of standard MCMC estimators as $\ell \to \infty$. The result is shown under weaker conditions in Middleton et al. (2020). In the same spirit Douc et al. (2023) provide the following CLT for $H_{k:\ell}$ as $\ell \to \infty$, where the asymptotic variance is the same as for standard MCMC.

Theorem 1.2.2. *Under Assumption 1 with $\kappa > 1$, let $h \in L^m(\pi)$ for some $m > 2\kappa/(\kappa - 1)$. Then for any $k \in \mathbb{N}$,*

$$\sqrt{\ell - k + 1}\,(H_{k:\ell} - \pi(h)) \xrightarrow{d} \mathit{Normal}(0, v(P, h)), \tag{1.2.10}$$

as $\ell \to \infty$, where $v(P, h)$ is the asymptotic variance in the CLT for MCMC averages (1.1.2).

The asymptotic equivalence with regular MCMC as $\ell \to \infty$ should be expected since the initialization bias vanishes as $\ell \to \infty$, and given the form of the bias cancellation term in (1.2.7): the sum is over $\max(0, \tau - L - k)$ terms (irrespective of ℓ) while the weights in (1.2.8) decrease as $(\ell - k)^{-1}$. Thus the bias cancellation term disappears when $\ell - k$ increases. Regarding cost: if we count the cost of sampling from the MCMC transition P as one unit, and the cost of sampling from the coupled transition $\bar{P}$ as one unit if the chains have already met and two units if they have not, then the random cost of running Algorithm 1 with parameters ℓ and L to compute $H_{k:\ell}$ equals $\max(L, \ell - (\tau - L)) + 2(\tau - L)$ units. This behaves as ℓ when $\ell \to \infty$. Thus both cost and variance of $H_{k:\ell}$ are equivalent to those of MCMC as $\ell \to \infty$. By carefully choosing k, ℓ, L we can hope to obtain unbiased MCMC estimators with an efficiency close to that of MCMC.

Cost and parallel computing. Some users might prefer less efficient but cheaper estimators when enough parallel machines are available to produce them. Consider the task of obtaining C estimates using M parallel machines. When M is much smaller than C, each machine produces many estimates and the computing times even out across machines so that the speed-up is close to linear in M. On the other hand, if $M \geq C$, then each

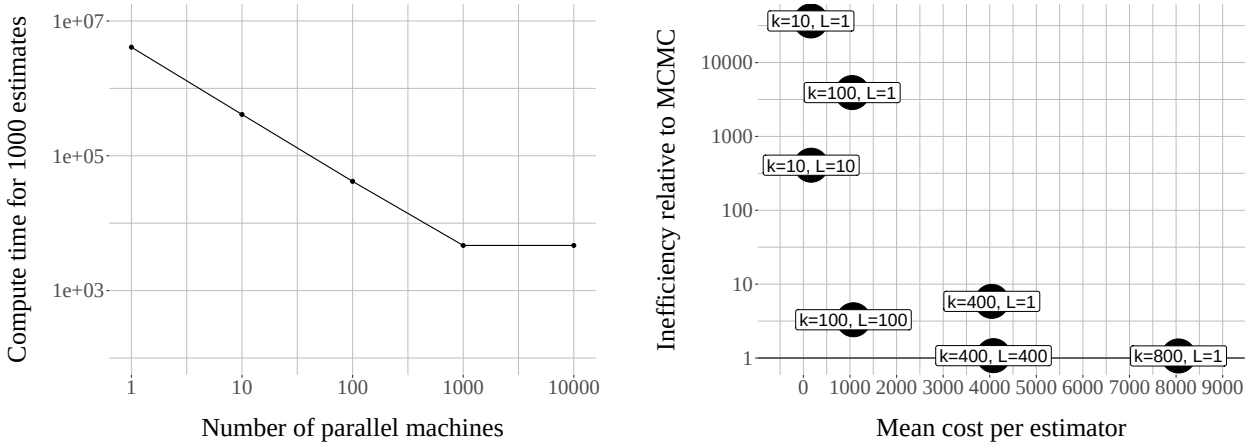

Figure 1.5: Left: compute time to generate 1000 independent copies of $H_{k:\ell}$, with $k = L = 400, \ell = 10k$, using parallel machines. Right: inefficiency relative to standard MCMC versus average cost of $H_{k:\ell}$, for different choices of k, and L set to either 1 or k, always with $\ell = 10k$. The configuration $k = 800, L = 800$ is not shown as it would be overlaid with $k = 800, L = 1$.

machine produces one estimate, and there is no speed-up in increasing M further. Careful: running unbiased MCMC on $M \gg C$ machines and retaining the C estimators that are first completed would introduce a bias, since the estimator is not independent of its cost. Figure 1.5 (left) illustrates the speed-up associated with an increasing number of machines, for the task of producing 1000 estimates. If each machine produces a single estimate, the total time is driven by the longest run, which is in average an increasing function of C. For example if τ has Geometric tails, the average maximum cost of unbiased MCMC behaves as $\log(C)$. Handling of budget constraints, such as hard or soft deadlines, on parallel machines is discussed in Glynn and Heidelberger (1990, 1991).

Choice of length ℓ. We proceed to proposing guidance for the tuning parameters. First we simplify the choice by recommending that ℓ is set as a large multiple of k, for example $\ell = 10k$. This is because a portion k/ℓ of iterations is simply discarded in the construction of (1.2.7), and we would like to limit this apparent waste. Thus, we are left with the choice of k and L; increasing k will automatically increase ℓ and $\ell - k$, thus decreasing the magnitude of the weights v_t in the bias cancellation term.

Choice of burn-in k and lag L. The bias cancellation is exactly zero in the event $\{\tau - L < k\}$. By setting k as a large quantile of $\tau - L$, we ensure that the event occurs

with high probability. The increase of the lag L, compared to the choice $L = 1$ in Glynn and Rhee (2014); Jacob et al. (2020b), is advocated in Vanetti and Doucet (2020). From the expression of the weights in (1.2.8), increasing L decreases the weights in the bias cancellation term, and thus brings $H_{k:\ell}$ closer to regular MCMC. Furthermore, setting $L = k$ leads to a minor increase of cost per estimator compared to $L = 1$, and thus the efficiency is typically improved, sometimes drastically.

Concrete guideline. In our experience, satisfactory tuning can be done as follows. First generate 1000 independent meeting times with lag $L = 1$ (by lack of a better guess). Then set k as a large (e.g. 99%) quantile of $\tau - L$, which is the number of calls to $\bar{P}$ before observing the meeting $X_\tau = Y_{\tau-L}$. Finally, redefine $L := k$, and set $\ell = 10k$. Figure 1.5 (right) shows how different choices of k, L (always with $\ell = 10k$) lead to vastly different costs and efficiencies, for the estimation of $\pi(h)$ with $h : x \mapsto x$. In the figure, inefficiency is divided by the asymptotic variance $v(P, h)$ of the MCMC estimate, estimated using the method of Section 1.3.3. The relative inefficiency goes to one with increasing k, and setting $L := k$ instead of $L := 1$ is often worthwhile.

1.3 Beyond the estimation of stationary expectations

Unbiased MCMC provides estimators of stationary expectations $\pi(h)$, and also enables other methods that can be of interest for MCMC users, including non-asymptotic convergence diagnostics (Section 1.3.1), estimators of nested expectations (Section 1.3.2) and asymptotic variances (Section 1.3.3).

1.3.1 Convergence diagnostics

Upper bounds on the distance to stationarity. As a by-product of the unbiased estimator in (1.2.2) we can construct upper bounds on the total variation distance $|\pi_k - \pi|_{\mathrm{TV}}$, for any finite k, that can be estimated from samples of meeting times, as first proposed in Section 6 of Jacob et al. (2020b), and improved with the use of $L > 1$ in Biswas et al. (2019),

and with control variates in Craiu and Meng (2022). A simple way of deriving such bound is to write $|\pi_k - \pi|_{\mathrm{TV}} = \frac{1}{2}\sup_{|h|\leq 1} |\pi(h) - \pi_k(h)|$. Since H_k in (1.2.2) is unbiased, we can replace $\pi(h) - \pi_k(h)$ by $\mathbb{E}[\sum_{j=1}^{\infty}(h(X_{k+jL}) - h(Y_{k+(j-1)L}))]$, where the terms in the sum are zero if $k + jL \geq \tau$. There remains $\max(0, \lceil(\tau - L - k)/L\rceil)$ non-zero terms, corresponding to indices $j \geq 1$ such that $k + jL < \tau$. Finally, each non-zero term of the form $h(x) - h(y)$ can be upper-bounded by 2 if h is such that $\sup_x |h(x)| \leq 1$. Thus, we obtain

$$\forall k \geq 0 \quad |\pi_k - \pi|_{\mathrm{TV}} \leq \mathbb{E}[\max(0, \lceil(\tau - L - k)/L\rceil)]. \tag{1.3.1}$$

The right-hand side can be estimated by replacing the expectation by an empirical average over C independent meeting times $\tau_1, \ldots, \tau_C$, for any value of k. This is obtained by running Algorithm 1 with lag L and $\ell = 0$, C times independently. The empirical upper bound $C^{-1}\sum_{c=1}^{C} \max(0, \lceil(\tau_c - L - k)/L\rceil)$ is exactly zero for all $k \geq \max_c \tau_c - L$, so it is enough to evaluate it at integers k less than $\max_c \tau_c - L$. Figure 1.6 shows these bounds obtained for different lags L. Increasing the lag yields lower bounds, but with diminishing returns; as $L \to \infty$ the bounds are still not sharp, as they depend on the coupling employed. In practice, we can first generate meeting times with $L = 1$, and then redefine L as a large (e.g. 99%) empirical quantile of $\tau - L$. A similar reasoning lead to upper bounds on other distances than total variation: Biswas et al. (2019) consider 1-Wasserstein bounds and Papp and Sherlock (2022a) general W_p bounds.

Practical significance. We emphasize the convenience of (1.3.1) compared to the usual bounds encountered in the literature on Markov chains. In continuous state spaces, the coupling inequality due to Wolfgang Doeblin (Lindvall, 2002) reads: $|\pi_k - \pi|_{\mathrm{TV}} \leq \mathbb{P}_{\pi_0 \otimes \pi}(\tau > k)$ for all $k \geq 0$, where the meeting time corresponds to a pair of chains (without any lag), started from π_0 and π, but by definition MCMC users can rarely sample from π. In discrete state spaces, one can also write $\max_x |P^k(x, \cdot) - \pi|_{\mathrm{TV}} \leq \max_{x,y} \mathbb{P}_{x,y}(\tau > k)$ where the meeting time corresponds to chains started from states x, y (Corollary 5.3 in Levin and Peres, 2017). Optimizing over the states x, y could be computationally difficult. In contrast, the upper bounds in (1.3.1) only involve pairs of chains started from an arbitrary π_0.

Limitation. As a warning, the following describes a situation where the use of (1.3.1)

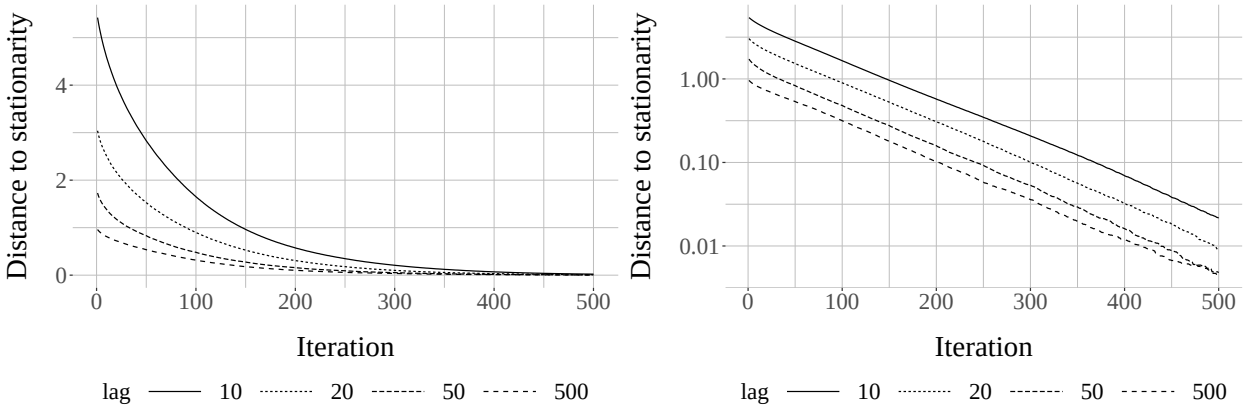

Figure 1.6: Upper bounds on $|\pi_k - \pi|_{\mathrm{TV}}$ for different times k, obtained using (1.3.1) and 10^4 independent copies of meeting times associated with different lags. The bounds are tighter with a larger lag, up to a certain point. Left: y-axis in linear scale. Right: y-axis in logarithmic scale.

would fail to provide reliable bounds on $|\pi_k - \pi|_{\mathrm{TV}}$. Suppose that the target π is multimodal, and that chains tend to get stuck in local modes. Assume further that the initial distribution π_0 puts its mass entirely in a local mode of π. The user might then observe a small empirical average of the meeting time, even after many independent runs. Yet the expectation of the meeting time could be much larger. Indeed, there could be a small probability that one chain moves to a different mode before meeting the second chain, and in that event, the meeting time could take large values, driving the expectation upward. This is illustrated in Section 5.1 of Jacob et al. (2020b). The risk is mitigated by specifying an initial distribution π_0 that is spread out relative to the modes of π, or by increasing the lag L (Biswas et al., 2019).

1.3.2 Nested expectations

Two-step target distributions. Consider expectations with respect to a joint distribution on $\mathbb{X}_1 \times \mathbb{X}_2$ defined as $\pi_{12}(x_1, x_2) = \pi_1(x_1)\pi_2(x_2|x_1)$. Suppose that $\pi_1(x_1)$ can be evaluated up to a normalizing constant Z_1, and that $\pi_2(x_2|x_1)$ can be evaluated up to a normalizing constant $Z_2(x_1)$, which is not constant with respect to x_1. The unnormalized density of $\pi_{12}(x_1, x_2)$ involves the term $Z_2(x_1)$. If $Z_2(x_1)$ cannot be evaluated, then standard MCMC algorithms such as MRTH cannot be implemented (Plummer, 2015). The setting occurs

commonly in data analysis (e.g. Blocker and Meng, 2013; Liu et al., 2009). For example x_1 could be missing values or generated regressors that act as input in a second model, in which $\pi_2(x_2|x_1)$ is the distribution of a parameter of interest x_2. Then $\pi_{12}(x_1, x_2)$ would correspond to a Bayesian version of two-step estimation, as opposed to a Bayesian analysis using a joint model of all unknown quantities treated simultaneously.

Two-step MCMC. A direct MCMC strategy to approximate $\pi_{12}(x_1, x_2)$ would start by running an MCMC algorithm with transition P_1, targeting π_1, for t_1 steps. Then, run MCMC algorithms with transition P_{2,x_1} targeting $\pi_2(\cdot|x_1)$, for a subset of the states x_1 visited by the first chain. Each second-stage run could go for $t_2(x_1)$ steps. From all of these chains one can indeed approximate $\pi_{12}(x_1, x_2)$ consistently as long as t_1 and each $t_2(x_1)$ go to infinity. Such scheme raises questions because the second stage involves transition kernels P_{2,x_1} that depend on x_1. How long should we run each of the second-stage chain? How should we construct confidence intervals for the final estimates?

Unbiased approximation to the rescue. Some of these difficulties can be bypassed to some extent with unbiased MCMC. First, obtain $\hat{\pi}_1 = \sum_{n=1}^{N} \omega_{1,n} \delta_{X_{1,n}}$, an unbiased approximation of π_1. Optionally, sub-sample the measure to reduce the number of atoms, as described in Section 1.2.4. Then obtain for each $n \in \{1, \ldots, N\}$ an unbiased approximation $\hat{\pi}_2(\cdot|X_{1,n}) = \sum_{m=1}^{M_n} \omega_{2,n,m} \delta_{X_{2,n,m}}$ of $\pi_2(\cdot|x_1 = X_{1,n})$. Under adequate assumptions on the couplings at both stages and on h, $\hat{\pi}_{12}(h) = \sum_{n=1}^{N} \omega_{1,n} \sum_{m=1}^{M_n} \omega_{2,n,m} h(X_{1,n}, X_{2,n,m})$ has expectation equal to $\int h(x_1, x_2) \pi_{12}(dx_1, dx_2)$. Averaging C independent copies of $\hat{\pi}_{12}(h)$ leads to consistent approximations as $C \to \infty$, and confidence intervals can simply be constructed from the CLT. If we use the same tuning parameters k, ℓ, L for all second-stage approximations, their inefficiencies will be uneven as x_1 varies, but it does not affect the consistency as $C \to \infty$. Rainforth et al. (2018) provide relevant discussions on the efficiency of nested Monte Carlo schemes.

Normalizing constants. Nested expectations occur in the estimation of normalizing constants via thermodynamic integration or path sampling (Gelman and Meng, 1998). Introduce a sequence of distributions (π_λ) on $\mathbb{X}$, continuously indexed by $\lambda \in [0, 1]$, with $\pi_\lambda(x) = \exp(-U_\lambda(x))/Z_\lambda$. For example π_λ could be a posterior distribution where the like-

lihood was raised to the power λ. The thermodynamic integration identity reads:

$$\log(Z_1/Z_0) = -\int_0^1 \pi_\lambda(\nabla_\lambda U_\lambda) d\lambda, \tag{1.3.2}$$

which is an integral with respect to $\lambda \in [0,1]$ where the integrand is itself an expectation with respect to π_λ. Thus one can sample λ uniformly in $[0,1]$ and then approximate $-\pi_\lambda(\nabla_\lambda U_\lambda)$ with unbiased MCMC to obtain an unbiased estimator of $\log(Z_1/Z_0)$ (Rischard et al., 2018). Wang and Wang (2022) consider the problem of estimating a nonlinear function g of an expectation $\pi(h)$, and take the estimation of Z_1/Z_0 as an example. They develop generic unbiased estimators of $g(\pi(h))$ by combining unbiased MCMC with unbiased multilevel Monte Carlo (Blanchet et al., 2019).

1.3.3 Asymptotic variance

Unbiased estimators of the asymptotic variance. Unbiased MCMC and its connection to the Poisson equation elicited in Douc et al. (2023), see Section 1.2.2, lead to the construction of unbiased estimators of $v(P,h)$, the asymptotic variance in the CLT (1.1.2). A standard way of establishing the CLT for Markov chain averages (Douc et al., 2018, Chapter 21) is to write $\sum_{s=0}^{t-1}\{h(X_s)-\pi(h)\} = \sum_{s=1}^{t}\{g(X_s) - Pg(X_{s-1})\} + g(X_0) - g(X_t)$, where g is a solution of the Poisson equation (1.2.3), and then to observe that $\{g(X_s) - Pg(X_{s-1})\}_{s\geq 1}$ forms a martingale difference sequence. The CLT for martingale difference sequences yields

$$v(P,h) = \mathbb{E}_\pi[\{g(X_1) - Pg(X_0)\}^2] = 2\underbrace{\pi(\{h-\pi(h)\}g)}_{\text{(a)}} - \underbrace{(\pi(h^2) - \pi(h)^2)}_{\text{(b)}}. \tag{1.3.3}$$

The expression involves expectations with respect to π and the solution g of the Poisson equation. Using successful couplings, Douc et al. (2023) combine unbiased estimators G of evaluations g, from (1.2.6) in Section 1.2.2, with unbiased MCMC approximations $\hat{\pi}$ of π as in (1.2.9), to deliver estimators $\hat{v}(P,h)$ with $\mathbb{E}[\hat{v}(P,h)] = v(P,h)$. Considering the simpler problem of estimating $\pi(g)$, the idea goes as follows: first, run unbiased MCMC to obtain $\hat{\pi} = \sum_{n=1}^{N} \omega_n \delta_{Z_n}$ approximating π. Then sample I uniformly in $\{1,\ldots,N\}$, and generate $G(Z_I, y)$ with expectation $g(Z_I, y)$, as in (1.2.6) where y is an arbitrary state.

Finally, compute $N\omega_I G(Z_I, y)$, which has expectation equal to $\pi(g)$ under conditions similar to those of Theorem 1.2.1 on the meeting time τ and the function h.

In contrast to the bounds on $|\pi_k - \pi|_{\mathrm{TV}}$ presented in Section 1.3.1, that could be loose if the coupling is ill-chosen, here we can construct estimators with expectation equal to $v(P, h)$, thus the choice of coupling only affects variance and cost.

Practical significance. Estimation of $v(P, h)$ is required to compare the efficiency of unbiased MCMC relative to regular MCMC. It is also a key quantity to compare the performance of different MCMC algorithms. Unbiased estimators of $v(P, h)$ enable such efficiency comparisons without ever relying on long runs. Averages of C independent runs converge with the usual Monte Carlo rate. This compares favorably to classical estimators of $v(P, h)$. Indeed, commonly-used estimators of $v(P, h)$, such as batch means and spectral variance estimators, converge at a sub-Monte Carlo rate, e.g. $T^{-2/3}$ for batch means (Flegal and Jones, 2010). On the other hand, the unbiased estimators in Douc et al. (2023) require a successful coupling of the algorithm under consideration, and not just generated trajectories.

1.4 Design of successful coupling of MCMC algorithms

To implement unbiased MCMC, users need to design a successful coupling of their MCMC algorithm. Focusing on Markovian couplings, this amounts to constructing a coupled transition $\bar{P}$ to plug in Algorithm 1 and for which Assumption 1 is satisfied. Concretely, we need to be able to sample $(X, Y) \sim \bar{P}((x, y), \cdot)$, where (x, y) represent the current positions of the chains, such that 1) $X \sim P(x, \cdot)$ and $Y \sim P(y, \cdot)$, and 2) there is a possibility of meeting i.e. $\bar{P}((x, y), \{X = Y\}) > 0$ at least from some pairs (x, y). In Section 1.4.1 we review the more basic task of coupling random variables, before dealing with MCMC transitions in Section 1.4.2. References to realistic examples are provided in Section 1.4.3.

1.4.1 Couplings of random variables

Maximal couplings. A coupling of (X, Y) with $X \sim p$ and $Y \sim q$ is *maximal* if $\mathbb{P}(X = Y)$ is maximal and thus equal to $1 - |p - q|_{\mathrm{TV}}$. There may be more than one maximal coupling. Algorithm 2 is a modification by Gerber and Lee (2020) of the γ-coupling of Johnson (1998) with an extra parameter $\eta \in (0, 1]$. The scheme requires samples from p and q, and evaluations of the ratio of their densities. The probability of $\{X = Y\}$ is maximal only when to $\eta = 1$. However, the cost of running Algorithm 2, which contains a while loop, has a variance that goes to infinity when $\eta = 1$ and when $|p-q|_{\mathrm{TV}}$ goes to zero. With $\eta < 1$, the coupling is sub-maximal, but the variance of the cost is upper bounded uniformly over p and q. Under Algorithm 2, conditionally on $\{X \neq Y\}$, X is independent of Y.

Algorithm 2 Sampling a coupling of p and q, with parameter $\eta \in (0, 1]$. The coupling maximizes $\mathbb{P}(X = Y)$ when $\eta = 1$, but the variance of the cost is bounded when $\eta < 1$.

1. Sample $X \sim p$.
2. Sample $W \sim \mathrm{Uniform}(0, 1)$.
   (a) If $W \leq \min(\eta, q(X)/p(X))$, set $Y = X$.
   (b) Otherwise sample $Y^\star \sim q$ and $W^\star \sim \mathrm{Uniform}(0, 1)$ until $W^\star > \eta p(Y^\star)/q(Y^\star)$, and set $Y = Y^\star$.
3. Return (X, Y).

Algorithm 3 samples the same pairs (X, Y) as Algorithm 2 with $\eta = 1$, but via a mixture representation. Algorithm 3 is applicable when $\int \min(p(x), q(x))dx$ can be computed, and when $\tilde{p}$ and $\tilde{q}$ defined on line 3 can be sampled from. Its appeal is that its cost is deterministic.

Algorithm 3 Sampling a maximal coupling of p and q via a mixture. Note that $|p-q|_{\mathrm{TV}} = 1 - \int p \wedge q$, where $p \wedge q$ represents the point-wise minimum between p and q.

1. Draw $U \sim \mathrm{Uniform}(0, 1)$.
2. If $U \leq \int p \wedge q$, draw $X \sim \nu$ with $\nu : x \mapsto (p \wedge q)(x)/\int p \wedge q$, and set $Y = X$.
3. Otherwise, draw $X \sim \tilde{p}$ and $Y \sim \tilde{q}$ independently, with $\tilde{p} \propto p - p \wedge q$, $\tilde{q} \propto q - p \wedge q$.
4. Return (X, Y).

Synchronous couplings. The above algorithms generate X and Y independently in the event $\{X \neq Y\}$, and thus revert to the independent coupling when $|p - q|_{\mathrm{TV}}$ is close to

one. A natural way of introducing dependencies among random variables is to use common random numbers to generate them. In one dimension, if X has quantile function F_p^- and Y has quantile function F_q^-, then one can sample $U \sim \text{Uniform}(0,1)$ and set $X = F_p^-(U)$ and $Y = F_q^-(U)$. The resulting joint distribution minimizes $\mathbb{E}[(X-Y)^2]$: it is an *optimal transport* coupling (Villani, 2008), or equivalently it maximizes $\mathbb{C}\text{ov}(X,Y)$ (Glasserman and Yao, 1992). Typically, such couplings assign zero probability to the event $\{X = Y\}$. Figure 1.7 (left) illustrates with bivariate Normals the common random numbers or *synchronous coupling*, where $X = \mu_1 + \Sigma^{1/2}Z$ and $Y = \mu_2 + \Sigma^{1/2}Z$ with common $Z \sim \text{Normal}(0, I)$.

Reflection couplings. Spherically symmetric random variables are invariant by reflections with respect to planes passing through their center. From this observation, reflection couplings can be designed for e.g. Normal or Student distributions with means μ_1,μ_2 and equal variance Σ, in any dimension. The sample Y is defined by reflection of X with respect to the hyperplane bisecting the segment (μ_1, μ_2). The resulting coupling is synchronous in all directions orthogonal to the difference $(\mu_1 - \mu_2)$, along which it is anti-synchronous: $X - \mu_1$ and $Y - \mu_2$ either point toward each other, or face opposite directions. Bou-Rabee et al. (2020) propose a coupling that both maximizes $\mathbb{P}(X = Y)$ and reverts to a reflection coupling in the event $\{X \neq Y\}$, described in Algorithm 4 and with a deterministic cost. Figure 1.3 includes an implementation and Figure 1.7 (right) represents a sample from a reflection coupling.

Algorithm 4 A coupling for distributions p and q obtained from a common spherically symmetric distribution s, rescaled with a common covariance Σ, and shifted by μ_1 and μ_2 respectively. For example $p = \text{Normal}(\mu_1, \Sigma)$ and $q = \text{Normal}(\mu_2, \Sigma)$ if $s = \text{Normal}(0, I)$.

---

1. Let $\Delta = \Sigma^{-1/2}(\mu_1 - \mu_2)$ and $e = \Delta/|\Delta|$ where $|\cdot|$ is the L_2 norm.
2. Sample $\dot{X} \sim s$, and $W \sim \text{Uniform}(0,1)$.
3. If $s(\dot{X})W \leq s(\dot{X} + \Delta)$, set $\dot{Y} = \dot{X} + \Delta$.
4. Else set $\dot{Y} = \dot{X} - 2(e^T\dot{X})e$.
5. Set $X = \Sigma^{1/2}\dot{X} + \mu_1, Y = \Sigma^{1/2}\dot{Y} + \mu_2$, and return (X, Y).

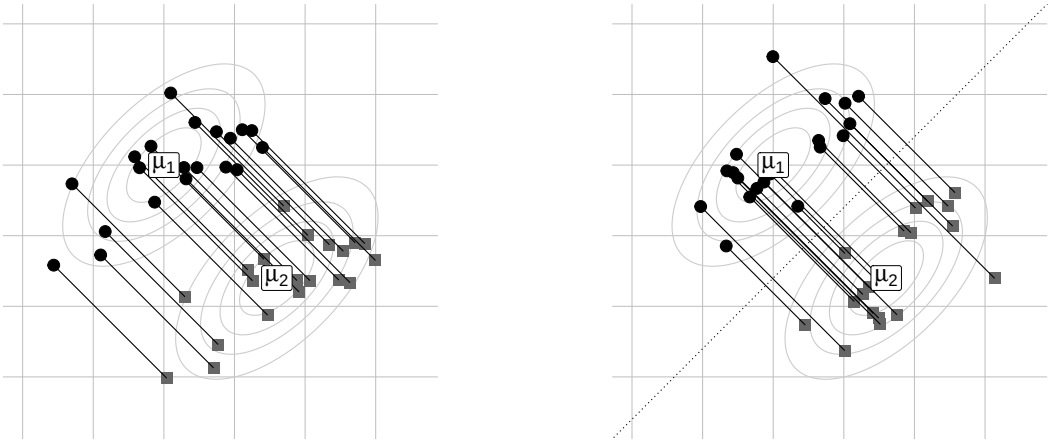

Figure 1.7: Left: common random numbers or *synchronous* coupling of Normal(μ_1, Σ) and Normal(μ_2, Σ). Each draw (X, Y) is represented by a segment. Right: *reflection* coupling. The dotted line represents the line bisecting the segment (μ_1, μ_2). The lengths $|X - Y|$ are constant under the synchronous coupling, but vary under the reflection coupling.

1.4.2 Coupling MCMC transitions

Coupling the constituents of a transition. MCMC algorithms describe how to obtain X_t conditional on $X_{t-1} = x$ through a succession of steps. With MRTH (e.g. Figure 1.3), a proposal $X^\star$ is sampled from a transition $q(x, \cdot)$ (step 1), then U is sampled from Uniform$(0, 1)$ and X_t is set to $X^\star$ if $U < \pi(X^\star)q(X^\star, x)/\pi(x)q(x, X^\star)$, or to x otherwise (step 2). Couplings of the entire transition can be obtained by coupling each of the steps, e.g. coupling proposals $X^\star \sim q(x, \cdot)$ and $Y^\star \sim q(y, \cdot)$, and then coupling the Uniforms employed for accepting or rejecting the proposals. For example Johnson (1998) uses maximal couplings as in Algorithm 2 for the proposals, and a common Uniform for acceptance. Wang et al. (2021) refine the coupling of the Uniforms to maximize the probability of $\{X_t = Y_t\}$. O'Leary and Wang (2021) show that all couplings of MRTH transitions can be obtained by certain stepwise couplings. Since there are often many coupling possibilities for each step, it is impossible to test all combinations. Below we provide some guiding principles.

Contracting before meeting. For some MCMC algorithms it may be possible to sample from a maximal coupling of $P(x, \cdot)$ and $P(y, \cdot)$ (e.g. Wang et al., 2021, for MRTH). However, even the maximal probability of $\{X = Y\}$, which is $1 - |P(x, \cdot) - P(y, \cdot)|_{\mathrm{TV}}$, is very small unless x and y are close to one another. Under an independent coupling, this

would rarely occur. Thus, the aim is first to bring the chains closer, so that they may then have a decent chance to meet. A coupling of P may alternate between different strategies depending on the current states x and y: for example one can employ a *contractive coupling* of P (as described below) if $|x - y|$ is large and a maximal coupling of P if $|x - y|$ is small. Eberle (2016) refers to such alternation as *mixed couplings*.

Contractive couplings. In the context of Markov chains, the transition $X_t \sim P(X_{t-1}, \cdot)$ can be represented as $X_t = \psi(X_{t-1}, U_t)$, where U_t is a source of randomness and ψ is a deterministic function. Then the *synchronous coupling* refers to the computation of $X_t = \psi(X_{t-1}, U_t)$ and $Y_t = \psi(Y_{t-1}, U_t)$ using the same random variable U_t at time t. It has long been observed that synchronous couplings of MCMC algorithms can be contractive (Agapiou et al., 2018; Johnson, 1996; Neal, 1999; Neal and Pinto, 2001), in the sense that the generated chains tend to get closer to one another. Assuming strong convexity of the potential function, it is known that common noise terms result in contraction for Langevin diffusions (pages 22-23 in Villani, 2008), for unadjusted Langevin (e.g. Appendix A in Wibisono, 2018), and for Hamiltonian Monte Carlo (e.g. Mangoubi and Smith, 2017) for at least some tuning parameters. Contraction from synchronous couplings has been observed in the context of Gibbs samplers, sometimes in high dimensions, e.g. Biswas et al. (2022) for linear regression with horseshoe priors, and Atchadé and Wang (2023) for regression with spike-and-slab priors. It does not always work either: for example two Brownian motions started at different positions and driven by the same noise always remain at a constant distance.

Reflection couplings were introduced to analyze Brownian motions on Euclidean spaces, leading to the smallest possible meeting times (Hsu and Sturm, 2013; Lindvall and Rogers, 1986). Reflections were later employed to obtain contraction for various processes: Eberle (2016) for a class of diffusion processes, Eberle et al. (2019) for Langevin dynamics, Bou-Rabee et al. (2020) for Hamiltonian Monte Carlo, for example. Jacob et al. (2020b) observe good performance of Algorithm 4 for random walk proposals in MRTH, on spherical Normal distributions as the dimension increases. Papp and Sherlock (2022b) establish this formally and propose another coupling, termed *gradient common random number coupling*, which is shown to work optimally for a class of target distributions.

Mixing different MCMC transitions to enable meetings. For an MCMC algorithm with transition P_1, it may be possible to design a contractive coupling $\bar{P}_1$ without being able to induce meetings. For example with Hamiltonian Monte Carlo, contraction can result from the use of common momentum variables (e.g Mangoubi and Smith, 2017). However, to obtain meetings would require pairs of momentum variables such that two Hamiltonian trajectories, propagated with these momentum variables, would end up at the same final position (see Figure 1 in Bou-Rabee and Eberle, 2023). A way to bypass this difficulty is to introduce another transition P_2 with a coupling $\bar{P}_2$ that induces meetings when chains are close. Heng and Jacob (2019) then propose to use the mixture $w_1 P_1 + w_2 P_2$, with coupling given by the mixture $w_1 \bar{P}_1 + w_2 \bar{P}_2$. The resulting chains contract thanks to $\bar{P}_1$ and have a chance to meet thanks to $\bar{P}_2$. Careful: it would not be legal to employ $\bar{P}_1$ when the chains are distant and $\bar{P}_2$ when they are close, as this would violate the marginal constraint that each chain evolves according to $w_1 P_1 + w_2 P_2$.

1.4.3 References to couplings of realistic MCMC algorithms

Successful couplings have been developed for a number of popular MCMC algorithms.

Discrete state spaces. Convergence diagnostics are challenging on discrete spaces, for which visualization is difficult; there, unbiased MCMC could be particularly useful. Jacob et al. (2020b) present a coupling of the Gibbs sampler studied in Yang et al. (2016) for Bayesian variable selection in high dimension. Nguyen et al. (2022) couple Gibbs samplers to perform Bayesian data clustering, where the states are partitions of finite sets. Kelly et al. (2023) couple MCMC samplers for phylogenetic inference, where the state space is that of discrete tree topologies along with parameters and latent variables.

Particle filtering and importance sampling. Conditional particle filters for smoothing in state space models are coupled in Jacob et al. (2020a). Lee et al. (2020) extend the methodology and propose a detailed study of the meeting times. Particle marginal Metropolis–Hastings for Bayesian inference in state space models (Andrieu et al., 2010) is coupled in Middleton et al. (2020). Particle independent Metropolis–Hastings is coupled in

(Middleton et al., 2019), with the curious implication that the bias of self-normalized importance sampling estimators can be removed in finite time; and likewise for general sequential Monte Carlo samplers. Ruiz et al. (2021) couple variants of iterated sampling importance resampling to fit variational auto-encoders.

Gradient-based MCMC. Heng and Jacob (2019); Xu et al. (2021) consider couplings of simple variants of Hamiltonian Monte Carlo with applications to logistic regression and log-Gaussian Cox point processes in non-trivial dimensions. Reflection couplings as in Figure 1.3 or Algorithm 4 can be directly used for Langevin Monte Carlo. Corenflos et al. (2023) propose couplings of some piecewise deterministic MCMC algorithms such as the bouncy particle sampler (Bouchard-Côté et al., 2018).

Tempering. For multimodal targets, a standard strategy is *tempering*. Jacob et al. (2020b) couple a parallel tempering version of a Gibbs sampler for the Ising model. Zhu and Atchadé (2020) consider simulated tempering for sparse canonical correlation analysis.

1.5 Comments and outstanding questions

1.5.1 Usefulness

Access to unbiased signed measures approximating the target π facilitates parallel computing: instead of long chains, unbiased MCMC users rely on large numbers of independent runs. The lack of bias has other appeals. For example iterative optimization methods may require the approximation of an integral at each iteration. Such approximation should preferably be unbiased to prevent accumulation of bias over the iterations (e.g. Tadić and Doucet, 2011). The usefulness of unbiased MCMC is investigated in the context of a Monte Carlo Expectation-Maximization scheme in Chen et al. (2018), and of a stochastic gradient optimization for variational auto-encoders in Ruiz et al. (2021).

Access to unbiased estimators enables various statistical tools that are primarily developed for independent variables. For example one can readily replace empirical averages by more

robust estimators of expectations. Nguyen et al. (2022) consider trimmed means, and one could naturally employ *median-of-means* estimators (Lecué and Lerasle, 2020; Lugosi and Mendelson, 2019) to aggregate unbiased MCMC estimators that have two finite moments under conditions stated in Theorem 1.2.1.

Unbiased estimators can also be plugged into the framework of multi-arm bandits, for example to identify the algorithm with minimal asymptotic variance among a collection of MCMC algorithms, or to identify the model with largest marginal likelihood in a collection of models. One could view each algorithm (or model) as an arm, and each unbiased estimator of an asymptotic variance (or a normalizing constant) as an observed loss. Then, *best arm identification* techniques (Audibert et al., 2010) can be used to find, as efficiently as possible, the arm associated with the smallest expected loss.

1.5.2 Applicability

There exists a world, of a size to be determined, between standard MCMC and perfect sampling, where unbiased estimators can be obtained but not exact samples (Glynn, 2016). Access to that world demands successful couplings of MCMC algorithms. Currently, such construction is endeavored algorithm-by-algorithm, by mixing ingredients such as maximal couplings, common random numbers and reflections. There is no guarantee that such *ad hoc* constructions can always be found. According to the theory (e.g. Pitman, 1976), for ergodic chains there always exist couplings such that $|\pi_t - \pi_0| = \mathbb{P}(\tau > t)$, but there may not be implementable in settings of relevance for MCMC practitioners. It is however possible to plug arbitrary Markov transitions into mixtures of kernels, as described in Section 1.4.2, or in an SMC sampler, and then to remove its bias via a generic coupling of particle independent Metropolis–Hastings (Middleton et al., 2019).

A successful coupling of inhomogeneous Markov transitions, e.g. for adaptive MCMC algorithms, remains elusive. For a stochastic process (X_t) with marginals converging to π in the sense that $\pi(h) = \lim_{t\to\infty} \mathbb{E}[h(X_t)]$, assuming that (Y_t) is a copy of (X_t) such that

$\sum_{t\geq 1} \mathbb{E}\left[|h(X_t) - h(Y_{t-1})|\right]$ is finite, then $\pi(h)$ has the representation

$$\pi(h) = \mathbb{E}[h(X_0) + \sum_{t\geq 1} (h(X_t) - h(Y_{t-1}))]. \tag{1.5.1}$$

Many adaptive MCMC algorithms are known to have converging marginals (Andrieu and Thoms, 2008; Atchade et al., 2011). In principle the debiasing device could be applied to (1.5.1), but it is unclear how to construct a faithful coupling of (X_t) and (Y_t) that would lead to an unbiased estimator with a finite computing time. Random truncation techniques as in Section 1.2.1 could be used, but good performance would depend on a choice of truncation variable that assumes detailed knowledge of (X_t).

The benefits of unbiased estimators are also reaped under different frameworks that do not require successful couplings, such as regeneration (Mykland et al., 1995). Techniques such as Brockwell and Kadane (2005) to automate the identification of regeneration times could also suggest coupling strategies, e.g. where chains could meet when they simultaneously visit regenerative atoms.

Links to code repositories and complementary information can be found on the companion website at `https://pierrejacob.quarto.pub/unbiased-mcmc`.

Bibliography

Agapiou, S., Roberts, G. O., and Vollmer, S. J. (2018). Unbiased Monte Carlo: Posterior estimation for intractable/infinite-dimensional models. *Bernoulli*, 24(3):1726–1786.

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(3):269–342.

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and computing*, 18:343–373.

Atchade, Y., Fort, G., Moulines, É., and Priouret, P. (2011). Adaptive Markov chain Monte Carlo: theory and methods. *Bayesian time series models*, 1.

Atchadé, Y. and Wang, L. (2023). A fast asynchronous Markov chain Monte Carlo sampler for sparse Bayesian inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkad078.

Audibert, J.-Y., Bubeck, S., and Munos, R. (2010). Best arm identification in multi-armed bandits. In *COLT*, pages 41–53.

Biswas, N., Bhattacharya, A., Jacob, P. E., and Johndrow, J. E. (2022). Coupling-based convergence assessment of some Gibbs samplers for high-dimensional Bayesian regression with shrinkage priors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(3):973–996.

Biswas, N., Jacob, P. E., and Vanetti, P. (2019). Estimating convergence of Markov chains

with L-lag couplings. In *Advances in Neural Information Processing Systems*, pages 7389–7399.

Blanchet, J. H., Glynn, P. W., and Pei, Y. (2019). Unbiased multilevel Monte Carlo: Stochastic optimization, steady-state simulation, quantiles, and other applications. *arXiv preprint arXiv:1904.09929*.

Blocker, A. W. and Meng, X.-L. (2013). The potential and perils of preprocessing: Building new foundations. *Bernoulli*, 19(4):1176 – 1211.

Bou-Rabee, N. and Eberle, A. (2023). Mixing time guarantees for unadjusted Hamiltonian Monte Carlo. *Bernoulli*, 29(1):75–104.

Bou-Rabee, N., Eberle, A., and Zimmer, R. (2020). Coupling and convergence for Hamiltonian Monte Carlo. *The Annals of Applied Probability*, 30(3):1209–1250.

Bouchard-Côté, A., Vollmer, S. J., and Doucet, A. (2018). The bouncy particle sampler: A nonreversible rejection-free Markov chain Monte Carlo method. *Journal of the American Statistical Association*, 113(522):855–867.

Brockwell, A. E. and Kadane, J. B. (2005). Identification of regeneration times in MCMC simulation, with application to adaptive schemes. *Journal of Computational and Graphical Statistics*, 14(2):436–458.

Chen, W., Ma, L., and Liang, X. (2018). Blind identification based on expectation-maximization algorithm coupled with blocked Rhee–Glynn smoothing estimator. *IEEE Communications Letters*, 22(9):1838–1841.

Corenflos, A., Sutton, M., and Chopin, N. (2023). Debiasing piecewise deterministic Markov process samplers using couplings. *arXiv preprint arXiv:2306.15422*.

Cowles, M. K. and Rosenthal, J. S. (1998). A simulation approach to convergence rates for Markov chain Monte Carlo algorithms. *Statistics and Computing*, 8:115–124.

Craiu, R. V. and Meng, X.-L. (2022). Double happiness: enhancing the coupled gains of L-lag coupling via control variates. *Statistica Sinica*, 32:1–22.

Douc, R., Jacob, P. E., Lee, A., and Vats, D. (2023). Solving the Poisson equation using coupled Markov chains. *arXiv preprint arXiv:2206.05691*.

Douc, R., Moulines, É., Priouret, P., and Soulier, P. (2018). *Markov chains*. Springer International Publishing.

Durmus, A. and Moulines, É. (2022). On the geometric convergence for MALA under verifiable conditions. *arXiv preprint arXiv:2201.01951*.

Durmus, A., Moulines, É., and Saksman, E. (2017). On the convergence of Hamiltonian Monte Carlo. *arXiv preprint arXiv:1705.00166*.

Eberle, A. (2016). Reflection couplings and contraction rates for diffusions. *Probability Theory and Related Fields*, 166(3):851–886.

Eberle, A., Guillin, A., and Zimmer, R. (2019). Couplings and quantitative contraction rates for Langevin dynamics. *The Annals of Probability*, 47(4):1982–2010.

Flegal, J. M. and Jones, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *The Annals of Statistics*, 38(2):1034 – 1070.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185.

Gerber, M. and Lee, A. (2020). Discussion on the paper by Jacob, O'Leary, and Atchadé. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):584–585.

Glasserman, P. and Yao, D. D. (1992). Some guidelines and guarantees for common random numbers. *Management Science*, 38(6):884–908.

Glynn, P. W. (1983). Randomized estimators for time integrals. Technical report, University of Wisconsin–Madison.

Glynn, P. W. (2016). Exact simulation vs exact estimation. In *2016 Winter Simulation Conference (WSC)*, pages 193–205. IEEE.

Glynn, P. W. and Heidelberger, P. (1990). Bias properties of budget constrained simulations. *Operations Research*, 38(5):801–814.

Glynn, P. W. and Heidelberger, P. (1991). Analysis of parallel replicated simulations under a completion time constraint. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 1(1):3–23.

Glynn, P. W. and Rhee, C.-H. (2014). Exact estimation for Markov chain equilibrium expectations. *Journal of Applied Probability*, 51(A):377–389.

Glynn, P. W. and Whitt, W. (1992). The asymptotic efficiency of simulation estimators. *Operations Research*, 40(3):505–520.

Heng, J. and Jacob, P. E. (2019). Unbiased Hamiltonian Monte Carlo with couplings. *Biometrika*, 106(2):287–302.

Hsu, E. P. and Sturm, K.-T. (2013). Maximal coupling of Euclidean Brownian motions. *Communications in Mathematics and Statistics*, 1:93–104.

Jacob, P. E., Lindsten, F., and Schön, T. B. (2020a). Smoothing with couplings of conditional particle filters. *Journal of the American Statistical Association*, 115(530):721–729.

Jacob, P. E., O'Leary, J., and Atchadé, Y. F. (2020b). Unbiased Markov chain Monte Carlo methods with couplings. *Journal of the Royal Statistical Society Series B (with discussion)*, 82(3):543–600.

Jacob, P. E. and Thiery, A. H. (2015). On nonnegative unbiased estimators. *The Annals of Statistics*, 43(2):769 – 784.

Johnson, V. E. (1996). Studying convergence of Markov chain Monte Carlo algorithms using coupled sample paths. *Journal of the American Statistical Association*, 91(433):154–166.

Johnson, V. E. (1998). A coupling-regeneration scheme for diagnosing convergence in Markov chain Monte Carlo algorithms. *Journal of the American Statistical Association*, 93(441):238–248.

Kelly, L. J., Ryder, R. J., and Clarté, G. (2023). Lagged couplings diagnose Markov chain Monte Carlo phylogenetic inference. *Annals of Applied Statistics (to appear)*.

Kontoyiannis, I. and Dellaportas, P. (2009). Notes on using control variates for estimation with reversible MCMC samplers. *arXiv preprint arXiv:0907.4160*.

Lecué, G. and Lerasle, M. (2020). Robust machine learning by median-of-means: Theory and practice. *The Annals of Statistics*, 48(2):906 – 931.

Lee, A., Singh, S. S., and Vihola, M. (2020). Coupled conditional backward sampling particle filter. *The Annals of Statistics*, 48(5):3066 – 3089.

Levin, D. A. and Peres, Y. (2017). *Markov chains and mixing times*, volume 107. American Mathematical Soc.

Lindvall, T. (2002). *Lectures on the coupling method*. Courier Corporation.

Lindvall, T. and Rogers, L. C. G. (1986). Coupling of multidimensional diffusions by reflection. *The Annals of Probability*, pages 860–872.

Liu, F., Bayarri, M., Berger, J., et al. (2009). Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis*, 4(1):119–150.

Lugosi, G. and Mendelson, S. (2019). Sub-Gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783 – 794.

Mangoubi, O. and Smith, A. (2017). Rapid mixing of Hamiltonian Monte Carlo on strongly log-concave distributions. *arXiv preprint arXiv:1708.07114*.

McCartan, C. and Imai, K. (2020). Sequential Monte Carlo for sampling balanced and compact redistricting plans. *arXiv preprint arXiv:2008.06131*.

McLeish, D. (2011). A general method for debiasing a Monte Carlo estimator. *Monte Carlo methods and applications*, 17(4):301–315.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Middleton, L., Deligiannidis, G., Doucet, A., and Jacob, P. E. (2019). Unbiased smoothing using particle independent Metropolis–Hastings. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2378–2387. PMLR.

Middleton, L., Deligiannidis, G., Doucet, A., and Jacob, P. E. (2020). Unbiased Markov chain Monte Carlo for intractable target distributions. *Electronic Journal of Statistics*, 14(2):2842–2891.

Mykland, P., Tierney, L., and Yu, B. (1995). Regeneration in Markov chain samplers. *Journal of the American Statistical Association*, 90(429):233–241.

Neal, R. M. (1999). Circularly-coupled Markov chain sampling. Technical report, Department of Statistics, University of Toronto.

Neal, R. M. and Pinto, R. L. (2001). Improving Markov chain Monte Carlo estimators by coupling to an approximating chain. Technical report, Department of Statistics, University of Toronto.

Nguyen, T. D., Trippe, B. L., and Broderick, T. (2022). Many processors, little time: MCMC for partitions via optimal transport couplings. In *International Conference on Artificial Intelligence and Statistics*, pages 3483–3514. PMLR.

O'Leary, J. and Wang, G. (2021). Metropolis–Hastings transition kernel couplings. *arXiv preprint arXiv:2102.00366*.

Papp, T. and Sherlock, C. (2022a). Bounds on Wasserstein distances between continuous distributions using independent samples. *arXiv preprint arXiv:2203.11627*.

Papp, T. P. and Sherlock, C. (2022b). A new and asymptotically optimally contracting coupling for the random walk Metropolis. *arXiv preprint arXiv:2211.12585*.

Pinto, R. L. and Neal, R. M. (2001). Improving Markov chain Monte Carlo estimators by coupling to an approximating chain. Technical report, University of Toronto.

Pitman, J. (1976). On coupling of Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 35(4):315–322.

Plummer, M. (2015). Cuts in Bayesian graphical models. *Statistics and Computing*, 25:37–43.

Propp, J. G. and Wilson, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures & Algorithms*, 9(1-2):223–252.

Rainforth, T., Cornish, R., Yang, H., Warrington, A., and Wood, F. (2018). On nesting Monte Carlo estimators. In *International Conference on Machine Learning*, pages 4267–4276. PMLR.

Rhee, C.-H. and Glynn, P. W. (2015). Unbiased estimation with square root convergence for SDE models. *Operations Research*, 63(5):1026–1043.

Rischard, M., Jacob, P. E., and Pillai, N. (2018). Unbiased estimation of log normalizing constants with applications to Bayesian cross-validation. *arXiv preprint arXiv:1810.01382*.

Robert, C. P. (1995). Convergence control methods for Markov chain Monte carlo algorithms. *Statistical Science*, 10(3):231–253.

Roberts, G. O. and Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71.

Roberts, G. O. and Tweedie, R. L. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110.

Rosenthal, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 90(430):558–566.

Rosenthal, J. S. (2000). Parallel computing and Monte Carlo algorithms. *Far east journal of theoretical statistics*, 4(2):207–236.

Ruiz, F. J., Titsias, M. K., Cemgil, T., and Doucet, A. (2021). Unbiased gradient estimation for variational auto-encoders using coupled Markov chains. In *Uncertainty in Artificial Intelligence*, pages 707–717. PMLR.

Rychlik, T. (1990). Unbiased nonparametric estimation of the derivative of the mean. *Statistics & probability letters*, 10(4):329–333.

Tadić, V. B. and Doucet, A. (2011). Asymptotic bias of stochastic gradient search. In *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pages 722–727. IEEE.

Vanetti, P. and Doucet, A. (2020). Discussion on the paper by Jacob, O'Leary, and Atchadé. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):584–585.

Vihola, M. (2018). Unbiased estimators and multilevel Monte Carlo. *Operations Research*, 66(2):448–462.

Villani, C. (2008). *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg.

Wang, G., O'Leary, J., and Jacob, P. E. (2021). Maximal couplings of the Metropolis–Hastings algorithm. In *International Conference on Artificial Intelligence and Statistics*, pages 1225–1233. PMLR.

Wang, G. and Wang, T. (2022). Unbiased multilevel Monte Carlo methods for intractable distributions: MLMC meets MCMC. *arXiv preprint arXiv:2204.04808*.

Wibisono, A. (2018). Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Conference on Learning Theory*, pages 2093–3027. PMLR.

Xu, K., Fjelde, T. E., Sutton, C., and Ge, H. (2021). Couplings for Multinomial Hamiltonian Monte Carlo. In *International Conference on Artificial Intelligence and Statistics*, pages 3646–3654. PMLR.

Yang, Y., Wainwright, M. J., and Jordan, M. I. (2016). On the computational complexity of high-dimensional Bayesian variable selection. *The Annals of Statistics*, 44(6):2497 – 2532.

Zhu, Q. and Atchadé, Y. F. (2020). Minimax quasi-Bayesian estimation in sparse canonical correlation analysis via a Rayleigh quotient function. *arXiv preprint arXiv:2010.08627*.