Couplings for Multinomial Hamiltonian Monte Carlo

Kai Xu[*] University of Edinburgh
Tor Erlend Fjelde[*] University of Cambridge
Charles Sutton University of Edinburgh
Hong Ge University of Cambridge

Abstract

Hamiltonian Monte Carlo (HMC) is a popular sampling method in Bayesian inference. Recently, Heng & Jacob (2019) studied Metropolis HMC with couplings for unbiased Monte Carlo estimation, establishing a generic parallelizable scheme for HMC. However, in practice a different HMC method, multinomial HMC, is considered as the go-to method, e.g. as part of the no-U-turn sampler. In multinomial HMC, proposed states are not limited to end-points as in Metropolis HMC; instead points along the entire trajectory can be proposed. In this paper, we establish couplings for multinomial HMC, based on optimal transport for multinomial sampling in its transition. We prove an upper bound for the meeting time – the time it takes for the coupled chains to meet – based on the notion of local contractivity. We evaluate our methods using three targets: 1,000 dimensional Gaussians, logistic regression and log-Gaussian Cox point processes. Compared to Heng & Jacob (2019), coupled multinomial HMC generally attains a smaller meeting time, and is more robust to choices of step sizes and trajectory lengths, which allows re-use of existing adaptation methods for HMC. These improvements together paves the way for a wider and more practical use of coupled HMC methods.

1 Introduction

Markov chain Monte Carlo (MCMC) is a standard tool to draw samples from target distributions known up to a normalising constant (Metropolis et al., 1953; Geman and Geman, 1984). Such samples are commonly used to estimate an integral of interest. Specifically, for a probability distribution π on $\mathbb{R}^d$ and a measurable function of interest $h : \mathbb{R}^d \mapsto \mathbb{R}$, we want to estimate

$$H = \int \pi(x)h(x)\mathrm{d}x = \mathbb{E}_{x\sim\pi}\left[h(x)\right]. \tag{1}$$

Approximating this integral H in (1) is at the core of many statistics and machine learning problems. For example in Bayesian inference, Monte Carlo samples are used to estimate some posterior predictive distribution, or perform model comparison (Gelman et al., 2013). Or in energy-based modelling, MCMC samples are to estimate gradients used to update model parameters (Teh et al., 2003; Xie et al., 2016; Qiu et al., 2019).

Under the framework of MCMC, a Markov chain is simulated to obtain correlated samples from π, and then Monte Carlo integration is used to estimate H using these samples. However, such estimators are unbiased only when the underlying Markov chain has converged to the equilibrium, which is challenging to verify in practice. Therefore MCMC with couplings has attracted research attention recently thanks to its ability to debias Monte Carlo estimators (Jacob et al., 2020). In particular, Heng and Jacob (2019) focused on the Metropolis-Hastings (MH) adjusted HMC variant, which proposes the end-point of a simulated Hamiltonian trajectory as the new state, followed by an MH correction step. We refer to this HMC variant as *coupled Metropolis HMC*. Heng and Jacob (2019) noticed that coupled Metropolis HMC is sensitive to the choice of HMC parameters such as integrator step sizes and Hamiltonian trajectory lengths. More specifically, parameters (e.g. trajectory lengths) optimal for sampling efficiency (e.g. effective sample size) can require a large number of HMC iterations to achieve meeting; on the other hand, optimal parameters for coupling can lead to poor mixing (Heng and Jacob, 2019).

Building upon the recent work of Heng and Jacob (2019), we propose two novel couplings based on a different, more robust implementation of HMC. We refer to our methods as *coupled multinomial HMC* and demonstrate several advantages of these methods. First, coupled multinomial HMC meets faster in general. Intuitively, like all MH algorithms, the previous coupled

[*] equal contributions

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

Algorithm 1: Sample a pair of coupled chains

Sample $(X_0, Y_0) \sim \bar{\pi}_0$ ($\bar{\pi}_0$ is a coupling of π_0);
Sample $X_1 \sim \mathcal{K}(X_0, \cdot)$;
Set $N = 1$;
while $X_N \neq Y_{N-1}$ **do**
  Sample $(X_{N+1}, Y_N) \sim \bar{\mathcal{K}}((X_N, Y_{N+1}), \cdot)$;
  Set $N = N + 1$;
Set $\tau = N$ and output $\{(X_n)_{n=0}^{\tau}, (Y_n)_{n=0}^{\tau-1}\}$;

HMC method can only propose a point from the initial or the last integration step, which leads to two drawbacks for couplings: (i) it may well be that intermediate points are the closest between two chains and (ii) rejection rates of proposals are quite sensitive to step sizes of Hamiltonian dynamics solvers. Multinomial coupling allows coupled chains to accept intermediate points that are potentially closer together, so they meet quicker. We therefore design couplings to minimize the expected distance between coupled chains within each transition to encourage faster meeting. Second, coupled multinomial HMC is less sensitive to Hamiltonian integration step sizes. For Metropolis HMC, a small enough step size has to be used to ensure a large enough acceptance probability in the MH adjustment step. However, for multinomial HMC, intermediate points can be proposed even though the end-points would have been rejected in Metropolis HMC. We argue that this robustness is crucial for practical use of coupled HMC algorithms. Thirdly, we prove that the meeting time of coupled multinomial HMC decays geometrically, which is a sufficient condition to use the unbiased estimator from Jacob et al. (2020). Finally, we perform extensive simulations to verify the improved meeting and robustness of our proposed method.

2 Background

2.1 Unbiased MCMC with couplings

For two distributions p and q, we denote $\Gamma(p, q)$ as their *couplings*, i.e. for any $\gamma \in \Gamma(p, q)$, the marginals of γ are p and q. For a π-invariant Markov kernel $\mathcal{K}$ defined on $\left(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\right)$, its coupled kernel $\bar{\mathcal{K}}$, defined on $\left(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d) \times \mathcal{B}(\mathbb{R}^d)\right)$, by construction has $\mathcal{K}$ as its marginals, where $\mathcal{B}$ denotes the Borel σ-algebra. Additionally, given an initial distribution π_0 and some $\bar{\pi}_0 \in \Gamma(\pi_0, \pi_0)$, a pair of coupled chains $X = (X_n)_{n\geq 0}$, $Y = (Y_n)_{n\geq 0}$ that share the same equilibrium distribution π can be simulated by Algorithm 1 (Jacob et al., 2020) until meeting at iteration $\tau := \inf\{n \geq 1 : X_n = Y_{n-1}\}$ (the *meeting time*). The main design choice in this algorithm is the construction of $\bar{\mathcal{K}}$. Jacob et al. (2020) established that if $\bar{\mathcal{K}}$ satisfies certain conditions, a pair of coupled chains X, Y from Algorithm 1 can be used to obtain unbiased estimates of (1) with finite variance and finite computation cost as

$$H_k(X, Y) = h(X_k) + \sum_{n=k+1}^{\tau-1} (h(X_n) - h(Y_{n-1})) \quad (2)$$

where $k \in \mathbb{N}$ is a parameter to choose. The first term in (2) is a standard, single-sample MCMC estimate and the second term can be seen as a debiasing term to correct the bias introduced by non-converged chains. This estimator is built on the pioneering works from Glynn and Rhee (2014), derived using telescoping sums. In practice, we use a time-averaged version of (2), which is still unbiased but with lower variance, e.g. in Section 5.

2.2 Hamiltonian Monte Carlo

In an HMC kernel, new states are proposed by simulating Hamiltonian dynamics (Neal, 2011). For a Hamiltonian system with a position variable $q \in \mathbb{R}^d$ and a momentum variable $p \in \mathbb{R}^d$, the trajectory $\mathbf{t} := (q(t), p(t))_{t\in\mathbb{R}_+}$ can be described by the following ordinary differential equations

$$\begin{aligned} \frac{\mathrm{d}q}{\mathrm{d}t} &= +\nabla_p \mathcal{E}\left(q(t), p(t)\right), \\ \frac{\mathrm{d}p}{\mathrm{d}t} &= -\nabla_q \mathcal{E}\left(q(t), p(t)\right) = -\nabla U\left(q(t)\right) \end{aligned} \quad (3)$$

where the *potential* $U : \mathbb{R}^d \mapsto \mathbb{R}_+$ is chosen s.t. the target $\pi(q) \propto \exp\left(-U(q)\right)$, the *kinetic* term $K : \mathbb{R}^d \mapsto \mathbb{R}_+$ is assumed to have a form of $K(p) = \frac{1}{2} p^\top M p$, where M is the mass matrix, and the *Hamiltonian* is defined as $\mathcal{E}(q, p) := U(q) + K(p)$.[1] The extended target $\bar{\pi}$ for *phase points* $z := (q, p)$ on the *phase space* $\mathbb{R}^d \times \mathbb{R}^d$ is then defined as having density $\propto \exp\left(-\mathcal{E}(q, p)\right)$.

To describe the dynamics more succinctly, we consider the *flow* map $\Phi_t(q_0, p_0) = (q(t), p(t))$ for (3) initialized at $(q_0, p_0) := \big(q(0), p(0)\big) \in \mathbb{R}^d \times \mathbb{R}^d$. Following Heng and Jacob (2019), we write $\Phi_t^\circ(q_0, p_o) = q(t)$ and $\Phi_t^*(q_0, p_0) = p(t)$ for the flow projected onto its position and momentum spaces, respectively. The flow map Φ_t is in general not available in closed form and requires discretization in time via numerical integrators as approximations. A standard choice for HMC is the *leapfrog integrator* that, given an initial phase point (q_0, p_0), iterates:

$$\begin{aligned} p_{\ell+1/2} &:= p_\ell - \frac{\varepsilon}{2} \nabla U(q_\ell) \\ q_{t+1} &:= q_\ell + \varepsilon p_{\ell+1/2} \\ p_{t+1} &:= p_{\ell+1/2} - \frac{\varepsilon}{2} \nabla U(q_{\ell+1}) \end{aligned}$$

[1] Unless otherwise specified we let $M^{-1} = I_d$ throughout, though we note that M can be chosen using existing adaption methods, e.g. (Carpenter et al., 2017), or as in Riemannian HMC (Girolami and Calderhead, 2011).

for $\ell = 0, \dots, L-1$ with a step size $\varepsilon > 0$ and leapfrog steps $L \in \mathbb{N}$. We denote $\hat{\Phi}_{\varepsilon,\ell}(q_0, p_0) := (q_\ell, p_\ell)$ as the numerical flow map approximated by a leapfrog integrator with a step size ε for ℓ steps, and similarly $\hat{\Phi}^{\circ}_{\varepsilon,\ell}$ and $\hat{\Phi}^{*}_{\varepsilon,\ell}$ for projected maps onto position and momentum.

Metropolis HMC One can design an MCMC kernel by proposing the *end-point* of a Hamiltonian trajectory in (3). In practice, a discretized trajectory is obtained by leapfrog integration. Due to numerical errors in the simulation, in order to ensure the kernel π-invariant, the proposal needs to be adjusted by a Metropolis-Hasting step (Metropolis et al., 1953; Neal, 2011). Denoting $\mathcal{N}_{\mathrm{d}}$ as the d-dimensional standard Gaussian, the kernel $Q \sim \mathcal{K}^{\mathrm{MH}}_{\varepsilon,L}(Q_0, \cdot)$ for Metropolis HMC follows

$$P_0 \sim \mathcal{N}_{\mathrm{d}}, \quad (q_L, p_L) = \hat{\Phi}_{\varepsilon,L}(Q_0, P_0),$$
$$Q = \begin{cases} q_L & \text{with prob. } \min\{1, \exp(\Delta_{\mathcal{E}})\} \\ Q_0 & \text{otherwise} \end{cases} \tag{4}$$

where $\Delta_{\mathcal{E}} := -\mathcal{E}(q_L, p_L) + \mathcal{E}(Q_0, P_0)$ is the energy difference between the origin and the proposal.

Multinomial HMC Betancourt (2018) describes a trajectory variant of HMC, which we refer as *multinomial* HMC and denote $\mathcal{K}^{\mathrm{Mult}}_{\varepsilon,L}(Q_0, \cdot)$. In multinomial HMC, all intermediate points of a numerical trajectory can be proposed as the next state:

$$P_0 \sim \mathcal{N}_{\mathrm{d}}, \mathbf{t} \sim \mathbb{P}_{\varepsilon,L}(\cdot \mid Q_0, P_0), (Q, P) \sim \mathbb{P}(\cdot \mid \mathbf{t}) \tag{5}$$

where $\mathbf{t} := [(q_{-L_{\mathrm{b}}}, p_{-L_{\mathrm{b}}}), \dots, (Q_0, P_0), \dots, (q_{L_{\mathrm{f}}}, p_{L_{\mathrm{f}}})]$. The *trajectory sampling* $\mathbf{t} \sim \mathbb{P}_{\varepsilon,L}(\cdot \mid Q_0, P_0)$ follows

$$L_{\mathrm{f}} \sim \mathcal{U}(\{0, \dots, L\}), \quad L_{\mathrm{b}} = L - L_{\mathrm{f}},$$
$$(q_\ell, p_\ell) = \begin{cases} \hat{\Phi}_{\varepsilon,\ell}(Q_0, +P_0) & \text{for } \ell = 1, \dots, L_{\mathrm{f}} \\ \hat{\Phi}_{\varepsilon,\ell}(Q_0, -P_0) & \text{for } \ell = 1, \dots, L_{\mathrm{b}} \end{cases} \tag{6}$$

The *intra-trajectory sampling* $Q, P \sim \mathbb{P}(\cdot \mid \mathbf{t})$ follows a multinomial distribution (Betancourt, 2018) as

$$\mathbb{P}\big((Q, P) = (q_\ell, p_\ell) \mid \mathbf{t}\big) = \sigma\big((q_\ell, p_\ell), \mathbf{t}\big) \tag{7}$$

where $\sigma(z, \mathbf{t}) := \exp\left(-\mathcal{E}(z)\right) / \sum_{z' \in \mathbf{t}} \exp\left(-\mathcal{E}(z')\right)$.

2.3 Coupled MCMC kernels

The coupled HMC kernel in (Heng and Jacob, 2019) *and* the coupled kernels proposed in this work can be unified through Algorithm 2. Specifically, letting (i) In Line 2, $\mathbb{P}_{L_{\mathrm{f}}}(L_{\mathrm{f}} = L) = 1$ and (ii) In Line 5, $(i, j) \mid (\mathbf{t}^1, \mathbf{t}^2) \sim \bar{P}_\ell$ in Algorithm 2, we recover the $\bar{\mathcal{K}}^{\mathrm{MH}}_{\varepsilon,L}$ from Heng and Jacob (2019), where $\bar{P}_\ell$ follows the generative process

$$u \sim \mathcal{U}([0, 1]),$$
$$i = \begin{cases} L & \text{if } u < \alpha^1 \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad j = \begin{cases} L & \text{if } u < \alpha^2 \\ 0 & \text{otherwise} \end{cases}$$

Algorithm 2: Coupled HMC kernels

1 Sample $P_0 \sim \mathcal{N}_{\mathrm{d}}$;
2 Sample $L_{\mathrm{f}} \sim \mathbb{P}_{L_{\mathrm{f}}}$ and set $L_{\mathrm{b}} = L - L_{\mathrm{f}}$;
3 **for** $c = 1, 2$ **do**
4 Integrate using leapfrog to obtain $\mathbf{t}^c = [\hat{\Phi}_{\varepsilon,-L_{\mathrm{b}}}(Q_0^c, -P_0), \dots (Q_0^c, P_0), \dots \hat{\Phi}_{\varepsilon,L_{\mathrm{f}}}(Q_0^c, P_0)]$
5 Sample next state indices $(i, j) \mid (\mathbf{t}^1, \mathbf{t}^2) \sim \bar{P}_\ell$;
6 Set $(Q^1, P^1) = \mathbf{t}^1_i, (Q^2, P^2) = \mathbf{t}^2_j$;
7 Output (Q^1, Q^2);

where $\alpha^c = \exp\left(-\mathcal{E}(\mathbf{t}^c_L) + \mathcal{E}(\mathbf{t}^c_0)\right)$ for $c = 1, 2$. This corresponds to using *common random number* (CRN) in the MH correction steps in (4). For coupled multinomial HMC kernels studied in this work, which we denote $\bar{\mathcal{K}}^{\gamma}_{\varepsilon,L}$, we make different choices for Line 2 and Line 5 in Algorithm 2. In short, Line 2 will be a coupled version of (6) and Line 5 will correspond to a coupling γ of two multinomial distributions as (7). We will discuss them in detail in Section 3.

Although coupled HMC kernels can bring two chains within a small neighborhood of each other, the probability of *exact* meeting is zero, thus failing to satisfy conditions for using (2). To alleviate this issue Heng and Jacob (2019) instead propose a mixture of coupled random-walk Metropolis-Hastings (RWMH) $\bar{\mathcal{K}}_\sigma$ and coupled HMC $\bar{\mathcal{K}}_{\varepsilon,L}$ to trigger "exact meeting". The coupled RWMH kernel $\bar{\mathcal{K}}_\sigma$ with proposal variance $\sigma^2 I_d$ uses maximal coupling (Johnson, 1998; Jacob et al., 2020) to encourage two chains meet exactly when they are close. For completeness, we provide it as Algorithm 3 in Appendix A. The overall mixture kernel, denoted $\bar{\mathcal{K}}_{\varepsilon,L,\sigma}$, is then defined as

$$\bar{\mathcal{K}}_{\varepsilon,L,\sigma}\left(\bar{x}, \bar{A}\right) = (1 - \alpha)\, \bar{\mathcal{K}}_{\varepsilon,L}\left(\bar{x}, \bar{A}\right) + \alpha\, \bar{\mathcal{K}}_{\sigma}\left(\bar{x}, \bar{A}\right) \tag{8}$$

for $\alpha \in (0, 1)$, $\bar{x} := (x, y) \in \mathbb{R}^d \times \mathbb{R}^d$ and $\bar{A} := (A, B) \in \mathcal{B}(\mathbb{R}^d) \times \mathcal{B}(\mathbb{R}^d)$. That is, with probability α we use the coupled RWMH kernel and with probability $1-\alpha$ we use the HMC kernel. Heng and Jacob (2019) proves that, under certain assumptions, if the *relaxed meeting time* $\tau_\delta := \inf\{n \geq 0 : \|X_n - Y_{n-1}\| \leq \delta\}$ of the coupled HMC kernel $\bar{\mathcal{K}}_{\varepsilon,L}$ has geometric tails for any $\delta > 0$, the chains meet exactly with non-zero probability under $\bar{\mathcal{K}}_{\varepsilon,L,\sigma}$ for any $\alpha \in (0, 1)$, warranting the use of (2).

The key conditions that ensure the unbiasedness, finite variance and finite computation cost of (2) are (i) the coupled chains marginally converge to the target and (ii) the two chains meet sufficiently quickly and stay together after meeting; see Jacob et al. (2020) for explicit definitions. Suppose our proposed HMC kernels satisfy (i) by construction, to ensure the method satisfies (ii) it is sufficient to prove that the relaxed meeting time has geometric tails. This we establish in Section 4.

3 Optimal Transport Couplings for Multinomial HMC

Recall that in order to use the multinomial HMC kernel $\mathcal{K}^{\text{Mult}}_{\varepsilon,L}$ in Algorithm 2, we need to specify how Line 2 and Line 5 are performed. First, the number of leapfrog steps forward and backward sampled in Line 2 follows (6), inheriting from multinomial HMC, and is shared between the two chains. In other words, both chains simulate forward and backward for the same number of steps, making them "aligned in-time". Second, Line 5 correspond to a coupling of the intra-trajectory multinomial sampling step in (7). To ensure that the marginal chains are equivalent to the original multinomial kernel, it is sufficient to sample (i, j) such that the corresponding marginal *categorical distributions* $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ of (7) for indices i, j are preserved

$$\boldsymbol{\mu} : \mathcal{C}\text{at}(\ell = i) = \sigma(\mathbf{t}^1_\ell, \mathbf{t}^1), \quad \boldsymbol{\nu} : \mathcal{C}\text{at}(\ell = j) = \sigma(\mathbf{t}^2_\ell, \mathbf{t}^2)$$

Here we overload the notations $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ also to refer to their corresponding *probability vectors*.

To this end, our method is fully specified by providing an algorithm to sample (i, j) such that $i \sim \boldsymbol{\mu}$ and $j \sim \boldsymbol{\nu}$. The collection of such joint distributions for (i, j) are couplings of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, i.e. $\Gamma(\boldsymbol{\mu}, \boldsymbol{\nu})$.

3.1 Optimal transport couplings

To repeat, our aim is to construct coupled kernels in which the coupled chains from Algorithm 1 meet in a relatively small number of MCMC steps, i.e. short meeting time. Unfortunately, it is not clear how to directly minimize the meeting time. Intuitively, one might expect a kernel which, informally, "brings chains closer" to also have an improved meeting time. Naturally this brings us to consider the following problem:

$$\gamma := \arg\min_{\gamma'} \sum_{i,j} \gamma'_{ij} D_{ij} \quad \text{s.t.} \quad \gamma' \in \Gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) \tag{9}$$

where $D_{ij} = d(\mathbf{t}^1_i, \mathbf{t}^2_j)$ is the distance in the *position space* between the i-th point in the first trajectory and the j-th point in the second. This is an example of a *Kantorovich problem*, a well-studied family of problems from optimal transport (Villani, 2003). In the case where $d_2^p(x, y) = \|x - y\|_2^p$, we will refer to the minimizer as the W_p*-coupling* due to the role it plays in the Wasserstein distance wrt. Euclidean metric $W_p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \left(\inf_{\gamma \in \Gamma(\boldsymbol{\mu}, \boldsymbol{\nu})} \mathbb{E}_{(X,Y)\sim\gamma} \|x - y\|_2^p\right)^{1/p}$.

In this work we will consider two different choices for the metric d: 1) Euclidean distance d_2 which gives rise to the W_2*-coupling*, and 2) 0-1 distance d_I which gives rise to the *maximal coupling*.

3.2 W_2-coupling

Arguably the most natural choice of metric d in (9) is the squared Euclidean distance $d_2^2(\mathbf{t}^1_i, \mathbf{t}^2_j) = \left\|q^1_i - q^2_j\right\|_2^2$, whose solution we denote γ°. Once we obtain γ°, sampling (i, j) is straightforward. For completeness, we provide the full algorithm as Algorithm 4 in Appendix B.

Computationally, the optimization in (9) can be solved by generic linear programming solvers or more specialized methods, e.g. as in Bonneel et al. (2011). Such solvers in general have a time complexity $\mathcal{O}(K^3)$ where K is the length of the probability vectors $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$. This can be alleviated by using an approximate solver which could introduce biases. Therefore, similarly to Jacob et al. (2016), we also describe a debiasing method that allows the use of approximate solvers in Appendix B.

3.3 Maximal coupling

For general choices of d (9) we do not have analytical solutions, but for the particular choice $d_I(\mathbf{t}^1_i, \mathbf{t}^2_j) = \mathbb{1}(i \neq j)$ we do. In this case, the solution is the well-known maximal coupling γ^* of two categorical distributions, which can be represented in its mixture view as

$$\gamma^* = \omega \frac{\boldsymbol{\mu} \wedge \boldsymbol{\nu}}{Z} + (1 - \omega) \frac{\boldsymbol{\mu} - (\boldsymbol{\mu} \wedge \boldsymbol{\nu}) + \boldsymbol{\nu} - (\boldsymbol{\mu} \wedge \boldsymbol{\nu})}{1 - Z} \tag{10}$$

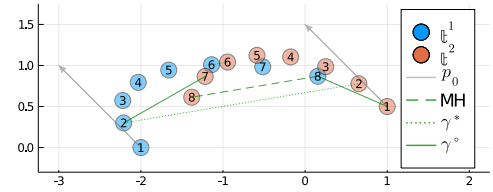
where $\wedge$ is the point-wise minimum operation, $\omega = \mathbb{P}(i = j)$ and $Z = \sum_i (\boldsymbol{\mu} \wedge \boldsymbol{\nu})_i$; sampling from γ^* is therefore tractable and straightforward. The process is summarized in Algorithm 6 in Appendix B.

By definition, for a maximal coupling γ^* the probability of choosing pairs with the same time-index in two trajectories is maximized; we refer to such pairs with same indices as "index-aligned" pairs. As we will see in Section 4, this property allows us to exploit Lemma 1 in Heng and Jacob (2019) to show that the distance between the two coupled chains decreases with non-zero probability when the potential is strongly convex, or, equivalently, the target is strongly log-concave.

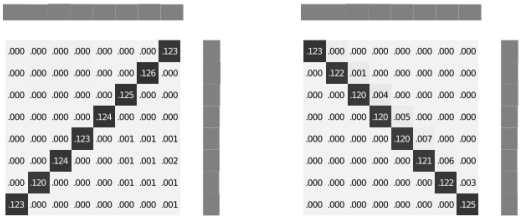
Though the idea of index-aligned pairs is useful to establish the theoretical results, it is not necessarily so in practice. Note that as the approximation of Hamiltonian simulation by numerical integration becomes more accurate when step sizes become smaller, the joint γ^* converges to the diagonal uniform distribution, i.e. $\gamma_{ii} \approx 1/K$ and $\gamma_{ij} \approx 0$ for $i \neq j$ for large K. It is easy to construct examples where this leads to sub-optimal behavior when the goal is to minimize distance between the proposed states; Figure 1 illustrates this nicely.

3.4 An illustration of different couplings

We now illustrate how different intra-trajectory couplings behave using a 2D Gaussian with zero mean and



(a) Coupled trajectories



(b) Maximal coupling

(c) W_2-coupling

Figure 1: An illustration of different HMC couplings. Green lines in Figure 1a indicate possible pairs from different methods. For coupled Metropolis HMC, the dashed line pairs the end-points of two trajectories, which has a relative large distance. The dotted line is for multinomial HMC with maximal coupling. Though there is a change that the 6-th points of two trajectories are paired, other index-aligned pairs are equally likely (Figure 1b). E.g. the pair of 2-th points has a large distance. In contrast, all pairs from multinomial HMC with W_2-coupling (solid lines) have relatively small distances, resulting in a small distance on average. To see this, we calculate the expected distances: they are 1.37 for W_2-coupling and 1.97 for maximal coupling, where the former is clearly smaller, as expected.

unit diagonal covariance. We start by simulating two Hamiltonian trajectories from $q_0^1 = [0.5, 2.0]$ and $q_0^2 = [0.5, -1.0]$ using the same momentum $p_0 = [1.0, 1.0]$ for 7 steps, obtaining two trajectories $\mathbf{t}^1$ and $\mathbf{t}^2$ in Figure 1a, where the arrows represent the initial momentum p_0. We then sample from our two couplings to generate 100,000 pairs of indices to estimate the joint distributions and to compute the marginals, which are shown in Figure 1b and Figure 1c. Note how the joint distributions differs: the ordering of the pairings are "reversed". This intuitively makes sense when looking at Figure 1a, in which, e.g. the closest point for $\mathbf{t}_1^1$ is $\mathbf{t}_8^2$. Finally, note that this U-turn example is chosen to highlight the differences between the two couplings. If there was no U-turn, the differences between the two couplings could potentially be smaller.

4 Theoretical analysis

We now establish geometric tails for the meeting time for the mixture kernel in (8) with the proposed coupled HMC kernels as the HMC component, thus satisfying the necessary conditions to use the estimator (2).

Proof sketch To prove geometric tails it turns out that it is sufficient to prove that the methods satisfy the conditions for Proposition 1 in Heng and Jacob (2019). Informally, the proposition states that once the chains enter a region S in the state space where the target density is strongly log-concave, there is a non-zero probability that the chains will end up in a δ-neighbourhood of each other in some $n_0 \in \mathbb{N}$ steps. The proof presented in Heng and Jacob (2019) obtains this statement for the coupled Metropolis HMC by arguing directly about the probabilities of such an event conditioned on the initial states being in S. Here we instead prove a slightly stronger property, *local contractivity*, from which the proposition follows immediately. Informally, local contractivity ensures that the distance between the chains will decrease on average when initialized in some region. We first prove that this holds for the maximal coupling by exploiting the fact that it maximizes the probability of picking index-aligned pairs, which, as mentioned before, is guaranteed to decrease the distance compared to the initial positions for strongly log-concave targets. Once this has been established, local contractivity for the W_2-coupling follows immediately since by definition W_2-coupling has a smaller expected distance than the maximal coupling. The remainder of the proof is essentially identical to Heng and Jacob (2019) where excursions from the set S is controlled with a geometric drift condition, from which we obtain geometric tails for the meeting time and thus validity of the methods.

Following Heng and Jacob (2019), we make two assumptions on the potential function $U : \mathbb{R}^d \mapsto \mathbb{R}$.

Assumption 1 (Regularity and growth of potential)**.** *The potential U is twice continuously differentiable and its gradient $\nabla U : \mathbb{R}^d \mapsto \mathbb{R}^d$ is globally β-Lipschitz, i.e. there exists $\beta > 0$ such that $\|\nabla U(q) - \nabla U(q')\| \leq \beta\|q - q'\|$ for all $q, q' \in \mathbb{R}^d$.*

Assumption 2 (Local strong convexity of potential)**.** *There exists a compact set $S \in \mathcal{B}(\mathbb{R}^d)$, with positive Lebesgue measure, s.t. the restriction of the potential U to S is α-strongly convex, i.e., $\exists\ \alpha > 0$ s.t. $(q - q')^\top (\nabla U(q) - \nabla U(q')) \geq \alpha\|q - q'\|^2$ for all $q, q' \in S$.*

Unless otherwise specified, we will let S denote the set in Assumption 2, $\bar{\mathcal{K}}_{\varepsilon,L}^{\gamma}$ denote a coupled HMC kernel as described in Algorithm 2 with (i) shared momentum, (ii) shared forward and backward simulation steps and (iii) $(i, j) \sim \gamma$ for intra-trajectory sampling, and $\mathrm{pr}_{\varepsilon,L}^{\gamma}$ denote the law of a coupled HMC kernel $\bar{\mathcal{K}}_{\varepsilon,L}^{\gamma}$. For functions $f : \mathbb{R}^d \to \mathbb{R}$, we will also use the notation $L_\ell(f) = \left\{x \in \mathbb{R}^d : f(x) \leq \ell\right\}$ for the levelsets of f and

$f_A := f\big|_A$ for the restriction of f to $A \in \mathcal{B}(\mathbb{R}^d)$.

4.1 Geometric tails via local contractivity

We first state the definition of local contractivity and Proposition 1 from Heng and Jacob (2019).

Condition 1 (Local contractivity)**.** *Given a compact set $S \in \mathcal{B}(\mathbb{R})$ with positive Lebesgue measure, we say the kernel $\bar{\mathcal{K}}^{\gamma}_{\varepsilon,L}$ is* locally contractive *on S with rate $\rho \in (0,1)$ if there exists $m \geq 1$ such that for any given $k_0 > 0$ there exists $\bar{\varepsilon} > 0$, $\bar{L} \in \mathbb{N}$ s.t.*

$$\mathbb{E}_{(l_1,l_2)\sim\gamma}\left[\left\|\hat{\Phi}^{\circ}_{\varepsilon,l_1}(q^1,p) - \hat{\Phi}^{\circ}_{\varepsilon,l_2}(q^2,p)\right\|^m\right] \leq \rho^m \|q^1 - q^2\|^m \tag{11}$$

for all $\varepsilon \in (0,\bar{\varepsilon})$, $L \in \mathbb{N}$ such that $\varepsilon L < \bar{\varepsilon}\bar{L}$ and for all $(q^1,q^2,p) \in S \times S \times L_{k_0}(K)$.

Informally, this is saying that there exists a step size and integration time such that a single application of $\bar{\mathcal{K}}^{\gamma}_{\varepsilon,L}$ decreases the distance between the two states on average. Furthermore, this property is preserved when decreasing the step size or the integration time. This last part is important since different parts of the analysis will require possibly smaller step sizes and/or integration times. Thus, by ensuring that all statements hold for all smaller step sizes and integration times, we can combine the statements by simply choosing the minimum of the step sizes and/or integration times required by the different statements.

Proposition 4.1 (Proposition 1, Heng and Jacob (2019))**.** *Suppose that the potential U satisfies Assumptions 1 and 2. Then for any $\delta > 0$, $u_0 > \inf_{q\in S} U(q)$, and $u_1 < \sup_{q\in S} U(q)$ with $u_0 < u_1$, there exists $\bar{\varepsilon} > 0$ and $\bar{L} \in \mathbb{N}$ such that for any $\varepsilon \in (0,\bar{\varepsilon})$ and $L \in \mathbb{N}$ satisfying $\varepsilon L < \bar{\varepsilon}\bar{L}$, there exists $v_0 \in (u_0,u_1)$, $n_0 \in \mathbb{N}$ and $\omega \in (0,1)$ such that*

$$\inf_{q^1,q^2\in S_0} \bar{\mathcal{K}}^{\gamma,n_0}_{\varepsilon,L}\left((q^1,q^2),D_\delta\right) \geq \omega \tag{12}$$

where $S_0 = L_{v_0}(U_S)$ is compact with positive Lebesgue measure,

$$\bar{\mathcal{K}}^{\gamma,n}_{\varepsilon,L}\left((q^1,q^2),A^1\times A^2\right) = \mathrm{pr}^{\gamma}_{\varepsilon,L}\left((Q^1_n,Q^2_n)\in A^1\times A^2 \mid (Q^1_0,Q^2_0) = (q^1,q^2)\right)$$

denotes the n-step transition probabilities of the coupled chain, and $D_\delta = \left\{(q,q') \in \mathbb{R}^d \times \mathbb{R}^d : \|q - q'\| \leq \delta\right\}$.

As noted earlier, this proposition is a key step in the proof of the coupled Metropolis HMC kernel in Heng and Jacob (2019). This ensures that once we reach the set $S_0 \subset S$, there is a non-zero probability that within some finite number of steps the chains will be δ-close, i.e. meet in the relaxed sense. If we can then also ensure that this set S_0 will be entered by the coupled chains sufficiently often, then we bound the tails of distribution over meeting times.

We now establish that indeed, for coupled multinomial HMC kernels, Condition 1 implies Proposition 4.1.

Lemma 4.1. *If $\bar{\mathcal{K}}^{\gamma}_{\varepsilon,L}$ satisfies Condition 1, then $\bar{\mathcal{K}}^{\gamma}_{\varepsilon,L}$ satisfies the conditions of Proposition 4.1.*

Proof. Observe that

$$\begin{aligned}
&\mathrm{pr}_{\varepsilon,L}\left(\left\|Q^1_1 - Q^2_1\right\| \leq \rho\left\|Q^1_0 - Q^2_0\right\| \mid (Q^1_0,Q^2_0) = (q^1,q^2)\right)\\
&= \mathbb{E}_{\bar{\mathcal{K}}^{\gamma}_{\varepsilon,L}}\left[\mathbb{1}\left\{\left\|Q^1_1 - Q^2_1\right\| \leq \rho\|q^1 - q^2\|\right\}\right]\\
&= \mathbb{E}_{P\sim\mathcal{N}(0,I)}\left[\mathbb{E}_{(l_1,l_2)\sim\gamma}\left[\mathbb{1}(R_{q^1,q^2,P}) \mid P\right]\right]
\end{aligned}$$

where we have let $R_{q^1,q^2,p}$ denote the set of events where we have contraction, i.e.

$$R_{q^1,q^2,p} = \left\{\left\|\hat{\Phi}^{\circ}_{\varepsilon,l_1}(q^1,p) - \hat{\Phi}^{\circ}_{\varepsilon,l_2}(q^2,p)\right\| \leq \rho\|q^1 - q^2\|\right\}$$

By Condition 1, $\mathbb{E}_{(l_1,l_2)\sim\gamma}\left[\mathbb{1}(R_{q^1,q^2,p})\right] > 0$ for any $(q^1,q^2,p) \in S \times S \times L_{k_0}(K)$, where $k_0 > 0$ is to be decided, since otherwise (11) would not hold. Hence a single application of the kernel $\bar{\mathcal{K}}^{\gamma}_{\varepsilon,L}$ will have a non-zero probability of decreasing the distance between the states if we are in S. The remainder of the proof ensures that parameters can be chosen such that there is a non-zero probability of staying within the set $S_0 \subset S$ for some $n_0 := \inf\left\{n \in \mathbb{N} : \rho^n\|q^1 - q^2\| \leq \delta\right\}$ applications of the kernel, i.e. $(Q^1_k,Q^2_k) \in S_0 \times S_0$ for all $k = 1,\ldots,n_0$. This finally allows us to conclude that there is a non-zero probability of entering D_δ if we are currently in the set S_0. See the Appendix C.1 for the full proof. □

Theorem 4.1 (Theorem 2, Heng and Jacob (2019))**.** *Suppose that the potential U satisfies Assumptions 1 and 2. Suppose that there exists $\bar{\varepsilon} > 0$ and $\bar{\sigma} > 0$ such that for any $\varepsilon \in (0,\bar{\varepsilon})$, $L \in \mathbb{N}$ and $\sigma \in (0,\bar{\sigma})$, there exists a measurable function $V : \mathbb{R}^d \to [1,\infty)$, $\lambda \in (0,1)$, $b < \infty$ and $\mu > 0$ such that*

$$\mathcal{K}_{\varepsilon,L}(V)(x) \leq \lambda V(x) + b, \quad Q_\sigma(V)(x) \leq \mu\left(V(x)+1\right)$$

for all $x \in \mathbb{R}^d$, $\pi_0(V) < \infty$, $\lambda_0 = (1-\gamma)\lambda + \gamma(1+\mu) < 1$ and $\left\{x \in \mathbb{R}^d : V(x) \leq \ell_1\right\} \subseteq \left\{x \in S : U(x) \leq \ell_0\right\}$, for some $\ell_0 \in \{\inf_{x\in S} U(x), \sup_{x\in S} U(x)\}$ and $\ell_1 > 1$ satisfying $\lambda_0 + 2\left((1-\gamma)b + \gamma\mu\right)(1-\lambda_0)^{-1}(1+\ell_1)^{-1} < 1$. Then there exists $\varepsilon_0 \in (0,\bar{\varepsilon})$, $L_0 \in \mathbb{N}$ 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

$$\mathrm{pr}^{\gamma}_{\varepsilon,L,\sigma}(\tau > n) \leq C_0\kappa_0^n$$

for some $C_0 \in \mathbb{R}_+$ and $\kappa_0 \in (0,1)$ and for $n \in \mathbb{N}_0$, where $\mathrm{pr}^{\gamma}_{\varepsilon,L,\sigma}$ denotes the law of the kernel $\bar{\mathcal{K}}^{\gamma}_{\varepsilon,L,\sigma}$ for a given coupled HMC kernel $\bar{\mathcal{K}}^{\gamma}_{\varepsilon,L}$.

Proof. The proof is identical to Heng and Jacob (2019) via Lemma 4.1. □

4.2 Local contractivity for W_2-coupling and maximal coupling

Now we establish Condition 1 for coupled multinomial HMC kernels with maximal coupling γ^* and W_2-coupling γ°, ensuring that Theorem 4.1 applies to the resulting mixture kernels $\bar{\mathcal{K}}^*_{\varepsilon,L,\sigma}$ and $\bar{\mathcal{K}}^\circ_{\varepsilon,L,\sigma}$.

We first restate a slight variation of Lemma 1 from Heng and Jacob (2019), which tells us that the states reached by *exact* flows with shared momentum is closer than the initial states for sufficiently small integration times.

Lemma 4.2. *Suppose that the potential U satisfies Assumptions 1 and 2. For any compact set $A \subset S \times S \times \mathbb{R}^d$, there exists a trajectory length $T > 0$ such that*

$$\left\|\Phi^\circ_t(q^1,p) - \Phi^\circ_t(q^2,p)\right\| \leq \rho\left\|q^1 - q^2\right\| \tag{13}$$

for all $t \in [-T,T] \setminus \{0\}$ and all $(q^1,q^2,p) \in A$.

Proof. See Appendix C.2 for the detailed proof. □

Note that Lemma 4.2 is a statement about the distance between the integrated states *at the same integration time t*. As an immediate consequence the expected distance with respect to a joint distribution with probability mass only along the diagonals satisfies a similar property, controlling for numerical errors (see Appendix C.3). Therefore our strategy in proving local contractivity for $\bar{\mathcal{K}}^*_{\varepsilon,l}$ and $\bar{\mathcal{K}}^\circ_{\varepsilon,l}$ is to ensure that as we decrease the stepsize the probability mass on the diagonals, i.e. $\mathbb{P}(i=j)$, can be made close to 1.

Maximal coupling To establish Condition 1 for coupled multinomial HMC with maximal coupling γ^*, $\bar{\mathcal{K}}^*_{\varepsilon,l}$, we first introduce a bound on the total variation distance between the trajectory distributions $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$.

Proposition 4.2. *Suppose that U satisfies Assumptions 1 and 2. For any $\delta > 0$, there exists $\varepsilon_0 > 0$, $L_0 \in \mathbb{N}$ s.t. for all $\varepsilon \in (0,\varepsilon_0)$, $L \in \mathbb{N}$ satisfying $\varepsilon L < \varepsilon_0 L_0$, we have*

$$\mathrm{D_{TV}}(\boldsymbol{\mu},\boldsymbol{\nu}) = \mathbb{P}(i \neq j) < \delta. \tag{14}$$

Proof. The proof uses the 1-Lipschitz property of the softmax function (Gao and Pavel, 2018) in (7) to bound the probability differences introduced by numerical errors. See Appendix C.4. □

With Proposition 4.2, we can establish local contractivity for coupled multinomial HMC kernels with γ^*.

Lemma 4.3. *$\bar{\mathcal{K}}^*_{\varepsilon,l}$ satisfies Condition 1.*

Proof. Due to Proposition 4.2, for a given integration time, we can choose step size arbitrary small to increase probability of picking parallel-in-time pairs, whose contractivity is established in Proposition C.1. See Appendix C.5 for the complete proof. □

W_2-coupling Similarly to Lemma 4.3, for coupled multinomial HMC with γ°, $\bar{\mathcal{K}}^\circ_{\varepsilon,l}$, we have:

Lemma 4.4. *$\bar{\mathcal{K}}^\circ_{\varepsilon,l}$ satisfies Condition 1.*

Proof. The definition of γ° from (9) implies

$$\begin{aligned}\mathbb{E}_{(i,j)\sim\gamma^\circ}\left[\left\|\hat{\Phi}^\circ_{\varepsilon,l_i}(q^1,p) - \hat{\Phi}^\circ_{\varepsilon,l_j}(q^2,p)\right\|^2\right] \leq \\ \mathbb{E}_{(i,j)\sim\gamma^*}\left[\left\|\hat{\Phi}^\circ_{\varepsilon,l_i}(q^1,p) - \hat{\Phi}^\circ_{\varepsilon,l_j}(q^2,p)\right\|^2\right]\end{aligned} \tag{15}$$

The rest of the proof follows that of Lemma 4.3. □

Lemmas 4.3 and 4.4 together with Theorem 4.1 then establishes geometric tails for the meeting time of the resulting mixture kernels $\bar{\mathcal{K}}^*_{\varepsilon,l}$ and $\bar{\mathcal{K}}^\circ_{\varepsilon,l}$, respectively.

5 Experiments

In this section, we evaluate the performance of the proposed coupled HMC kernels. The coupled Metropolis HMC by Heng and Jacob (2019) is used as a baseline. Following Heng and Jacob (2019), we combine coupled HMC kernels with a coupled RWMH kernel to obtain exact couplings: we take the standard deviation of the RWMH kernel to be $\sigma = 10^{-3}$ and the mixture cofficient (i.e. probability of using RWMH) to be $\alpha = 1/20$. For the estimation task, we consider a more efficient (still unbiased) *time-averaged* variant of (2):

$$H_{k:m}(X,Y) = \frac{1}{m-k+1}\sum_{i=k}^{m} H_i(X,Y) \tag{16}$$

Furthermore, we run R independent pairs of coupled chains (X^r,Y^r), $r = 1,\ldots,R$, and estimate $H^\dagger$ as $\hat{H} = R^{-1}\sum_{r=1}^{R} H_{k:m}(X^{(r)},Y^{(r)})$. For $x \in \mathbb{R}^d$, we consider h as the first and second moments of x, i.e. $h_i(x) = x_i$ and $h_{d+i}(x) = x_i^2$ for $i = 1,\ldots,d$.

We consider three target distributions. The first target is a **1,000D Gaussian**. The second one is the posterior of a **Bayesian logistic regression** model on the German credit dataset (Asuncion and Newman, 2007). We apply the same pre-processing as in Heng and Jacob (2019), which results in a sampling space of $\mathbb{R}^{302}$. The last model considered is a **log-Gaussian Cox point process** that models tree locations in a forest. We discretize the forest using a 16×16 grid, resulting in a sampling space on $\mathbb{R}^{256}$. Note that the first two targets meet necessary conditions from Section 4. More details of these targets can be found in Appendix D.1.

Our implementation is based on AdvancedHMC.jl Xu et al. (2020) and is available at at `https://github.com/TuringLang/CoupledHMC.jl`, which also contains scripts to reproduce results in this paper.

5.1 Meeting time comparisons

We first investigate how the meeting time τ of our method changes under different step sizes ϵ and numbers of leapfrog steps L. For this purpose, we run all coupled HMC methods initialised at a random draw from $\mathcal{N}(0, I)$ for 1,000 iterations. For each method, we use different step sizes ε and leapfrog steps L: $(\varepsilon, L) \in \{0.05, 0.07, \ldots, 0.45\} \times \{5, 10, 15\}$ for 1,000D Gaussians, $(\varepsilon, L) \in \{0.01, 0.0125, \ldots, 0.04\} \times \{10, 20, 30\}$ for logistic regression and $(\varepsilon, L) \in \{0.05, 0.07, \ldots, 0.45\} \times \{10, 20, 30\}$ for log-Gaussian Cox point processes. Furthermore, we repeat each experiment for $R = 10$ times to estimate standard derivation. Figure 2 shows resulting meeting time together with standard derivation for varying ε and a fixed $L = 10$; figures for other L are similar so we defer those to Appendix E.1. It is worth noting that τ equal to 1,000 should be interpreted as coupled chains *did not meet within 1,000 iterations*.

Figure 2 shows clearly that both maximal coupling and W_2-coupling achieve smaller meeting time than the baseline for large step sizes. This robustness against large step sizes is useful in practice since it allows us to simulate a trajectory of a given length with less computation, by using larger ϵ rather than larger L. However, when the step size is sufficiently small Metropolis HMC will almost always accept the end-point, thus travel the full integration length T at every step. In contrast, multinomial HMC will put uniform mass on intermediate states which means that it travels $1/4T$ in expectation. Therefore Metropolis HMC will move towards the typical set faster and thus have a smaller meeting time compared to multinomial HMC. It is also worth noting the surge in meeting time for coupled Metropolis HMC in Figure 2a around $\varepsilon = 0.3$ can be explained by the similar phenomena observed in Figure 1, large trajectory length can lead to end-points close to their starting points, thus never meet. In particular, the trajectory length $3 = 0.3 \times 10$ is around $\pi \approx 3.14$, in which case trajectories are basically full circles ending close to where they start, a special case for Gaussians.

Furthermore, for logistic regression (Figure 2b), optimal parameters of HMC ($\epsilon = 0.03, L = 10$) leads to excessively long meeting time. This result is consistent with those in Heng and Jacob (2019). This is clearly undesirable: optimal parameters for sampling efficiency leads to non-contractive coupled chains. Besides, it is worth noting that maximal coupling is more robust to large step sizes than W_2-optimal coupling for the logistic regression model. To understand this, recall that W_2-coupling takes a local greedy approach but there is *no guarantee* it can lead to faster meeting through multiple transitions. With large step sizes, numerical errors in simulation are enlarged, leading to more probabilities assigned to non-diagonal entries in the coupling matrix, equivalently more freedom in the W_2-coupling. In such cases, the greedy effect of W_2-coupling is also enlarged but such greedy approach *turns out* to be less effective than maximal coupling for the logistic regression model, a target that satisfies Assumptions 1 and 2. In short, whether or not the greedy approach is preferable is target-dependent. Specifically, when a target satisfies Assumptions 1 and 2, one would expect maximal coupling to work well enough; when such assumptions fail, W_2-coupling can be more efficient, as seen in Appendix E.2.

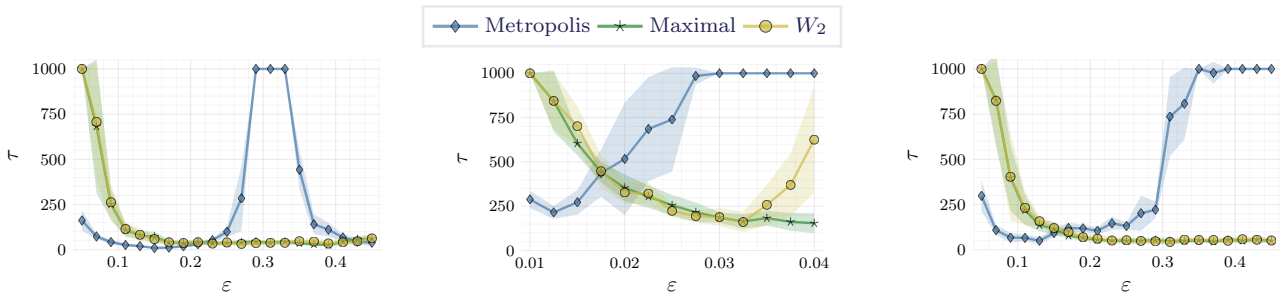
Finally, as motivated earlier, one can use existing adaption techniques to choose parameters ε, L. For example, one can use the adapted parameters from preliminary runs of NUTS. As an concrete example, NUTS-adapted ε, L for logistic regression are 0.022 and 22, and those for log-Gaussian Cox point processes are 0.28 and 16, which allows our method to meet relatively fast: 114 and 118 for the first model and 50 and 51 for the second one, for the two proposed kernels respectively.

5.2 Estimator efficiency comparisons

Although Monte Carlo estimates by (16) are unbiased, it can have large variances due to the use of coupled but often short Markov chains. In other words, making (16) unbiased comes at a cost of increased variance. Therefore, it is helpful to study the efficiency, or *inefficiency*, of the estimator under a joint effect of removed bias but increased variance, which we define next.

For a vector-valued function h, the variance of estimator $\hat{H}$ for coupled HMC is defined as $\sum_d \nu(h_d)$ where $\nu(h) = \mathbb{V}_r\left(H_{k:m}(h, X^r, Y^r)\right)$. Here r is the index of repeated runs. *Asymptotic inefficiency* is defined as $\sum_d i(h_d)$ where $i(h) = \hat{C}\nu(h)$ (Glynn and Whitt, 1992). Here $\hat{C} = \mathbb{E}_r\left[2(\tau^r - 1) + \max(1, m + 1 - \tau^r)\right]$ is the *expected cost* over R runs and τ^r is the meeting time for the r-th run. The asymptotic variance of (non-coupled) HMC can be approximated with the `spectrum0.ar` function of the `coda` R package (Plummer et al., 2006) using a long chain: 10,000 samples after a burn-in of 1,000 using $(\epsilon, L) = (0.03, 10)$ for logistic regression, and $(0.3, 10)$ for the log-Gaussian Cox point process model. Relative inefficiency is then defined as the ratio of asymptotic inefficiency over asymptotic variance.

We study inefficiency using logistic regression and log-Gaussian Cox point processes, both widely used in practice. Following Heng and Jacob (2019), we set ε and L to values resulting in smallest meeting time in Section 5.1. We first perform 100 runs of coupled HMC kernels to get an empirical distribution of meeting time τ, then we use this distribution to determine k and m following heuristics from Heng and Jacob (2019): take k as either the median or the 90% sample quantile of



(a) 1000D Gaussian (b) Logistic regression (c) Log-Gaussian Cox point process

Figure 2: Meeting time τ with different ε and $L = 10$ out of $R = 10$ runs with lines for average and shade for 1 standard deviation. Overall, coupled multinomial HMC attains smaller meeting time and is more robust to ε.

k	m	Metropolis[2]	Maximal	W_2
median	$5k$	2.40	17.69	3.18
	$10k$	2.39	7.37	3.73
90% quantile	$5k$	2.36	2.14	2.88
	$10k$	2.32	1.90	**0.94**
median	$5k$	6.03	4.84	7.62
	$10k$	4.81	4.01	6.00
90% quantile	$5k$	5.26	4.58	6.86
	$10k$	4.61	**3.83**	5.80

Table 1: Relative inefficiency with different k and m for logistic regression (top half) and log-Gaussian Cox point processes (bottom half). Bold indicates the one most close to 1. Note that for each method, k is different thus inefficiencies across different coupled kernels (across columns) are *not directly comparable*. Instead, we aim to study if the relative inefficiency can be made close to 1 with suitable parameters (across rows).

τ and m as a product of k and a constant, e.g. $5k$ or $10k$. We then perform $R = 100$ independent runs of coupled chains with different combinations of k and m – we expect a longer chain to have a smaller variance but larger computation budget. Also bear in mind that an ideal asymptotic inefficiency should be close to 1.

Table 1 shows asymptotic inefficiencies for varying k and m. For logistic regression, with suitable choices of k, m, the relative inefficiency can be made close to 1 (W_2-coupling); for the other model, the best (3.83) is attained by maximal coupling, both of which are superior to best of Metropolis. Overall, it demonstrates that optimal transport couplings can obtain better bias-variance trade-off than the baseline. Note that both optimal transport couplings *seem* to be inefficient for $m = 5k$ with k being the median. This is because the chains meet quickly compared to Metropolis, and are thus further from the stationary distribution. Table 1 also confirms that larger m reduces asymptotic inefficiency at the cost of more computation. Variance reduction can also be achieved by parallel execution.

[2]We kindly note that relative inefficiencies reported here for Metropolis on logistic regression differ from Heng and Jacob (2019) by a factor ≈ 2.0, as confirmed by the authors.

6 Related Work

Research on couplings for MCMC methods has a long history (Devroye, 1990; Johnson, 1996, 1998; Rosenthal, 1997; Meyn and Tweedie, 2012; Rowland et al., 2018; Nuesken and Pavliotis, 2018; Jacob et al., 2019; Biswas et al., 2019). Couplings for HMC has been more recently focusing on Metropolis HMC, e.g. Neal (2017). Most closely related to our work is Heng and Jacob (2019), in which coupling for Metropolis HMC is established. Bou-Rabee et al. (2020) studied the convergence of Metropolis HMC, and also proposed a new way to couple momentum variables, called *contractive coupling*, that does not rely on sharing them.[3]

Our analysis in Section 4 is also related to works on convergence analysis of HMC on log-concave targets (Mangoubi and Smith, 2017; Chen and Vempala, 2019).

7 Conclusion

In this paper, we develop two novel couplings for multinomial HMC. We provide theoretical analysis on the validity of the proposed methods, and perform simulations to demonstrate their advantages over existing methods in terms of meeting time and robustness to HMC parameters, which is an important step towards practical use of coupled HMC. We hope this work will help advance the research on searching more efficient coupled HMC methods, and a wider use of coupled HMC for probabilistic modeling in practice. For future work, we are interested in extending the coupling methods to more advanced HMC variants, e.g. the no-U-turn algorithm (Hoffman and Gelman, 2014).

[3]We also study the effect of contractive coupling in our methods empirically but defer this to Appendix E.2.

Acknowledgement

Hong Ge and Tor Erlend Fjelde acknowledge generous support from Huawei Research and donations from Microsoft Research. The proposed algorithms' implementation is based on the Turing probabilistic programming language, from which the authors benefit greatly. We would also like to thanks Cameron Pfiffer, Mohamed Tarek, Martin Trapp, Sharan Yalburgi for helpful comments on an earlier draft of this paper.

References

Asuncion, A. and Newman, D. (2007). UCI machine learning repository.

Betancourt, M. (2018). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434 [stat]*. arXiv: 1701.02434.

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*, volume 32.

Bonneel, N., Van De Panne, M., Paris, S., and Heidrich, W. (2011). Displacement interpolation using Lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, pages 1–12.

Bou-Rabee, N., Eberle, A., and Zimmer, R. (2020). Coupling and convergence for Hamiltonian Monte Carlo. *Annals of applied probability: an official journal of the Institute of Mathematical Statistics*, 30(3):1209–1250.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).

Chen, Z. and Vempala, S. S. (2019). Optimal convergence rate of Hamiltonian Monte Carlo for strongly logconcave distributions. *arXiv:1905.02313 [cs, stat]*.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300.

Devroye, L. (1990). Coupled samples in simulation.

Gao, B. and Pavel, L. (2018). On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv:1704.00805 [cs, math]*.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741.

Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214.

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

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

Hairer, E., Lubich, C., and Wanner, G. (2006). *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*. Number 31 in Springer series in computational mathematics. Springer, Berlin ; New York, 2nd ed edition.

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

Hoffman, M. D. and Gelman, A. (2014). The no-U-turn sampler: Adaptively setting path Lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623. arXiv: 1111.4246.

Jacob, P. E., Lindsten, F., and Schön, T. B. (2016). Coupling of particle filters. *CoRR*.

Jacob, P. E., Lindsten, F., and Schön, T. B. (2019). Smoothing with couplings of conditional particle filters. *Journal of the American Statistical Association*, pages 1–20.

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

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.

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

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.

Meyn, S. P. and Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2.

Neal, R. M. (2017). Circularly-coupled Markov chain sampling. *arXiv:1711.04399 [stat]*.

Nuesken, N. and Pavliotis, G. A. (2018). Constructing sampling schemes via coupling: Markov semigroups and optimal transport. *arXiv:1806.11026 [math]*.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11. Number: 1.

Qiu, Y., Zhang, L., and Wang, X. (2019). Unbiased contrastive divergence algorithm for training energy-based latent variable models. In *International Conference on Learning Representations*.

Rosenthal, J. S. (1997). Faithful couplings of Markov chains: now equals forever. *Advances in Applied Mathematics*, 18(3):372–381.

Rowland, M., Choromanski, K. M., Chalus, F., Pacchiano, A., Sarlos, T., Turner, R. E., and Weller, A. (2018). Geometrically coupled Monte Carlo sampling. In *Advances in Neural Information Processing Systems 31*, pages 195–206.

Teh, Y. W., Welling, M., Osindero, S., and Hinton, G. E. (2003). Energy-based models for sparse overcomplete representations. *Journal of Machine Learning Research*, 4(Dec):1235–1260.

Thorisson, H. (2000). *Coupling, Stationarity, and Regeneration*. Probability and Its Applications. Springer New York.

Villani, C. (2003). *Topics in optimal transportation*. Number 58. American Mathematical Soc.

Xie, J., Lu, Y., Zhu, S.-C., and Wu, Y. (2016). A theory of generative convnet. In *International Conference on Machine Learning*, pages 2635–2644.

Xu, K., Ge, H., Tebbutt, W., Tarek, M., Trapp, M., and Ghahramani, Z. (2020). AdvancedHMC.jl: A robust, modular and efficient implementation of advanced HMC algorithms. In *Symposium on Advances in Approximate Bayesian Inference*, pages 1–10.