Biometrika (2019), **00**, 0, *pp.* 1–16
Printed in Great Britain
doi: 10.1093/biomet/asy074

Unbiased Hamiltonian Monte Carlo with couplings

BY J. HENG

ESSEC Business School, 5 Nepal Park, Singapore 139408, Republic of Singapore
b00760223@essec.edu

AND P. E. JACOB

Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, Massachusetts 02138, U.S.A.
pjacob@fas.harvard.edu

SUMMARY

We propose a method for parallelization of Hamiltonian Monte Carlo estimators. Our approach involves constructing a pair of Hamiltonian Monte Carlo chains that are coupled in such a way that they meet exactly after some random number of iterations. These chains can then be combined so that the resulting estimators are unbiased. This allows us to produce independent replicates in parallel and average them to obtain estimators that are consistent in the limit of the number of replicates, rather than in the usual limit of the number of Markov chain iterations. We investigate the scalability of our coupling in high dimensions on a toy example. The choice of algorithmic parameters and the efficiency of our proposed approach are then illustrated on a logistic regression with 300 covariates and a log-Gaussian Cox point processes model with low- to fine-grained discretizations.

Some key words: Coupling; Hamiltonian Monte Carlo method; Parallel computing; Unbiased estimation.

1. INTRODUCTION

1.1. *Parallel computation with Hamiltonian Monte Carlo*

Hamiltonian Monte Carlo is a Markov chain Monte Carlo method for approximating integrals with respect to a target probability distribution π on $\mathbb{R}^d$. Originally proposed by Duane et al. (1987) in the physics literature, it was later introduced into statistics by Neal (1993) and is now widely adopted as a standard sampling tool (Lelièvre et al., 2010; Brooks et al., 2011). Various aspects of its theoretical properties have been studied; see Betancourt et al. (2017) and Betancourt (2017) for geometric properties, Livingstone et al. (2016) and Durmus et al. (2017) for ergodicity results, and Beskos et al. (2013), Mangoubi & Smith (2017) and Bou-Rabee et al. (2018) for scaling results with respect to the dimension d. These results suggest that Hamiltonian Monte Carlo compares favourably with other Markov chain Monte Carlo algorithms such as random walk Metropolis–Hastings and Metropolis-adjusted Langevin algorithms in high dimensions. In practice, Hamiltonian Monte Carlo is at the core of the no-U-turn sampler (Hoffman & Gelman, 2014) implemented in the software Stan (Carpenter et al., 2016).

If one could initialize from the target distribution, the usual estimators based on any Markov chain Monte Carlo algorithm would be unbiased, and one could simply average over independent

chains (Rosenthal, 2000). Except for certain applications in which this can be achieved with perfect simulation methods (Casella et al., 2001; Huber, 2016), Markov chain Monte Carlo estimators are ultimately consistent in the limit of the number of iterations. Algorithms that rely on such asymptotics face the risk of becoming obsolete if computational power continues to increase through the number of available processors and not through clock speed.

Several methods have been proposed to address this limitation with varying generality (Mykland et al., 1995; Neal, 2017; Glynn & Rhee, 2014). Our approach builds upon recent work of Jacob et al. (2017), which introduces unbiased estimators based on Metropolis–Hastings algorithms and Gibbs samplers. The present article describes how to design unbiased estimators for Hamiltonian Monte Carlo and some of its variants (Girolami & Calderhead, 2011). The proposed method is widely applicable and involves a simple coupling between a pair of Hamiltonian Monte Carlo chains. Coupled chains are run for a random but almost surely finite number of iterations, and are combined in such a way that the resulting estimators will be unbiased. One can produce independent copies of these estimators in parallel and average them to obtain consistent approximations in the limit of the number of replicates. This also yields confidence intervals that are valid in the number of replicates by the central limit theorem; see also Glynn & Heidelberger (1991) for central limit theorems parameterized by the number of processors or time budget.

1.2. *Notation*

Given a sequence $(x_n)_{n\geqslant 0}$ and integers $k < m$, we adopt the convention that $\sum_{n=m}^{k} x_n = 0$. The set of natural numbers is denoted by $\mathbb{N}$ and the set of nonnegative real numbers by $\mathbb{R}_+$. The d-dimensional vector of zeros is denoted by 0_d and the $d \times d$ identity matrix by I_d. The Euclidean norm of a vector $x \in \mathbb{R}^d$ is written as $|x| = (\sum_{i=1}^{d} x_i^2)^{1/2}$. Given a subset $A \subseteq \Omega$, the indicator function $\mathbb{I}_A : \Omega \to \{0, 1\}$ is defined by $\mathbb{I}_A(x) = 1$ if $x \in A$ and 0 if $x \in \Omega \setminus A$. For a smooth function $f : \mathbb{R}^d \to \mathbb{R}$, we denote its gradient by $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ and its Hessian by $\nabla^2 f : \mathbb{R}^d \to \mathbb{R}^{d\times d}$. The gradients of a function $(x, y) \mapsto f(x, y)$ with respect to the variables x and y are written as $\nabla_x f$ and $\nabla_y f$, respectively. Given functions $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^d \to \mathbb{R}^n$, we define the composition $f \circ g : \mathbb{R}^d \to \mathbb{R}^m$ by $(f \circ g)(x) = f\{g(x)\}$ for all $x \in \mathbb{R}^d$. The Borel σ-algebra of $\mathbb{R}^d$ is denoted by $\mathcal{B}(\mathbb{R}^d)$; on the product space $\mathbb{R}^d \times \mathbb{R}^d$, $\mathcal{B}(\mathbb{R}^d) \times \mathcal{B}(\mathbb{R}^d)$ denotes the product σ-algebra. The Gaussian distribution on $\mathbb{R}^d$ with mean vector μ and covariance matrix Σ is denoted by $\mathcal{N}(\mu, \Sigma)$ and its density by $x \mapsto \mathcal{N}(x; \mu, \Sigma)$. The uniform distribution on $[0, 1]$ is denoted by $\mathrm{Un}[0, 1]$. We use the shorthand $X \sim \eta$ to refer to a random variable with distribution η. On a measurable space $(\Omega, \mathcal{F})$, given a measurable function $\varphi : \Omega \to \mathbb{R}$, a probability measure η and a Markov transition kernel M, we define the integral $\eta(\varphi) = \int_\Omega \varphi(x)\eta(\mathrm{d}x)$ and the function $M(\varphi)(x) = \int_\Omega \varphi(y)M(x, \mathrm{d}y)$ for $x \in \Omega$.

1.3. *Unbiased estimation with couplings*

Suppose that $h : \mathbb{R}^d \to \mathbb{R}$ is a measurable function of interest and consider the task of approximating the integral $\pi(h) = \int h(x)\pi(\mathrm{d}x) < \infty$. Following Glynn & Rhee (2014) and Jacob et al. (2017), we will construct a pair of coupled Markov chains $X = (X_n)_{n\geqslant 0}$ and $Y = (Y_n)_{n\geqslant 0}$ with the same marginal law, associated with an initial distribution π_0 and a π-invariant Markov transition kernel K defined on $\{\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\}$. To do so, we introduce a Markov transition kernel $\bar{K}$ on $\{\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d) \times \mathcal{B}(\mathbb{R}^d)\}$ that admits K as its marginals, i.e., $\bar{K}\{(x, y), A \times \mathbb{R}^d\} = K(x, A)$ and $\bar{K}\{(x, y), \mathbb{R}^d \times A\} = K(y, A)$ for all $x, y \in \mathbb{R}^d$ and $A \in \mathcal{B}(\mathbb{R}^d)$. After initializing $(X_0, Y_0) \sim \bar{\pi}_0$ with a coupling that has π_0 as its marginals, we simulate $X_1 \sim K(X_0, \cdot)$ and $(X_{n+1}, Y_n) \sim \bar{K}\{(X_n, Y_{n-1}), \cdot\}$ for all integers $n \geqslant 1$. We will write pr for the law of the coupled

chain $(X_n, Y_n)_{n\geqslant 0}$ and use E to denote expectation with respect to pr. We now make the following assumptions.

Assumption 1 (Convergence of marginal chain). As $n \to \infty$, $E\{h(X_n)\} \to \pi(h)$. Furthermore, there exist $\kappa_1 > 0$ and $C_1 < \infty$ such that $E\{h(X_n)^{2+\kappa_1}\} < C_1$ for all integers $n \geqslant 0$.

Assumption 2 (Tail of meeting time). The meeting time $\tau = \inf\{n \geqslant 1 : X_n = Y_{n-1}\}$ satisfies a geometric tail condition of the form $\mathrm{pr}(\tau > n) \leqslant C_2\kappa_2^n$ for some constants $C_2 \in \mathbb{R}_+$ and $\kappa_2 \in (0, 1)$ and all integers $n \geqslant 0$.

Assumption 3 (Faithfulness). The coupled chains are faithful (Rosenthal, 1997), i.e., $X_n = Y_{n-1}$ for all integers $n \geqslant \tau$.

Under these assumptions, the random variable defined as

$$H_k(X, Y) = h(X_k) + \sum_{n=k+1}^{\tau-1} \{h(X_n) - h(Y_{n-1})\} \tag{1}$$

for any integer $k \geqslant 0$ is an unbiased estimator of $\pi(h)$ with finite variance (Jacob et al., 2017, Proposition 3.1). Computation of (1) can be performed with $\tau-1$ applications of $\bar{K}$ and $\max(1, k+1-\tau)$ applications of K; thus the computational cost has a finite expectation under Assumption 2. The first term, $h(X_k)$, is in general biased since the chain $(X_n)_{n\geqslant 0}$ may not have reached stationarity by iteration k. The second term acts as a bias correction and is equal to zero when $k \geqslant \tau - 1$.

As the estimators $H_k(X, Y)$, for various values of k, can be computed from a single realization of the coupled chains, this prompts the definition of a time-averaged estimator $H_{k:m}(X, Y) = (m - k + 1)^{-1} \sum_{n=k}^{m} H_n(X, Y)$ for integers $k \leqslant m$. The latter inherits the unbiasedness and finite-variance properties, and can be rewritten as

$$H_{k:m}(X, Y) = M_{k:m}(X) + \sum_{n=k+1}^{\tau-1} \min\left(1, \frac{n-k}{m-k+1}\right) \{h(X_n) - h(Y_{n-1})\}, \tag{2}$$

where $M_{k:m}(X) = (m-k+1)^{-1} \sum_{n=k}^{m} h(X_n)$ can be viewed as the usual Markov chain estimator with m iterations and a burn-in period of $k-1$. As before, the second term plays a bias correction role and is equal to zero when $k \geqslant \tau - 1$. Hence, if the value of k is sufficiently large, we can expect the variance of $H_{k:m}(X, Y)$ to be close to that of $M_{k:m}(X)$. Moreover, the cost of computing (2), which involves $\tau - 1$ applications of $\bar{K}$ and $\max(1, m+1-\tau)$ applications of K, becomes comparable to m iterations under K for sufficiently large m. Therefore we can expect the asymptotic inefficiency of $H_{k:m}(X, Y)$, given by the product of the expected computational cost and the variance of $H_{k:m}(X, Y)$ (Glynn & Whitt, 1992), to approach the asymptotic variance of the underlying Markov chain as m increases We refer to Jacob et al. (2017, § 3.1) for a more detailed discussion on the effects of k and m, and recall their proposed guideline of taking k to be a large quantile of the meeting time τ and m as a large multiple of k.

In practice, our proposed method involves simulating R pairs of coupled Markov chains $(X^{(r)}, Y^{(r)}) = (X_n^{(r)}, Y_n^{(r)})_{n\geqslant 0}$ $(r = 1, \ldots, R)$ completely in parallel, with each pair taking a random time to compute depending on their meeting time. As this produces R independent replicates $H_{k:m}(X^{(r)}, Y^{(r)})$ $(r = 1, \ldots, R)$ of the unbiased estimator (2), one can compute the average

$R^{-1}\sum_{r=1}^{R} H_{k:m}(X^{(r)}, Y^{(r)})$ to approximate $\pi(h)$. By appealing to the usual central limit theorem for independent and identically distributed random variables, confidence intervals that are justified as $R \to \infty$ can also be constructed.

Explicit constructions of coupled chains satisfying Assumptions 1–3 for Markov kernels K defined by Metropolis–Hastings algorithms and Gibbs samplers are given in Jacob et al. (2017, § 4) and Jacob et al. (2019). The focus of the present article is to propose a coupling strategy tailored to Hamiltonian Monte Carlo chains, so as to enable the use of the unbiased estimators (1) and (2). We will show in § 5 that this approach applies to realistic settings and retains the benefits of Hamiltonian Monte Carlo in terms of scaling with dimension.

2. HAMILTONIAN DYNAMICS

2.1. *Hamiltonian flows*

Suppose that the target distribution has the form $\pi(\mathrm{d}q) \propto \exp\{-U(q)\}\,\mathrm{d}q$, where the potential function $U : \mathbb{R}^d \to \mathbb{R}_+$ satisfies the following conditions.

Assumption 4 (Regularity and growth of potential). The potential U is twice continuously differentiable and its gradient $\nabla U : \mathbb{R}^d \to \mathbb{R}^d$ is globally β-Lipschitz, i.e., there exists $\beta > 0$ such that $|\nabla U(q) - \nabla U(q')| \leqslant \beta|q - q'|$ for all $q, q' \in \mathbb{R}^d$.

These conditions imply at most quadratic growth of the potential, or equivalently that the tails of the target distribution are no lighter than Gaussian.

We now introduce Hamiltonian flows on the phase space $\mathbb{R}^d \times \mathbb{R}^d$, which consists of position variables $q \in \mathbb{R}^d$ and momentum variables $p \in \mathbb{R}^d$. We will be concerned with a Hamiltonian function $\mathcal{E} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ of the form $\mathcal{E}(q,p) = U(q) + |p|^2/2$. We use the identity mass matrix here and will rely on preconditioning in § 5.4 to incorporate curvature properties of π. The time evolution of a particle $\{q(t), p(t)\}_{t\in\mathbb{R}_+}$ under Hamiltonian dynamics is described by the ordinary differential equations

$$\frac{\mathrm{d}}{\mathrm{d}t} q(t) = \nabla_p \mathcal{E}\{q(t), p(t)\} = p(t), \quad \frac{\mathrm{d}}{\mathrm{d}t} p(t) = -\nabla_q \mathcal{E}\{q(t), p(t)\} = -\nabla U\{q(t)\}. \tag{3}$$

Under Assumption 4, (3) with an initial condition $\{q(0), p(0)\} = (q_0, p_0) \in \mathbb{R}^d \times \mathbb{R}^d$ admits a unique solution globally on $\mathbb{R}_+$ (Lelièvre et al., 2010, p. 14). Therefore the flow map $\Phi_t(q_0, p_0) = \{q(t), p(t)\}$ is well-defined for any $t \in \mathbb{R}_+$, and we will write its projection onto the position and momentum coordinates as $\Phi_t^{\circ}(q_0, p_0) = q(t)$ and $\Phi_t^{*}(q_0, p_0) = p(t)$, respectively.

It is worth recalling that Hamiltonian flows have the following properties.

Property 1 (Reversibility). For any $t \in \mathbb{R}_+$, the inverse flow map satisfies $\Phi_t^{-1} = M \circ \Phi_t \circ M$, where $M(q,p) = (q, -p)$ denotes momentum reversal.

Property 2 (Energy conservation). The Hamiltonian function satisfies $\mathcal{E} \circ \Phi_t = \mathcal{E}$ for any $t \in \mathbb{R}_+$.

Property 3 (Volume preservation). For any $t \in \mathbb{R}_+$ and $A \in \mathcal{B}(\mathbb{R}^{2d})$, $\mathrm{Leb}_{2d}\{\Phi_t(A)\} = \mathrm{Leb}_{2d}(A)$, where Leb_{2d} denotes the Lebesgue measure on $\mathbb{R}^{2d}$.

These properties imply that the extended target distribution on phase space, $\tilde{\pi}(\mathrm{d}q, \mathrm{d}p) \propto \exp\{-\mathcal{E}(q,p)\}\,\mathrm{d}q\,\mathrm{d}p$, is invariant under the Markov semigroup induced by the flow; that is, for

any $t \in \mathbb{R}_+$ the pushforward measure $\Phi_t \sharp \tilde{\pi}$, defined by $\Phi_t \sharp \tilde{\pi}(A) = \tilde{\pi}\{\Phi_t^{-1}(A)\}$ for $A \in \mathcal{B}(\mathbb{R}^{2d})$, is equal to $\tilde{\pi}$.

2.2. *Coupled Hamiltonian dynamics*

We now consider the coupling of two particles $\{q^i(t), p^i(t)\}_{t\in\mathbb{R}_+}$ $(i = 1, 2)$ evolving under (3) with initial conditions $\{q^i(0), p^i(0)\} = (q_0^i, p_0^i)$ $(i = 1, 2)$. We first draw some insights from a Gaussian example.

Example 1. Let π be a Gaussian distribution on $\mathbb{R}$ with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. In this case we have $U(q) = (q-\mu)^2/(2\sigma^2)$ and $\nabla U(q) = (q-\mu)/\sigma^2$, and the solution of (3) is

$$\Phi_t(q_0, p_0) = \left\{ \begin{array}{c} \mu + (q_0 - \mu)\cos\left(\dfrac{t}{\sigma}\right) + \sigma p_0 \sin\left(\dfrac{t}{\sigma}\right) \\ p_0 \cos\left(\dfrac{t}{\sigma}\right) - \dfrac{1}{\sigma}(q_0 - \mu)\sin\left(\dfrac{t}{\sigma}\right) \end{array} \right\}.$$

Hence the difference between the particle positions is

$$q^1(t) - q^2(t) = (q_0^1 - q_0^2)\cos\left(\frac{t}{\sigma}\right) + \sigma(p_0^1 - p_0^2)\sin\left(\frac{t}{\sigma}\right).$$

If we set $p_0^1 = p_0^2$, then $|q^1(t) - q^2(t)| = |\cos(t/\sigma)|\,|q_0^1 - q_0^2|$; so for any nonnegative integer n, the particles meet exactly whenever $t = (2n+1)\pi\sigma/2$, and contraction occurs for any $t \neq \pi n\sigma$.

This example motivates a coupling that simply assigns particles the same initial momentum. Moreover, it also reveals that certain trajectory lengths will result in greater contraction than others. We now examine the utility of this approach more generally. Define $\Delta(t) = q^1(t) - q^2(t)$ to be the difference between particle locations and observe that

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|\Delta(t)|^2 = \Delta(t)^{\mathrm{T}}\{p^1(t) - p^2(t)\}.$$

Therefore, by imposing that $p^1(0) = p^2(0)$, the function $t \mapsto |\Delta(t)|$ admits a stationary point at time $t = 0$. This is geometrically intuitive, as the trajectories at time zero are parallel to one another for an infinitesimally small amount of time. To characterize this stationary point, we compute

$$\frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}|\Delta(t)|^2 = -\Delta(t)^{\mathrm{T}}\big[\nabla U\{q^1(t)\} - \nabla U\{q^2(t)\}\big] + |p^1(t) - p^2(t)|^2$$

and make the following assumption.

Assumption 5 (Local convexity of potential). There exists a compact set $S \in \mathcal{B}(\mathbb{R}^d)$, with positive Lebesgue measure, such that the restriction of U to S is α-strongly convex, i.e., there exists $\alpha > 0$ such that $(q - q')^{\mathrm{T}}\{\nabla U(q) - \nabla U(q')\} \geqslant \alpha|q - q'|^2$ for all $q, q' \in S$.

Under Assumption 5, we have

$$\frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}|\Delta(0)|^2 \leqslant -\alpha|\Delta(0)|^2 + |p^1(0) - p^2(0)|^2$$

if $q_0^1, q_0^2 \in S$ with $q_0^1 \neq q_0^2$. Therefore, by taking $p^1(0) = p^2(0)$, it follows from the second derivative test that $t = 0$ is a strict local maximum point. Continuity of $t \mapsto |\Delta(t)|^2$ implies that there exists a trajectory length $T > 0$ such that for any $t \in (0, T]$ there exists $\rho \in [0, 1)$ satisfying

$$\left|\Phi_t^\circ(q_0^1, p_0) - \Phi_t^\circ(q_0^2, p_0)\right| \leqslant \rho\, |q_0^1 - q_0^2|. \tag{4}$$

We note the dependence of T on the initial positions q_0^1 and q_0^2 and the momentum p_0. We now strengthen the above claim.

LEMMA 1. *Suppose that the potential U satisfies Assumptions 4 and 5. For any compact set $A \subset S \times S \times \mathbb{R}^d$, there exists a trajectory length $T > 0$ such that for any $t \in (0, T]$ there exists $\rho \in [0, 1)$ satisfying (4) for all $(q_0^1, q_0^2, p_0) \in A$.*

Although the qualitative result in Lemma 1 is sufficient for our purposes, we note that more quantitative results of this type have been established recently by Mangoubi & Smith (2017, Theorem 6) and Bou-Rabee et al. (2018, Theorem 2.1) to study the mixing time of the Hamiltonian Monte Carlo method. The preceding results show that the trajectory length T yielding contraction of the coupled system and the corresponding contraction rate ρ do not depend on d but only on the constants α and β in Assumptions 4 and 5. This suggests that such a coupling strategy can be effective in high dimensions as long as the Hessian of U is sufficiently well-conditioned.

3. COUPLED HAMILTONIAN MONTE CARLO

3.1. *Leap-frog integrator*

As the flow defined by (3) is typically intractable, time discretizations are required. The leap-frog symplectic integrator is a standard choice as it preserves Properties 1 and 3. Given a step size $\varepsilon > 0$ and a number of leap-frog steps $L \in \mathbb{N}$, this scheme initializes at $(q_0, p_0) \in \mathbb{R}^d \times \mathbb{R}^d$ and iterates

$$p_{\ell+1/2} = p_\ell - \frac{\varepsilon}{2}\nabla U(q_\ell), \quad q_{\ell+1} = q_\ell + \varepsilon p_{\ell+1/2}, \quad p_{\ell+1} = p_{\ell+1/2} - \frac{\varepsilon}{2}\nabla U(q_{\ell+1})$$

for $\ell = 0, \ldots, L-1$. We write the leap-frog iteration as $\hat{\Phi}_\varepsilon(q_\ell, p_\ell) = (q_{\ell+1}, p_{\ell+1})$ and the corresponding approximation of the flow as $\hat{\Phi}_{\varepsilon,\ell}(q_0, p_0) = (q_\ell, p_\ell)$ for $\ell = 0, \ldots, L$. As before, we denote by $\hat{\Phi}^\circ_{\varepsilon,\ell}(q_0, p_0) = q_\ell$ and $\hat{\Phi}^*_{\varepsilon,\ell}(q_0, p_0) = p_\ell$ the projections onto the position and momentum coordinates, respectively.

It can be established that the leap-frog scheme is of order two (Hairer et al., 2005, Theorem 3.4); that is, for sufficiently small ε,

$$\left|\hat{\Phi}_{\varepsilon,L}(q_0, p_0) - \Phi_{\varepsilon L}(q_0, p_0)\right| \leqslant C_3(q_0, p_0, L)\varepsilon^2, \tag{5}$$

$$\left|\mathcal{E}\{\hat{\Phi}_{\varepsilon,L}(q_0, p_0)\} - \mathcal{E}(q_0, p_0)\right| \leqslant C_4(q_0, p_0, L)\varepsilon^2 \tag{6}$$

for some positive constants C_3 and C_4 that depend continuously on the initial condition (q_0, p_0) for any number L of leap-frog iterations. To simplify our exposition we will assume throughout that (5) and (6) hold. We refer to the book on geometric numerical integration by Hairer et al. (2005) and to the survey by Bou-Rabee & Sanz-Serna (2018) for additional assumptions under which these error bounds hold.

We now discuss how the above constants behave with respect to dimension and integration length. Firstly, in the simplified setting of a target distribution with independent and identical

marginals and appropriate growth conditions on the potential, the results of Beskos et al. (2013, Propositions 5.3 and 5.4) indicate that these constants would scale as $d^{1/2}$. Hence, if we scale the step size ε as $d^{-1/4}$, advocated by Beskos et al. (2013) in this setting, we can expect these errors to be stable in high dimensions. Secondly, while the constant associated with the pathwise error bound (5) will typically grow exponentially with L (Leimkuhler & Matthews, 2015, § 2.2.3), the constant of the Hamiltonian error bound (6) can be stable over exponentially long time intervals εL (Hairer et al., 2005, Theorem 8.1). Although the Hamiltonian is not conserved exactly under time discretization, one can employ a Metropolis–Hastings correction as described in the following section.

3.2. *Coupled Hamiltonian Monte Carlo kernel*

Hamiltonian Monte Carlo (Duane et al., 1987; Neal, 1993) is a Metropolis–Hastings algorithm that targets π using time-discretized Hamiltonian dynamics as proposals. In view of § 2.2, we consider coupling two Hamiltonian Monte Carlo chains $(Q_n^1, Q_n^2)_{n\geqslant 0}$ by initializing $(Q_0^1, Q_0^2) \sim \bar{\pi}_0$ and evolving the chains jointly according to the following procedure.

Algorithm 1. Coupled Hamiltonian Monte Carlo step given (Q_{n-1}^1, Q_{n-1}^2).

Sample momentum $P_n^* \sim \mathcal{N}(0_d, I_d)$ and $U_n \sim \mathrm{Un}[0,1]$ independently.
For $i = 1, 2$:
  Set $(q_0^i, p_0^i) = (Q_{n-1}^i, P_n^*)$.
  Perform leap-frog integration to obtain $(q_L^i, p_L^i) = \hat{\Phi}_{\varepsilon,L}(q_0^i, p_0^i)$.
  If $U_n < \alpha\{(q_0^i, p_0^i), (q_L^i, p_L^i)\}$, set $Q_n^i = q_L^i$.
  Otherwise, set $Q_n^i = Q_{n-1}^i$.
Output (Q_n^1, Q_n^2).

Since the leap-frog integrator preserves Properties 1 and 3, the Metropolis–Hastings acceptance probability is

$$\alpha\{(q,p),(q',p')\} = \min\big[1,\ \exp\{\mathcal{E}(q,p) - \mathcal{E}(q',p')\}\big] \tag{7}$$

for $(q,p), (q',p') \in \mathbb{R}^d \times \mathbb{R}^d$. Iterating the above yields two marginal chains $(Q_n^1)_{n\geqslant 0}$ and $(Q_n^2)_{n\geqslant 0}$ that are π-invariant. Algorithm 1 amounts to running two Hamiltonian Monte Carlo chains with common random numbers; this has been considered in Neal (2017) to remove the burn-in bias, and in Mangoubi & Smith (2017) and Bou-Rabee et al. (2018) to analyse mixing properties.

We denote the associated coupled Markov transition kernel on the position coordinates by $\bar{K}_{\varepsilon,L}\{(q^1,q^2), A^1 \times A^2\}$ for $q^1, q^2 \in \mathbb{R}^d$ and $A^1, A^2 \in \mathcal{B}(\mathbb{R}^d)$. Marginally we have $\bar{K}_{\varepsilon,L}\{(q^1,q^2), A^1 \times \mathbb{R}^d\} = K_{\varepsilon,L}(q^1, A^1)$ and $\bar{K}_{\varepsilon,L}\{(q^1,q^2), \mathbb{R}^d \times A^2\} = K_{\varepsilon,L}(q^2, A^2)$, where $K_{\varepsilon,L}$ denotes the Markov transition kernel of the marginal Hamiltonian Monte Carlo chain. If we supplement Assumption 4 with the existence of a local minimum of U, then aperiodicity, Lebesgue irreducibility and Harris recurrence of $K_{\varepsilon,L}$ follow from Durmus et al. (2017, Theorem 2); see also Cances et al. (2007) and Livingstone et al. (2016) for previous related work. Hence ergodicity follows from Meyn & Tweedie (2009, Theorem 13.0.1), and Assumption 1 is met for test functions satisfying $\pi(h^{2+\kappa_1}) < \infty$ for some $\kappa_1 > 0$.

We will write the law of the coupled Hamiltonian Monte Carlo chain as $\mathrm{pr}_{\varepsilon,L}$ and use $E_{\varepsilon,L}$ to denote expectation with respect to $\mathrm{pr}_{\varepsilon,L}$. The following result establishes that the relaxed meeting time $\tau_\delta = \inf\{n \geqslant 0 : |Q_n^1 - Q_n^2| \leqslant \delta\}$, for any $\delta > 0$, has geometric tails.

THEOREM 1. *Suppose that the potential U satisfies Assumptions 4 and 5. Assume also that there exists $\tilde{\varepsilon} > 0$ such that for any $\varepsilon \in (0, \tilde{\varepsilon})$ and $L \in \mathbb{N}$, there exist a measurable function $V : \mathbb{R}^d \to [1, \infty)$, $\lambda \in (0, 1)$ and $b < \infty$ such that*

$$K_{\varepsilon,L}(V)(q) \leqslant \lambda V(q) + b \tag{8}$$

for all $q \in \mathbb{R}^d$, $\pi_0(V) < \infty$ and $\{q \in \mathbb{R}^d : V(q) \leqslant \ell_1\} \subseteq \{q \in S : U(q) \leqslant \ell_0\}$, for some $\ell_0 \in \{\inf_{q \in S} U(q), \sup_{q \in S} U(q)\}$ and $\ell_1 > 1$ satisfying $\lambda + 2b(1-\lambda)^{-1}(1+\ell_1)^{-1} < 1$. Then for any $\delta > 0$ there exist $\varepsilon_0 \in (0, \tilde{\varepsilon})$ and $L_0 \in \mathbb{N}$ such that for any $\varepsilon \in (0, \varepsilon_0)$ and $L \in \mathbb{N}$ satisfying $\varepsilon L < \varepsilon_0 L_0$,

$$\mathrm{pr}_{\varepsilon,L}(\tau_\delta > n) \leqslant C_0 \kappa_0^n \tag{9}$$

for some $C_0 \in \mathbb{R}_+$ and $\kappa_0 \in (0, 1)$ and for all integer $n \geqslant 0$.

The proof of Theorem 1 proceeds by first showing that the relaxed meeting can take place, in finitely many iterations, whenever both chains enter a region of the state space where the target distribution is strongly log-concave. As suggested in Neal (2017), one can expect good coupling behaviour if the chains spend enough time in this region of the state space; the second part of the proof makes this intuition precise by controlling excursions with the geometric drift condition (8). The latter can be established under additional assumptions on the potential U (Durmus et al., 2017, Theorem 9).

As Theorem 1 implies that the coupled chains can get arbitrarily close with sufficient frequency, one could potentially employ the unbiased estimation framework of Glynn & Rhee (2014), which introduces a truncation variable. To verify Assumption 2, which requires exact meetings, in the next section we combine the coupled Hamiltonian Monte Carlo kernel with another coupled kernel that is designed to trigger exact meetings when the two chains are close.

4. UNBIASED HAMILTONIAN MONTE CARLO

4.1. *Coupled random walk Metropolis–Hastings kernel*

Let K_σ denote the π-invariant Gaussian random walk Metropolis–Hastings kernel with proposal covariance $\sigma^2 I_d$. In the following we describe a coupling of $K_\sigma(x, \cdot)$ and $K_\sigma(y, \cdot)$ that results in exact meetings with high probability when $x, y \in \mathbb{R}^d$ are close (Johnson, 1998; Jacob et al., 2017) and σ is appropriately chosen.

We begin by sampling the proposals $X^* \sim \mathcal{N}(x, \sigma^2 I_d)$ and $Y^* \sim \mathcal{N}(y, \sigma^2 I_d)$ from the maximal coupling of these two Gaussian distributions (Jacob et al., 2017, § 4.1). Under the maximal coupling, the probability of $\{X^* \neq Y^*\}$ is equal to the total variation distance between the distributions $\mathcal{N}(x, \sigma^2 I_d)$ and $\mathcal{N}(y, \sigma^2 I_d)$. Analytical tractability in the Gaussian case allows us to write that distance as $\mathrm{pr}(2\sigma|Z| \leqslant \delta)$, where $Z \sim \mathcal{N}(0, 1)$ and $\delta = |x - y|$. By approximating the folded Gaussian cumulative distribution function (Pollard, 2005), we obtain

$$\mathrm{pr}(X^* = Y^*) = \mathrm{pr}(2\sigma|Z| > \delta) = 1 - (2\pi)^{-1/2} \frac{\delta}{\sigma} + O\left(\frac{\delta^2}{\sigma^2}\right) \tag{10}$$

as $\delta/\sigma \to 0$. Hence, to achieve $\mathrm{pr}(X^* = Y^*) = \theta$ for some desired probability θ, σ should be chosen approximately as $\delta/\{(2\pi)^{1/2}(1-\theta)\}$.

The proposed values X^* and Y^* are then accepted according to Metropolis–Hastings acceptance probabilities, i.e., if $U^* \leqslant \min\{1, \pi(X^*)/\pi(x)\}$ and $U^* \leqslant \min\{1, \pi(Y^*)/\pi(y)\}$,

respectively, where a common uniform random variable $U^* \sim \mathrm{Un}[0,1]$ is used for both chains. We denote the resulting coupled Markov transition kernel on $\{\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d) \times \mathcal{B}(\mathbb{R}^d)\}$ by $\bar{K}_\sigma$. If σ is small relative to the spread of the target distribution, the probability of accepting both proposals would be high. On the other hand, (10) shows that σ needs to be large compared to δ for the event $\{X^* = Y^*\}$ to occur with high probability. This leads to a trade-off; in practice, one can monitor acceptance probabilities of random walk Metropolis–Hastings chains from preliminary runs to get an idea of how small σ should be. Although most of the simulations in § 5 will use $\sigma = 10^{-3}$ as the default value, the sensitivity to the choice of σ of our proposed method will be investigated in § 5.3 and § 5.4.

4.2. *Combining coupled kernels*

We now combine the coupled Hamiltonian Monte Carlo kernel $\bar{K}_{\varepsilon,L}$ introduced in § 3.2 with the coupled random walk Metropolis–Hastings kernel $\bar{K}_\sigma$ from § 4.1, using the mixture

$$\bar{K}_{\varepsilon,L,\sigma}\{(x,y), A \times B\} = (1-\gamma)\bar{K}_{\varepsilon,L}\{(x,y), A \times B\} + \gamma \bar{K}_\sigma\{(x,y), A \times B\} \tag{11}$$

for $x, y \in \mathbb{R}^d$ and $A, B \in \mathcal{B}(\mathbb{R}^d)$, where $\gamma \in (0,1)$, $\varepsilon > 0$, $L \in \mathbb{N}$ and $\sigma > 0$ are appropriately chosen. The rationale for this choice is to enable exact meetings using the coupled random walk Metropolis–Hastings kernel when the chains are brought close together by the coupled Hamiltonian Monte Carlo kernel.

To address the choice of γ, in light of the efficiency considerations in § 1.3 we should understand how γ affects both the average meeting time, which we will investigate in § 5.3 and § 5.4, and the asymptotic inefficiency of the marginal kernel $K_{\varepsilon,L,\sigma} = (1-\gamma)K_{\varepsilon,L} + \gamma K_\sigma$. We now compare the asymptotic inefficiency of $K_{\varepsilon,L,\sigma}$ with that of $K_{\varepsilon,L}$. Assuming that evaluation of the potential and of its gradient have the same cost, the latter is given by the product of its cost, $L+2$, and its asymptotic variance, $v(h, K_{\varepsilon,L}) = \lim_{n\to\infty} \mathrm{var}_{\varepsilon,L}\{n^{-1/2}\sum_{i=1}^n h(X_i)\}$, where $X_0 \sim \pi$ and $X_n \sim K_{\varepsilon,L}(X_{n-1}, \cdot)$ for all integer $n \geqslant 1$. The expected cost of $K_{\varepsilon,L,\sigma}$ is $(1-\gamma)(L+2)+\gamma$, and we now consider its asymptotic variance $v(h, K_{\varepsilon,L,\sigma})$. By Peskun's ordering (Peskun, 1973), we have $v(h, K_{\varepsilon,L,\sigma}) \leqslant v(h, P_{\varepsilon,L})$ where $P_{\varepsilon,L} = (1-\gamma)K_{\varepsilon,L} + \gamma I$ with the identity kernel defined as $I(x,A) = \mathbb{I}_A(x)$ for $x \in \mathbb{R}^d$ and $A \in \mathcal{B}(\mathbb{R}^d)$. We then apply Łatuszyński & Roberts (2013, Corollary 1) to obtain $v(h, K_{\varepsilon,L,\sigma}) \leqslant \gamma(1-\gamma)^{-1}\,\mathrm{var}_\pi\{h(X)\} + (1-\gamma)^{-1}v(h, K_{\varepsilon,L})$. Hence, in summary, the relative asymptotic inefficiency can be bounded above by

$$\big\{1 + \gamma(1-\gamma)^{-1}(L+2)^{-1}\big\}\big[1 + \gamma\{1 + \Psi(h, K_{\varepsilon,L})\}^{-1}\big], \tag{12}$$

where $\Psi(h, K_{\varepsilon,L}) = 1 + 2\sum_{n=1}^\infty \mathrm{Corr}_{\varepsilon,L}\{h(X_0), h(X_n)\}$ denotes the integrated autocorrelation time of a stationary Hamiltonian Monte Carlo chain. In view of (12), we recommend choosing only small values of γ to reduce the loss of efficiency of the marginal chain; most of the simulations in § 5 will use $\gamma = 1/20$ as the default value.

We will write $Q_\sigma(x,A) = \int_A \mathcal{N}(y; x, \sigma^2 I_d)\,\mathrm{d}y$, with $x \in \mathbb{R}^d$ and $A \in \mathcal{B}(\mathbb{R}^d)$, for the Markov transition kernel of the Gaussian random walk; we will write the law of the resulting coupled chain $(X_n, Y_n)_{n\geqslant 0}$ as $\mathrm{pr}_{\varepsilon,L,\sigma}$ and use $E_{\varepsilon,L,\sigma}$ to denote expectation with respect to $\mathrm{pr}_{\varepsilon,L,\sigma}$. Algorithm 2 details the simulation of $(X_n, Y_n)_{n\geqslant 0}$ to compute the unbiased estimators described in § 1.3.

The mixture kernel $K_{\varepsilon,L,\sigma}$ inherits ergodicity properties from any of its components, so Assumption 1 can be satisfied following the discussion in § 3.2. Noting that the faithfulness property in Assumption 3 holds by construction, we now turn our attention to Assumption 2.

Algorithm 2. Computation of unbiased estimator $H_{k:m}(X, Y)$ of $\pi(h)$.

Initialize $(X_0, Y_0) \sim \bar{\pi}_0$ from a coupling with π_0 as marginals.
With probability γ, sample $X_1 \sim K_\sigma(X_0, \cdot)$; otherwise, sample $X_1 \sim K_{\varepsilon,L}(X_0, \cdot)$.
Set $n = 1$. While $n < \max(m, \tau)$:
    With probability γ, sample $(X_{n+1}, Y_n) \sim \bar{K}_\sigma\{(X_n, Y_{n-1}), \cdot\}$.
    Otherwise, sample $(X_{n+1}, Y_n) \sim \bar{K}_{\varepsilon,L}\{(X_n, Y_{n-1}), \cdot\}$.
    If $X_{n+1} = Y_n$, set $\tau = n + 1$.
    Increment $n \leftarrow n + 1$.
Compute $H_{k:m}(X, Y)$ using (2).

THEOREM 2. *Suppose that the potential U satisfies Assumptions 4 and 5. Assume also that there exist $\tilde{\varepsilon} > 0$ and $\tilde{\sigma} > 0$ such that for any $\varepsilon \in (0, \tilde{\varepsilon})$, $L \in \mathbb{N}$ and $\sigma \in (0, \tilde{\sigma})$, there exist a measurable function $V : \mathbb{R}^d \to [1, \infty)$, $\lambda \in (0, 1)$, $b < \infty$ and $\mu > 0$ such that*

$$K_{\varepsilon,L}(V)(x) \leqslant \lambda V(x) + b, \quad Q_\sigma(V)(x) \leqslant \mu\{V(x) + 1\}$$

for all $x \in \mathbb{R}^d$, $\pi_0(V) < \infty$, $\lambda_0 = (1 - \gamma)\lambda + \gamma(1 + \mu) < 1$ and $\{x \in \mathbb{R}^d : V(x) \leqslant \ell_1\} \subseteq \{x \in S : U(x) \leqslant \ell_0\}$, for some $\ell_0 \in \{\inf_{x\in S} U(x), \sup_{x\in S} U(x)\}$ and $\ell_1 > 1$ satisfying $\lambda_0 + 2\{(1 - \gamma)b + \gamma\mu\}(1 - \lambda_0)^{-1}(1 + \ell_1)^{-1} < 1$. Then there exist $\varepsilon_0 \in (0, \tilde{\varepsilon}), L_0 \in \mathbb{N}$ and $\sigma_0 > 0$ such that for any $\varepsilon \in (0, \varepsilon_0)$, $L \in \mathbb{N}$ satisfying $\varepsilon L < \varepsilon_0 L_0$ and $\sigma \in (0, \sigma_0)$, we have

$$\mathrm{pr}_{\varepsilon,L,\sigma}(\tau > n) \leqslant C_0 \kappa_0^n \tag{13}$$

for some $C_0 \in \mathbb{R}_+$ and $\kappa_0 \in (0, 1)$ and for all integers $n \geqslant 0$.

The proof of the above result proceeds in two parts, like the proof of Theorem 1, but requires slightly stronger assumptions to ensure that the mixture kernel still satisfies a geometric drift condition. The assumptions of Theorems 1 and 2 can be verified for target distributions given by multivariate Gaussian distributions and posterior distributions arising from Bayesian logistic regression; see the Supplementary Material. Although the above discussion guarantees the validity of the unbiased estimator computed by Algorithm 2 for a range of tuning parameters, its efficiency will depend on the distribution of the meeting time τ induced by the coupling and mixing properties of the marginal kernel $K_{\varepsilon,L,\sigma}$.

5. NUMERICAL ILLUSTRATIONS

5.1. *Preliminaries*

In practice, we will run Algorithm 2 R times independently in parallel to obtain the unbiased estimators $H_{k:m}(X^{(r)}, Y^{(r)})$ $(r = 1, \ldots, R)$. Following the framework of Glynn & Whitt (1992), we define the asymptotic inefficiency by $i(h, \bar{\pi}_0, \bar{K}_{\varepsilon,L,\sigma}) = E_{\varepsilon,L,\sigma}\{2(\tau - 1) + \max(1, m + 1 - \tau)\}\,\mathrm{var}_{\varepsilon,L,\sigma}\{H_{k:m}(X, Y)\}$, assuming that using $\bar{K}_{\varepsilon,L,\sigma}$ costs twice as much as $K_{\varepsilon,L,\sigma}$. This measure of efficiency accounts for the fact that, with a given computational budget, one can average over more estimators if each is cheaper to compute. We will approximate this inefficiency by empirical averages over the R realizations. For comparison, the asymptotic variance $v(h, K_{\varepsilon,L})$ of the standard Hamiltonian Monte Carlo estimator will be approximated with the `spectrum0.ar` function in the `R` (R Development Core Team, 2019) package `coda` (Plummer et al., 2006), using 10 000 iterations after a burn-in of 1000 for all examples. We will consider estimation

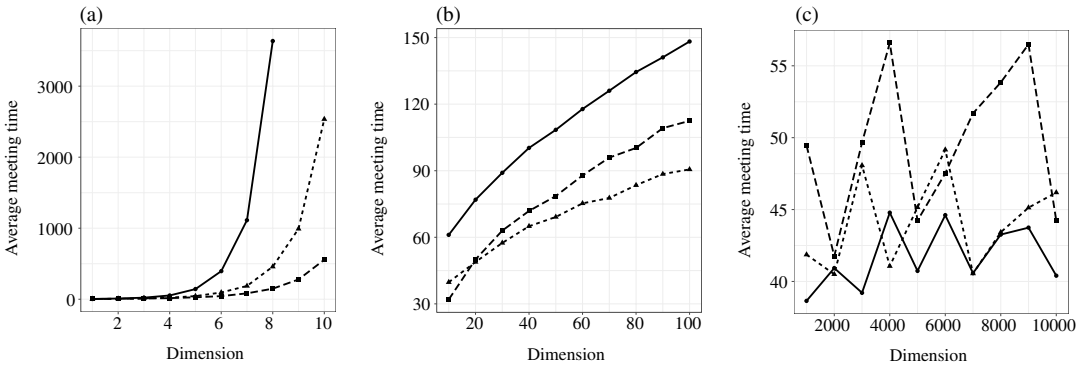

Fig. 1. Gaussian example of § 5.2: scaling of the average meeting time with dimension for 1000 coupled chains based on (a) random walk Metropolis–Hastings, (b) the Metropolis-adjusted Langevin algorithm, and (c) Hamiltonian Monte Carlo. In each panel the lines and symbols correspond to $C = 1$ (solid line with circles), $C = 1.5$ (dotted line with triangles) and $C = 2$ (dashed line with squares).

of first and second moments; that is, we set $h_i(x) = x_i$ and $h_{d+i}(x) = x_i^2$ for $i = 1, \ldots, d$, and compare $i(\bar{\pi}_0, \bar{K}_{\varepsilon,L,\sigma}) = \sum_{i=1}^{2d} i(h_i, \bar{\pi}_0, \bar{K}_{\varepsilon,L,\sigma})$ with $v(K_{\varepsilon,L}) = \sum_{i=1}^{2d} v(h_i, K_{\varepsilon,L})$ at possibly different parameter configurations. An important point to be illustrated in the following is that the parameters ε and L minimizing the asymptotic inefficiency $(L+2)v(K_{\varepsilon,L})$ may not necessarily be suitable for our proposed estimator. Lastly, we will apply the guideline of taking k as the 90% sample quantile of meeting times, obtained from a small number of preliminary runs, and setting $m = 10k$.

5.2. *Toy examples*

We first investigate the scalability of the proposed approach in high dimensions on a standard Gaussian target distribution on $\mathbb{R}^d$, by examining the average meeting time of stationary coupled chains generated by (11). For simplicity, $\sigma = 10^{-3}$ and $\gamma = 1/20$ are taken as the parameters' default values. To ensure stable acceptance probabilities as $d \to \infty$ (Beskos et al., 2013), we scale the step size as $\varepsilon = Cd^{-1/4}$ and select different constants $C > 0$ to induce a range of acceptance probabilities. The number of leap-frog steps is taken to be $L = 1 + \lfloor \varepsilon^{-1} \rfloor$, which fixes the integration time εL at approximately 1. For comparison, we consider (11) with $L = 1$, as this corresponds to the Metropolis-adjusted Langevin algorithm, and adopt the scaling $\varepsilon^2 = C^2 d^{-1/3}$ (Roberts & Rosenthal, 1998); see the Supplementary Material for an alternative coupling. Lastly, we also consider coupled chains generated solely by the coupled random walk Metropolis–Hastings kernel described in § 4.1, with proposal variance scaling as $\sigma^2 = C^2 d^{-1}$ (Roberts et al., 1997). The results displayed in Fig. 1 demonstrate the effectiveness of our coupling strategy in high dimensions, and illustrate the appeal of Hamiltonian Monte Carlo kernels in high-dimensional settings.

Next we consider a banana-shaped target distribution on $\mathbb{R}^2$, whose potential is given by the Rosenbrock function $U(x_1, x_2) = (1 - x_1)^2 + 10(x_2 - x_1^2)^2$ for $(x_1, x_2) \in \mathbb{R}^2$. The aim here is to examine the utility of our proposed coupling for a highly nonconvex potential and to explore the use of a new coupling for Hamiltonian Monte Carlo introduced in Bou-Rabee et al. (2018, § 2.3.2). In contrast to Algorithm 1, which assigns the same initial momentum to both chains, the latter samples an initial momentum $P_n^1 \sim \mathcal{N}(0_d, I_d)$ for the first chain and sets the initial

momentum for the second chain to

$$P_n^2 = \begin{cases} P_n^1 + \kappa \Delta_{n-1} & \min\left\{1, \dfrac{\mathcal{N}(\bar{\Delta}_{n-1}^{\mathrm{T}} P_n^1 + \kappa |\Delta_{n-1}|; 0, 1)}{\mathcal{N}(\bar{\Delta}_{n-1}^{\mathrm{T}} P_n^1; 0, 1)}\right\}, \\ P_n^1 - 2(\bar{\Delta}_{n-1}^{\mathrm{T}} P_n^1)\bar{\Delta}_{n-1} & \text{otherwise,} \end{cases}$$

where $\kappa > 0$ is a tuning parameter, $\Delta_{n-1} = Q_{n-1}^1 - Q_{n-1}^2$ is the difference between the chains at iteration $n-1$, and $\bar{\Delta}_{n-1} = \Delta_{n-1}/|\Delta_{n-1}|$ is the normalized difference. Leap-frog integration and Metropolis–Hastings acceptance of the output are then performed in the same way as in Algorithm 1; the resulting coupled Hamiltonian Monte Carlo kernel is then employed in the mixture (11). We simulate 1000 coupled chains initialized independently from the uniform distribution on $[-5, 5]^2$, using this new coupling with $\kappa = 1$ and the previous one which corresponds to $\kappa = 0$. Employing the same parameters $(\varepsilon, L, \sigma, \gamma) = (1/500, 500, 10^{-3}, 1/20)$ for both couplings, we observe that the new coupling reduces the average meeting time from 158 to 52. This example demonstrates that the proposed method can be used beyond convex potentials and that alternative couplings can result in significantly shorter meeting times.

5.3. *Logistic regression*

We now consider a Bayesian logistic regression on the classic German credit dataset, as in Hoffman & Gelman (2014). After including all pairwise interactions and performing standardization, the design matrix has 1000 rows and 300 columns. Given covariates $x_i \in \mathbb{R}^{300}$, intercept $a \in \mathbb{R}$ and coefficients $b \in \mathbb{R}^{300}$, each observation $y_i \in \{0, 1\}$ is modelled as an independent Bernoulli random variable with probability of success $\{1 + \exp(-a - b^{\mathrm{T}} x_i)\}^{-1}$. The prior is specified as $a \mid s^2 \sim \mathcal{N}(0, s^2)$, $b \mid s^2 \sim \mathcal{N}(0_{300}, s^2 I_{300})$ independently, where the variance parameter s^2 follows an exponential distribution with rate 0.01. The target π is the posterior distribution of parameters $(a, b, \log s^2)$ on $\mathbb{R}^d$ with $d = 302$.

Initializing coupled chains independently from $\pi_0 = \mathcal{N}(0_d, I_d)$, for each parameter configuration $(\varepsilon, L) \in \{0.01, 0.0125, \ldots, 0.04\} \times \{10, 20, 30\}$ we run five pairs of coupled Hamiltonian Monte Carlo chains for 1000 iterations. This computation can be done independently in parallel for each configuration and repetition; the output is displayed in Fig. 2(a). Although multiple configurations lead to contractive chains, this is not the case for $(\varepsilon, L) = (0.03, 10)$, which are optimal parameter values for Hamiltonian Monte Carlo. For configurations that yield distances less than 10^{-10}, we simulate 100 meeting times in parallel using the mixture kernel (11) with $\sigma = 10^{-3}$ and $\gamma = 1/20$. We then select the parameter configuration $(\varepsilon, L) = (0.0125, 10)$ that has the smallest average computational cost, taken as $L + 2$ times the average meeting time.

To illustrate the effects of σ and γ, we fix $(\varepsilon, L) = (0.0125, 10)$ and examine the distribution of meeting times as σ or γ varies. Decreasing σ leads to larger meeting times: conservatively small values of σ require more iterations before the chains get close enough for the maximal coupling to propose the same value with high probability. On the other hand, if σ is too large, large meeting times are observed as random walk proposals would be rejected with high probability. Figure 2(b) suggests that the effectiveness of our coupling is not highly sensitive to the choice of σ, provided that it is small enough. Similarly, Fig. 2(c) shows stable meeting times for the range of values of γ considered.

Finally, we produce $R = 1000$ coupled chains in parallel with $(\varepsilon, L, \sigma, \gamma) = (0.0125, 10, 10^{-3}, 1/20)$ and compare the inefficiency of our estimator with the asymptotic variance of the optimal Hamiltonian Monte Carlo estimator for various choices of k and m. The results,

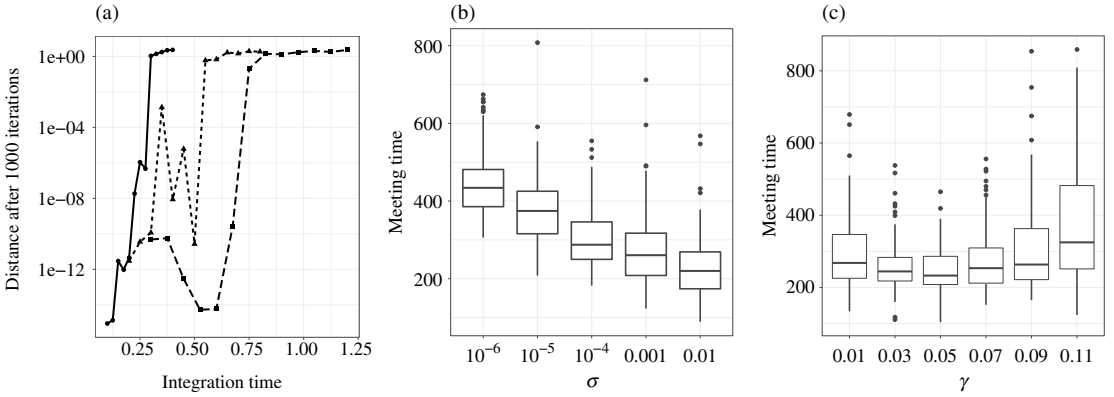

Fig. 2. Logistic regression example of § 5.3: (a) average distance between coupled chains at iteration 1000 plotted against integration time εL, for $L = 10$ (solid line with circles), $L = 20$ (dotted line with triangles) and $L = 30$ (dashed line with squares); boxplots of meeting times as the parameters (b) σ and (c) γ vary.

Table 1. *Relative inefficiency of the proposed estimator in the logistic regression example; the cost, variance and relative inefficiency were computed using* $R = 1000$ *independent runs, while the median and* 90% *quantile of the meeting time were computed with* 100 *preliminary runs*

k	m	Cost	Variance	Rel. ineff.
1	k	436	4.0×10^{2}	1989.07
1	$5k$	436	3.4×10^{2}	1671.93
1	$10k$	436	2.8×10^{2}	1403.28
median(τ)	k	458	7.4×10^{0}	38.22
median(τ)	$5k$	1258	1.1×10^{-1}	1.58
median(τ)	$10k$	2298	4.5×10^{-2}	1.18
90% quantile(τ)	k	553	6.0×10^{0}	38.11
90% quantile(τ)	$5k$	1868	5.8×10^{-2}	1.23
90% quantile(τ)	$10k$	3518	2.6×10^{-2}	1.05

Cost, the expected computational cost; Variance, the sum of variances when estimating first and second moments; Rel. ineff., relative inefficiency, the ratio of the asymptotic inefficiency $i(\bar{\pi}_0, \bar{K}_{\varepsilon,L,\sigma})$ with parameters $(\varepsilon, L, \sigma, \gamma) = (0.0125, 10, 10^{-3}, 1/20)$ to the asymptotic variance $v(K_{\varepsilon,L})$ with optimal parameters $(\varepsilon, L) = (0.03, 10)$.

summarized in Table 1, show that bias removal comes at a cost of increased variance, and that this can be reduced with appropriate choices of k and m. Our guideline for choosing k and m results in a relative inefficiency of 1.05 at an average computational cost of 3518 applications of $K_{\varepsilon,L,\sigma}$, or approximately 5 minutes of computing time with our implementation. Therefore, thanks to unbiasedness, we can safely average over independent copies of an estimator whose expected cost is of the order of a few thousand Hamiltonian Monte Carlo iterations.

5.4. *Log-Gaussian Cox point processes*

We end with a challenging high-dimensional application of Bayesian inference for log-Gaussian Cox point processes on a dataset containing the locations of 126 Scots pine saplings in a natural forest in Finland (Møller et al., 1998). After discretizing the plot into an $n \times n$ regular grid, the number of points $y_i \in \mathbb{N}$ in each grid cell is assumed to be conditionally independent given a latent intensity process Λ_i, $i \in \{1, \ldots, n\}^2$, and is modelled as Poisson distributed with mean

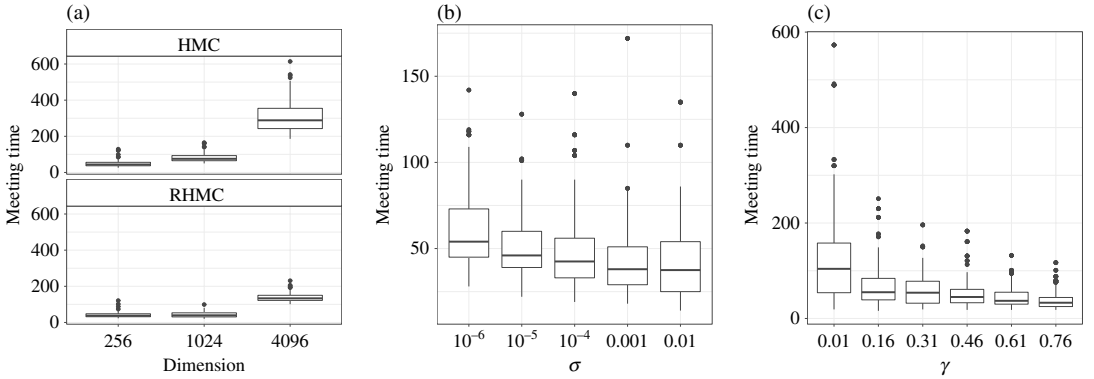

Fig. 3. Cox process example of § 5.4: boxplots of meeting times (a) for the Hamiltonian Monte Carlo, HMC, and Riemann manifold Hamiltonian Monte Carlo, RHMC, algorithms and for all three discretizations, and as the parameters (b) σ and (c) γ vary.

$a\Lambda_i$, where $a = n^{-2}$ is the area of each grid cell. The prior is specified by $\Lambda_i = \exp(X_i)$, where X_i, $i \in \{1, \ldots, n\}^2$, is a Gaussian process with mean $\mu \in \mathbb{R}$ and exponential covariance function $\Sigma_{i,j} = s^2 \exp\{-|i-j|/(nb)\}$ for $i, j \in \{1, \ldots, n\}^2$. We will use the parameter values $s^2 = 1.91$, $b = 1/33$ and $\mu = \log(126) - s^2/2$ estimated by Møller et al. (1998) and infer the posterior distribution of the latent process X_i, $i \in \{1, \ldots, n\}^2$, given the count data and these hyperparameter values. We consider three discretizations, with $n \in \{16, 32, 64\}$, which correspond to target distributions π on $\mathbb{R}^d$ with $d \in \{256, 1024, 4096\}$.

Owing to the high dimensionality of this model, the mixing of random walk Metropolis–Hastings is known to be prohibitively slow (Christensen & Waagepetersen, 2002), while the Metropolis-adjusted Langevin algorithm requires a computationally costly reparameterization to be effective (Christensen et al., 2005). We will consider the use of Hamiltonian Monte Carlo and Riemann manifold Hamiltonian Monte Carlo with metric tensor $\Sigma^{-1} + a \exp(\mu + s^2/2) I_d$ (Girolami & Calderhead, 2011). We proceed as in § 5.3 and seek parameter configurations $(\varepsilon, L) \in \{0.05, 0.07, \ldots, 0.45\} \times \{10, 20, 30\}$ that yield contractive coupled chains with small computational cost, when initialized independently from the prior distribution. Although both algorithms have multiple configurations that result in contractive chains, the parameters ε and L that were optimal for these methods only led to contractive coupled Riemann manifold Hamiltonian Monte Carlo chains for all three discretizations. By simulating 100 meeting times with $\sigma = 10^{-3}$ and $\gamma = 1/20$ for configurations yielding distances of less than 10^{-10}, for $d \in \{256, 1024, 4096\}$ we select $(\varepsilon, L) \in \{(0.11, 10), (0.15, 10), (0.17, 10)\}$ respectively for the Hamiltonian Monte Carlo method and $(\varepsilon, L) \in \{(0.11, 10), (0.11, 10), (0.13, 10)\}$ for the Riemann manifold Hamiltonian Monte Carlo method, which gave the smallest average computational cost for each algorithm. The corresponding meeting times plotted in Fig. 3(a) show the effectiveness of our coupling strategy even in high dimensions. Panels (b) and (c) of Fig. 3, which display the meeting times of coupled Riemann manifold Hamiltonian Monte Carlo chains for the finest discretization, also illustrate the robustness of our coupling with respect to the choices of σ and γ.

With the above parameters and our guideline for choosing k and m, we computed $R = 1000$ coupled chains in parallel for each algorithm and discretization. For $d = 256, 1024$ and 4096, the relative inefficiency was found to be 11.00, 5.43 and 2.73 respectively for the Hamiltonian Monte Carlo method, and 11.68, 7.85 and 3.72 for the Riemann manifold Hamiltonian Monte Carlo method. For the finest discretization, the average computational times were approximately 90 and 20 minutes with our implementation. Despite some loss of efficiency, the benefit of exploiting

parallel computation for this problem is apparent since one can only run respectively 4439 and 714 iterations of these algorithms for the same time.

6. Discussion

Construction of couplings could be explored for other variants of the Hamiltonian Monte Carlo method, such as those involving partial momentum refreshment (Horowitz, 1991), adaptation of tuning parameters (Hoffman & Gelman, 2014), different choices of kinetic energy (Livingstone et al., 2019), and combinations with new sampling paradigms (Pollock et al., 2017; Fearnhead et al., 2018; Vanetti et al., 2018). Other ways of leveraging parallel hardware for Hamiltonian Monte Carlo include the work of Calderhead (2014), which builds on the 2004 Norwegian University of Science and Technology technical report by H. Tjelmeland and focuses on parallel computation at each iteration of the algorithm.

Acknowledgement

The computations in this article were run on the Odyssey cluster supported by the Division of Science Research Computing Group at Harvard University. Jacob acknowledges support from the U.S. National Science Foundation. Both authors were supported by the U.S. Army Research Office. The authors are grateful to Oren Mangoubi, Radford Neal, Aaron Smith, and the reviewers for their insightful feedback.

Supplementary material

Supplementary material available at *Biometrika* online includes an alternative coupling for the Metropolis-adjusted Langevin algorithm, additional simulation results on truncated Gaussian distributions, and the proofs of Lemma 1 and Theorems 1 and 2. An `R` package implementing the numerical results is available at `https://github.com/pierrejacob/debiasedhmc`.

References

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J. M. & Stuart, A. (2013). Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli* **19**, 1501–34.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv:* 1701.02434.

Betancourt, M., Byrne, S., Livingstone, S. & Girolami, M. (2017). The geometric foundations of Hamiltonian Monte Carlo. *Bernoulli* **23**, 2257–98.

Bou-Rabee, N., Eberle, A. & Zimmer, R. (2018). Coupling and convergence for Hamiltonian Monte Carlo. *arXiv:* 1805.00452.

Bou-Rabee, N. & Sanz-Serna, J. M. (2018). Geometric integrators and the Hamiltonian Monte Carlo method. *Acta Numer.* **27**, 113–206.

Brooks, S. P., Gelman, A., Jones, G. & Meng, X. L. (2011). *Handbook of Markov Chain Monte Carlo*. Boca Raton, Florida: CRC Press.

Calderhead, B. (2014). A general construction for parallelizing Metropolis–Hastings algorithms. *Proc. Nat. Acad. Sci.* **111**, 17408–13.

Cances, E., Legoll, F. & Stoltz, G. (2007). Theoretical and numerical comparison of some sampling methods for molecular dynamics. *ESAIM Math. Mod. Numer. Anal.* **41**, 351–89.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P. & Riddell, A. (2016). Stan: A probabilistic programming language. *J. Statist. Software* **20**, 1–37.

Casella, G., Lavine, M. & Robert, C. P. (2001). Explaining the perfect sampler. *Am. Statistician* **55**, 299–305.

Christensen, O. F., Roberts, G. O. & Rosenthal, J. S. (2005). Scaling limits for the transient phase of local Metropolis–Hastings algorithms. *J. R. Statist. Soc.* B **67**, 253–68.

Christensen, O. F. & Waagepetersen, R. (2002). Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics* **58**, 280–6.

Duane, S., Kennedy, A. D., Pendleton, B. J. & Roweth, D. (1987). Hybrid Monte Carlo. *Phys. Lett.* B **195**, 216–22.

DURMUS, A., MOULINES, E. & SAKSMAN, E. (2017). On the convergence of Hamiltonian Monte Carlo. *arXiv:* 1705.00166.
FEARNHEAD, P., BIERKENS, J., POLLOCK, M. & ROBERTS, G. O. (2018). Piecewise deterministic Markov processes for continuous-time Monte Carlo. *Statist. Sci.* **33**, 386–412.
GIROLAMI, M. & CALDERHEAD, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Statist. Soc.* B **73**, 123–214.
GLYNN, P. W. & HEIDELBERGER, P. (1991). Analysis of parallel replicated simulations under a completion time constraint. *ACM Trans. Mod. Comp. Simul.* **1**, 3–23.
GLYNN, P. W. & RHEE, C.-H. (2014). Exact estimation for Markov chain equilibrium expectations. *J. Appl. Prob.* **51**, 377–89.
GLYNN, P. W. & WHITT, W. (1992). The asymptotic efficiency of simulation estimators. *Oper. Res.* **40**, 505–20.
HAIRER, E., WANNER, G. & LUBICH, C. (2005). *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*. New York: Springer.
HOFFMAN, M. D. & GELMAN, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15**, 1593–623.
HOROWITZ, A. M. (1991). A generalized guided Monte Carlo algorithm. *Phys. Lett.* B **268**, 247–52.
HUBER, M. (2016). *Perfect Simulation*, vol. 148 of *Monographs on Statistics & Applied Probability*. Boca Raton, Florida: CRC Press.
JACOB, P. E., LINDSTEN, F. & SCHÖN, T. B. (2017). Smoothing with couplings of conditional particle filters. *arXiv:* 1701.02002.
JACOB, P. E., O'LEARY, J. & ATCHADÉ, Y. F. (2019). Unbiased Markov chain Monte Carlo with couplings. *arXiv:* 1708.03625v2.
JOHNSON, V. E. (1998). A coupling-regeneration scheme for diagnosing convergence in Markov chain Monte Carlo algorithms. *J. Am. Statist. Assoc.* **93**, 238–48.
ŁATUSZYŃSKI, K. & ROBERTS, G. O. (2013). CLTs and asymptotic variance of time-sampled Markov chains. *Methodol. Comp. Appl. Prob.* **15**, 237–47.
LEIMKUHLER, B. & MATTHEWS, C. (2015). *Molecular Dynamics.* New York: Springer.
LELIÈVRE, T., ROUSSET, M. & STOLTZ, G. (2010). *Free Energy Computations: A Mathematical Perspective*. London: Imperial College Press.
LIVINGSTONE, S., BETANCOURT, M., BYRNE, S. & GIROLAMI, M. (2016). On the geometric ergodicity of Hamiltonian Monte Carlo. *arXiv:* 1601.08057.
LIVINGSTONE, S., FAULKNER, M. F. & ROBERTS, G. O. (2019). Kinetic energy choice in Hamiltonian/hybrid Monte Carlo. *Biometrika* to appear.
MANGOUBI, O. & SMITH, A. (2017). Rapid mixing of Hamiltonian Monte Carlo on strongly log-concave distributions. *arXiv:* 1708.07114.
MEYN, S. P. & TWEEDIE, R. L. (2009). *Markov Chains and Stochastic Stability*. Cambridge: Cambridge University Press, 2nd ed.
MØLLER, J., SYVERSVEEN, A. R. & WAAGEPETERSEN, R. P. (1998). Log Gaussian Cox processes. *Scand. J. Statist.* **25**, 451–82.
MYKLAND, P., TIERNEY, L. & YU, B. (1995). Regeneration in Markov chain samplers. *J. Am. Statist. Assoc.* **90**, 233–41.
NEAL, R. M. (1993). Bayesian learning via stochastic dynamics. In *Advances in Neural Information Processing Systems 5 (NIPS 1992)*. San Francisco: Morgan Kaufmann, pp. 475–82.
NEAL, R. M. (2017). Circularly-coupled Markov chain sampling. *arXiv:* 1711.04399.
PESKUN, P. H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika* **60**, 607–12.
PLUMMER, M., BEST, N., COWLES, K. & VINES, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News* **6**, 7–11.
POLLARD, D. (2005). Total variation distance between measures. In *Asymptopia*, chap. 3.
POLLOCK, M., FEARNHEAD, P., JOHANSEN, A. M. & ROBERTS, G. O. (2017). The scalable Langevin exact algorithm: Bayesian inference for big data. *arXiv:* 1609.03436v2.
R DEVELOPMENT CORE TEAM (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.
ROBERTS, G. O., GELMAN, A. & GILKS, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Prob.* **7**, 110–20.
ROBERTS, G. O. & ROSENTHAL, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Statist. Soc.* B **60**, 255–68.
ROSENTHAL, J. S. (1997). Faithful couplings of Markov chains: now equals forever. *Adv. Appl. Math.* **18**, 372–81.
ROSENTHAL, J. S. (2000). Parallel computing and Monte Carlo algorithms. *Far East J. Theor. Statist.* **4**, 207–36.
VANETTI, P., BOUCHARD-CÔTÉ, A., DELIGIANNIDIS, G. & DOUCET, A. (2018). Piecewise deterministic Markov chain Monte Carlo. *arXiv:* 1707.05296v2.

[*Received on* 13 *February* 2018. *Editorial decision on* 13 *August* 2018]