A Additional Background

A.1 Properties of Hamiltonian flow

The flow map Φ_t has the following properties:

1. (Reversibility). $\forall\ t \in \mathbb{R}_+$, the inverse flow map Φ_t^{-1} satisfies $\Phi_t^{-1} = R \circ \Phi_t \circ R$, where $R(q,p) = (q,-p)$ denotes the momentum reversal operation.
2. (Energy conservation). The Hamiltonian $\mathcal{E}$ of the system satisfies $\mathcal{E} \circ \Phi_t = \mathcal{E}$.
3. (Measure preservation). For any $t \in \mathbb{R}_+$ and $A \in \mathcal{B}(\mathbb{R}^{2d})$, we have $\mathrm{Leb}_{2d}(\Phi_t(A)) = \mathrm{Leb}_{2d}(A)$, where Leb_d denotes the Lebesgue measure on $\mathbb{R}^d$.

Together the properties ensures that the Markov kernel defined by the Hamiltonian flow leaves the extended target distribution $\bar{\pi}$ invariant.

A.2 Properties of leapfrog integration

The numerical flow map $\hat{\Phi}_{\varepsilon,L}$ enjoys the following two inequalities due to the simplicity of order-two leapfrog integrators (Hairer et al., 2006)

$$\left\|\hat{\Phi}_{\varepsilon,L}(q_0,p_0) - \Phi_{\varepsilon L}(q_0,p_0)\right\| \leq C_a(q_0,p_0,L)\varepsilon^2 \quad (17)$$

$$\left\|\mathcal{E}\left(\hat{\Phi}_{\varepsilon,L}(q_0,p_0)\right) - \mathcal{E}(q_0,p_0)\right\| \leq C_b(q_0,p_0,L)\varepsilon^2 \quad (18)$$

for some positive constants C_a and C_b. These two inequalities are used in several places through our theoretical analysis, e.g. in Section C.2 and Section C.4.

A.3 Coupled RWMH kernel

The coupled RWMH kernel from Heng and Jacob (2019) used in this paper is shown in Algorithm 3 for completeness. Note that here we slightly abuse notation, writing $\mathcal{K}_\sigma(X,Y)$ to mean denote the probability *density* of the probability measure $\mathcal{K}_\sigma(X,\cdot)$ evaluated at Y, where X and Y are random variables.

B Additional Algorithmic Details

B.2 Sampling from discrete joints

For completeness, we provide an algorithmic description of how to sample a pair of indices given their joint probability matrix in Algorithm 4.

B.3 Debiasing marginal-non-preserving joints

A side effect of using fixed-point iteration solvers or even approximate solvers (Cuturi, 2013) to solve (9) is that the solution does not belong to $\Gamma(\boldsymbol{\mu},\boldsymbol{\nu})$. We denote such solutions as J°, which indicates it is a joint probability matrix rather than a proper coupling. Therefore we need a way to ensure that when using J°, we still have $i \sim \boldsymbol{\mu}$ and $j \sim \boldsymbol{\nu}$ exactly, which we refer as a *debiasing* step. Inspired by the mixture view of the maximal coupling, the result of our debiasing algorithm, the debiased W_2-coupling $\hat{\gamma}^\circ$, can be as well viewed as a mixture

$$\hat{\gamma}^\circ = \alpha J^\circ + (1-\alpha)J^d$$

where α is the probability of sampling from J°, and J^d is the debiasing joint probability matrix. The algorithm aims to find the maximal probability α such that $\hat{\gamma}^\circ \in \Gamma(\boldsymbol{\mu},\boldsymbol{\nu})$, together with the corresponding debiasing matrix J^d. First, to find the maximal α, we see that

Algorithm 3: Coupled RWMH kernel with maximal coupling (Jacob et al., 2020)

Input: A pair of current states (X_0, Y_0) and a RWMH kernel $\mathcal{K}_\sigma$ with variance $\sigma^2 I_d$
Output: A pair of next states (X', Y')

Sample $X^* \sim \mathcal{K}_\sigma(X_0, \cdot)$;
Sample $w \mid X \sim \mathcal{U}([0, \mathcal{K}_\sigma(X_0, X^*)])$;
if $w \leq \mathcal{K}_\sigma(Y_0, X^*)$ **then**
  Set $Y^* = X^*$;
else
  repeat
    Sample $Y^* \sim \mathcal{K}_\sigma(Y_0, \cdot)$;
    Sample $w^* \mid Y^* \sim \mathcal{U}([0, \mathcal{K}_\sigma(Y_0, Y^*)])$;
  until $w^* > \mathcal{K}_\sigma(X_0, Y^*)$;
Sample $u \sim \mathcal{U}([0,1])$;
Set $X = X_0$ and $Y = Y_0$;
if $u \leq \min\{1, \pi(X^*)/\pi(X_0)\}$ **then**
  Set $X = X^*$
if $u \leq \min\{1, \pi(Y^*)/\pi(Y_0)\}$ **then**
  Set $Y = Y^*$
Output (X, Y);

Algorithm 4: Sampling from a discrete joint J

Input: A $M \times N$ matrix J that represents the joint of two categorical distributions
Output: A pair of indices $(i,j) \sim J$

for $i = 1, \ldots, M, j = 1, \ldots, N$ **do**
  Compute $k = M(i-1) + j$;
  Set $u_k = (i,j)$ and $\mathbf{v}_k = J_{ij}$;
Sample $k \sim \mathcal{C}\mathrm{at}(\mathbf{v})$;
Output u_k;

Algorithm 5: Maximally sampling from a joint $\hat{\gamma}$ while ensuring marginals to be $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$

Input: A $K \times K$ probability matrix $\hat{\gamma}$ and two K-length probability vectors $\boldsymbol{\mu}, \boldsymbol{\nu}$ to target ;
Output: A pair of indices (i, j) with $i \sim \boldsymbol{\mu}$ and $j \sim \boldsymbol{\nu}$ while maximally using $\hat{\gamma}$;
Compute $\boldsymbol{\mu}^\circ$ and $\boldsymbol{\nu}^\circ$ as marginals of $\hat{\gamma}$;
Compute α according to (20);
Sample $U \sim \mathcal{U}([0, 1])$;
if $U < \alpha$ **then**
  Sample $(i, j) \sim \hat{\gamma}$ using Algorithm 4;
else
  Compute $\boldsymbol{\mu}^d$ and $\boldsymbol{\nu}^d$ by solving (19);
  Sample $i \sim \boldsymbol{\mu}^d$ and $j \sim \boldsymbol{\nu}^d$;
Output (i, j);

$\hat{\gamma}^\circ \in \Gamma(\boldsymbol{\mu}, \boldsymbol{\nu})$ implies

$$\boldsymbol{\mu} = \alpha\boldsymbol{\mu}^\circ + (1-\alpha)\boldsymbol{\mu}^d, \quad \boldsymbol{\nu} = \alpha\boldsymbol{\nu}^\circ + (1-\alpha)\boldsymbol{\nu}^d \tag{19}$$

where $\boldsymbol{\mu}^\circ$ and $\boldsymbol{\nu}^\circ$ are marginals of J° and $\boldsymbol{\mu}^d$ and $\boldsymbol{\nu}^d$ are marginals of J^d. Since $\boldsymbol{\mu}^d$ and $\boldsymbol{\nu}^d$ are K-length probability vectors, we have $\mu_i^d > 0$ and $\nu_i^d > 0$ for all $i = 1, \dots, K$, which implies a set of constrains on α

$$\mu_i \geq \alpha\mu_i^\circ, \quad \text{and} \quad \nu_i \geq \alpha\nu_i^\circ \quad \text{for all } i = 1, \dots, K$$

Therefore, the maximal value of α is given by

$$\alpha = \min\{1, \frac{\mu_1}{\mu_1^\circ}, \dots, \frac{\mu_K}{\mu_K^\circ}, \frac{\nu_1}{\nu_1^\circ}, \dots, \frac{\nu_K}{\nu_K^\circ}\}. \tag{20}$$

With α found, we can solve (19) to find $\boldsymbol{\mu}^d$ and $\boldsymbol{\nu}^d$, and J^d can be chosen as any coupling of them, i.e. $J^d \in \Gamma(\boldsymbol{\mu}^d, \boldsymbol{\nu}^d)$, including the *independent coupling* that simply samples as $i \sim \boldsymbol{\mu}^d, j \sim \boldsymbol{\nu}^d$. We summarise in Algorithm 5 a sampling procedure of $\hat{\gamma}^\circ$ resulting from this debiasing approach. It is not hard to see that by construction, the approach satisfies (19) and yields $\hat{\gamma}^\circ \in \Gamma(\boldsymbol{\mu}, \boldsymbol{\nu})$, which, as a result, yields a coupled HMC kernel whose marginal kernels converge to the target. Also, when there is no bias, i.e. $J^\circ \in \Gamma(\boldsymbol{\mu}, \boldsymbol{\nu})$, we have $\alpha = 1$ from (20) and the algorithm reduces to exact W_2-coupling.

B.4 Sampling from discrete maximal maximal coupling

For completeness, we provide an algorithmic description of how to sample a pair if indices from the maximal coupling of two categorical distribution in Algorithm 6.

Algorithm 6: Maximal coupling of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$

Input: Two categorical distributions $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$;
Output: A pair of indices $(i, j) \sim \gamma^*$;
Compute $\omega = 1 - \mathrm{D_{TV}}(\boldsymbol{\mu}, \boldsymbol{\nu})$ and $Z = \sum_i (\boldsymbol{\mu} \wedge \boldsymbol{\nu})_i$;
Sample $u \sim \mathcal{U}([0, 1])$;
if $u \leq \omega$ **then**
  Sample $i \sim \mathcal{C}\mathrm{at}(\frac{\boldsymbol{\mu} \wedge \boldsymbol{\nu}}{Z})$ and set $j = i$;
else
  Sample $i \sim \mathcal{C}\mathrm{at}(\frac{\boldsymbol{\mu} - (\boldsymbol{\mu} \wedge \boldsymbol{\nu})}{1 - Z})$,
  $j \sim \mathcal{C}\mathrm{at}(\frac{\boldsymbol{\nu} - (\boldsymbol{\mu} \wedge \boldsymbol{\nu})}{1 - Z})$;
Output (i, j);

C Technical Details

C.1 Proof of Lemma 4.1

Proof. Suppose $\bar{\mathcal{K}}_{\varepsilon,L}^\gamma$ satisfies Condition 1 on the set S for some $\bar{\varepsilon} > 0$, $\bar{L} \in \mathbb{N}$.

First observe that

$$\begin{aligned}
&\mathrm{pr}_{\varepsilon,L}^\gamma \left(\left\| Q_1^1 - Q_1^2 \right\| \leq \rho \left\| Q_0^1 - Q_0^2 \right\| \mid (Q_0^1, Q_0^2) = (q^1, q^2) \right) \\
&= \mathbb{E}_{\bar{\mathcal{K}}_{\varepsilon,L}^\gamma} \left[\mathbb{1} \left\{ \left\| Q_1^1 - Q_1^2 \right\| \leq \rho \left\| q^1 - q^2 \right\| \right\} \right] \\
&= \mathbb{E}_{P \sim \mathcal{N}(0,I)} \left[\mathbb{E}_{(l_1,l_2) \sim \gamma} \left[\mathbb{1}(R_{q^1,q^2,P}) \mid P \right] \right]
\end{aligned}$$

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

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

By Condition 1 we know that there exists $\omega_1 \in (0, 1)$ such that

$$\mathbb{P}_{(l_1,l_2) \sim \gamma} \left(R_{q^1,q^2,p} \right) \geq \omega_1 \tag{21}$$

for all $(q^1, q^2, p) \in S \times S \times L_{k_0}(K)$, where $k_0 > 0$. By the tower property of expectation, this immediately implies that

$$\begin{aligned}
&\mathbb{E}_{P \sim \mathcal{N}(0,I)} \left[\mathbb{E}_{(l_1,l_2) \sim \gamma} \left[\mathbb{1}(R_{q^1,q^2}) \, \mathbb{1} \left\{ K(P) \leq k_0 \right\} \mid P \right] \right] \\
&\geq \mathbb{E}_{P \sim \mathcal{N}(0,I)} \left[\omega_1 \, \mathbb{1} \left\{ K(P) \leq k_0 \right\} \right] \\
&= \omega_1 \mathbb{P}_{P \sim \mathcal{N}(0,I)} \left(\left\{ K(P) \leq k_0 \right\} \right) \\
&> 0
\end{aligned}$$

where the last inequality follows from the fact that the level sets $L_{k_0}(K)$ are closed for any $k_0 > 0$ since K is continuous and bounded and therefore compact, in addition to having positive Lebesgue measure. Since (21) holds for all $(q^1, q^2, p) \in S \times S \times L_{k_0}(K)$ with $\omega_1 > 0$,

we have

$$\begin{aligned}\inf_{q^1,q^2\in S} \mathrm{pr}^{\gamma}_{\varepsilon,L}\big(\{\|Q_1^1-Q_1^2\|\le\rho\|Q_0^1-Q_0^2\|\}\\ \cap\{K(P)\le k_0\}\mid (Q_0^1,Q_0^2)=(q^1,q^2)\big)\\ \ge\omega_1\omega_2\\ >0\end{aligned} \tag{22}$$

where we have let $\omega_2=\mathbb{P}_{P\sim\mathcal{N}(0,I)}(K(P)\le k_0)$.

In words, for any initial points $(q^1,q^2)\in S\times S$, a single application of the kernel $\bar{\mathcal{K}}^{\gamma}_{\varepsilon,L}$ decreases the distance with non-zero probability. Equipped with this, proving the desired statement is just a matter of ensuring that we can indeed apply (22) repeatedly to get the states sufficiently close to each other. A straightforward approach to this is to simply choose the stepsize to be sufficiently small such that even when taking the required number of steps to get within the desired δ-ball, every step taken is still within a set where (22) holds. This is exactly the approach taken in Heng and Jacob (2019) and so the rest of the proof is essentially identical to the last paragraph in the proof of Proposition 1 in Heng and Jacob (2019).

Consider $u_0>\inf_{q\in S}U(q)$, and $u_1<\sup_{q\in S}U(q)$ with $u_0<u_1$, and let $A_\ell:=L_\ell(U_S)\times L_{u_1-\ell}(K)\subset L_{u_1}(\mathcal{E})$ for $\ell\in(u_0,u_1)$. Since continuity and convexity of U_S imply that this is a closed function, its level sets $L_\ell(U_S)$ are closed. Moreover, under the assumptions on U and S, it follows that these level sets are compact with positive Lebesgue measure. Note that if $(q,p)\in A_\ell$, due to energy conservation and continuity of U, the mapping $t\mapsto\Phi^\circ_t(q,p)$ imply that $\Phi^\circ_t(q,p)\in L_{u_1}(U_S)$ for any $t\in[-T,T]$. Due to time discretization, using (18) and compactness of A_ℓ we can only conclude that there exists $\eta_0>0$ such that $\hat{\Phi}^\circ_{\varepsilon,l}(q,p)\in L_{u_1+\eta_0}(U)$ for all $(q,p)\in A_\ell$ and $l=L_b,...,L_f$. Let $n_0=\min\{n\in\mathbb{N}:\rho^nB\le\delta\}$, where $B:=\sup_{q^1,q^2\in S}\|q^1-q^2\|$. By choosing $v_0\in(u_0,u_1)$, $k_0>0$, and $\eta_0>0$ small enough such that

$$v_0+(n_0+1)k_0+n_0\eta_0<u_1$$

holds, we have $Q_k^1,Q_k^2\in S$ for all $k=1,\dots,n_0$. Hence, by repeated application of (22),

$$\inf_{q^1,q^2\in S_0}\mathrm{pr}_{\varepsilon,L}\left(\left\|Q^1_{n_0}-Q^2_{n_0}\right\|\le\delta\mid(Q_0^1,Q_0^2)=(q^1,q^2)\right)>0$$

with $S_0=L_{v_0}(U_S)$, exactly as in Proposition 4.1. □

C.2 Proof of Lemma 4.2

Lemma C.1. *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 $T>0$ and a step size $\varepsilon_1>0$ s.t. for any $\varepsilon\in(0,\varepsilon_1]$ and any $t\in[-T,T]\setminus\{0\}$ with $l:=t/\varepsilon\in\mathbb{Z}$, there exists $\rho\in[0,1)$ satisfying*

$$\left\|\hat{\Phi}^\circ_{\varepsilon,l}(q_0^1,p_0)-\hat{\Phi}^\circ_{\varepsilon,l}(q_0^2,p_0)\right\|\le\rho\|q_0^1-q_0^2\| \tag{23}$$

for all $(q_0^1,q_0^2,p_0)\in A$.

Proof. As the leapfrog integrator is of order two (Hairer et al., 2006; Bou-Rabee et al., 2020), for any sufficently small step size ε and number of step l states above, we have $\left\|\hat{\Phi}_{\varepsilon,l}(q_0,p_0)-\Phi_t(q_0,p_0)\right\|\le C_1(q_0,p_0,t)\varepsilon^2$ and similar for its position-projected correspondence

$$\left\|\hat{\Phi}^\circ_{\varepsilon,l}(q_0,p_0)-\Phi^\circ_t(q_0,p_0)\right\|\le C_1(q_0,p_0,t)\varepsilon^2 \tag{24}$$

where $C_1(q_0,p_0,t)$ is some constant that only depends on q_0,p_0 and t.

By (Lemma 1, Heng and Jacob, 2019), with some fixed T, we have $\rho'\in[0,1)$ satisfying

$$\left\|\Phi^\circ_t(q_0^1,p_0)-\Phi^\circ_t(q_0^2,p_0)\right\|\le\rho'\|q_0^1-q_0^2\| \tag{25}$$

for any $t\in(0,T]$ and all $(q_0^1,q_0^2,p_0)\in A$. Since $\Phi^\circ_t(q_0^1,-p_0)=\Phi^\circ_{-t}(q_0^1,p_0)$, applying (Lemma 1, Heng and Jacob, 2019) again with the momentum variable negated, we have (25) for $t\in[-T,0)$. Therefore (25) holds for $t\in[-T,T]\setminus\{0\}$.

With these two intermediate results, we can now bound the left-hand side (LHS) of (13) for any $t\in[-T,T]\setminus\{0\}$ with $l=t/\varepsilon\in\mathbb{Z}$ and all $(q_0^1,q_0^2,p_0)\in A$

$$\begin{aligned}&\left\|\hat{\Phi}^\circ_{\varepsilon,l}(q_0^1,p_0)-\hat{\Phi}^\circ_{\varepsilon,l}(q_0^2,p_0)\right\|\\ =&\|\hat{\Phi}^\circ_{\varepsilon,l}(q_0^1,p_0)-\Phi^\circ_t(q_0^1,p_0)-\\ &\hat{\Phi}^\circ_{\varepsilon,l}(q_0^2,p_0)+\Phi^\circ_t(q_0^2,p_0)+\Phi^\circ_t(q_0^1,p_0)-\Phi^\circ_t(q_0^2,p_0)\|\\ \le&\left\|\hat{\Phi}^\circ_{\varepsilon,l}(q_0^1,p_0)-\Phi^\circ_t(q_0^1,p_0)\right\|+\\ &\left\|\hat{\Phi}^\circ_{\varepsilon,l}(q_0^2,p_0)-\Phi^\circ_t(q_0^2,p_0)\right\|+\left\|\Phi^\circ_t(q_0^1,p_0)-\Phi^\circ_t(q_0^2,p_0)\right\|\\ \le&\left(C(q_0^1,p_0,t)+C(q_0^2,p_0,t)\right)\varepsilon^2+\rho'\|q_0^1-q_0^2\|\end{aligned}$$

where the third line is a result of the triangle inequality and the last line comes from (24) and (25) respectively. As $\lim_{\varepsilon\to0}\left(C(q_0^1,p_0,t)+C(q_0^2,p_0,t)\right)\varepsilon^2=0$, for any $\rho\in(\rho',1)$, there exists a step size $\varepsilon_1>0$ such that for any $\varepsilon\le\varepsilon_1$, (13) holds. □

C.3 Proof of Proposition C.1

For the sake of presentation, in this section we only consider Condition 1 for $m=1$. To prove that γ^* satisfies Condition 1 for $m>1$ follows the exact reasoning since (13) in Lemma 4.2 still holds when both sides are raised to some positive power m.

Proposition C.1. *Suppose that U satisfies Assumptions 1 and 2. For any compact set $A \subset S \times S \times \mathbb{R}^d$ and any parallel-in-time joint $J^{\parallel} \in \mathbb{R}^{K\times K}$, there exists a trajectory length $T > 0$, a step size $\varepsilon_1 > 0$ s.t. for any $\varepsilon \in (0, \varepsilon_1]$ and any $L_1, L_2 \in \mathbb{N}$ with $L_1 + L_2 = K - 1$ and $\varepsilon L_1, \varepsilon L_2 < T$, there exists $\tilde{\rho} \in (0, 1)$ satisfying*

$$\mathbb{E}_{(i,j)\sim J^{\parallel}}\left[\left\|\hat{\Phi}^{\circ}_{\varepsilon,l_i}(q^1,p) - \hat{\Phi}^{\circ}_{\varepsilon,l_j}(q^2,p)\right\|\right] \leq \tilde{\rho}\left\|q^1 - q^2\right\| \tag{26}$$

for all $(q^1, q^2, p) \in A$, where l_k is the k-th entry of the vector $[-L_1, \ldots, 0, \ldots, L_2]$.

Proof. By definition, $J^{\parallel}$ has only diagonal entries, thus $(i, j) \sim J^{\parallel}$ is equivalent to (i, i) with $i \sim \mathrm{diag}(J^{\parallel})$. Denote the left-hand side of (26) as A_1, expanding and rearranging A_1 and applying Lemma 4.2, we have

$$\begin{aligned}
A_1 &= \sum_{k=0}^{L_1+L_2+1} \mathbb{P}(i=k) \times \left\|\hat{\Phi}^{\circ}_{\varepsilon,l_k}(q^1,p) - \hat{\Phi}^{\circ}_{\varepsilon,l_k}(q^2,p)\right\| \\
&= \sum_{k\neq L_1+1} \mathbb{P}(i=k) \times \left\|\hat{\Phi}^{\circ}_{\varepsilon,l_k}(q^1,p) - \hat{\Phi}^{\circ}_{\varepsilon,l_k}(q^2,p)\right\| \\
&\qquad + \mathbb{P}(i=L_1+1) \times \left\|\hat{\Phi}^{\circ}_{\varepsilon,0}(q^1,p) - \hat{\Phi}^{\circ}_{\varepsilon,0}(q^2,p)\right\| \\
&\leq \sum_{k\neq L_1+1} \mathbb{P}(i=k) \times \rho_{l_k} \times \left\|q^1 - q^2\right\| \\
&\qquad + \mathbb{P}(i=L_1+1) \times \left\|q^1 - q^2\right\| \\
&= \mathbb{E}_i\left[\rho_{l_i}\right] \times \left\|q^1 - q^2\right\| \\
&:= \tilde{\rho}\left\|q^1 - q^2\right\|
\end{aligned}$$

where we let $\rho_0 = 1$. As $\rho_l \in (0, 1)$ for $l \neq 0$ and $\rho_0 = 1$, $\mathbb{E}_i[\rho_{l_i}] = \sum_k \mathbb{P}(i = k) \times \rho_{l_k} \in (0, 1)$ by the property of convex combination. In other words, we have $\tilde{\rho} \in (0, 1)$. □

C.4 Proof of Proposition 4.2

Proof. For two length-K Hamiltonian trajectories $\mathbf{t}^1$ and $\mathbf{t}^2$, denote $\mathbf{x} = [\mathcal{E}(\mathbf{t}^1_1), \ldots, \mathcal{E}(\mathbf{t}^1_K)]$ and $\mathbf{y} = [\mathcal{E}(\mathbf{t}^2_1), \ldots, \mathcal{E}(\mathbf{t}^2_K)]$ as vectors of the Hamiltonian energy of all phasepoints. With the softmax function $\sigma(\mathbf{x})_i = \exp(-\mathbf{x}_i)/\sum_{i'}\exp(-\mathbf{x}_{i'})$, the entries of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ can be expressed as

$$\mu_i = \sigma(\mathbf{x})_i \quad \nu_j = \sigma(\mathbf{y})_j$$

By the Cauchy–Schwarz inequality, we have $\|\sigma(\mathbf{x}) - \sigma(\mathbf{y})\|_1 \leq \sqrt{K}\|\sigma(\mathbf{x}) - \sigma(\mathbf{y})\|$. With this, we can then upper-bound $\mathrm{D_{TV}}(\boldsymbol{\mu}, \boldsymbol{\nu})$ as

$$\begin{aligned}
\mathrm{D_{TV}}(\boldsymbol{\mu}, \boldsymbol{\nu}) &= \mathrm{D_{TV}}(\sigma(\mathbf{x}), \sigma(\mathbf{y})) \\
&= \frac{1}{2}\|\sigma(\mathbf{x}) - \sigma(\mathbf{y})\|_1 \\
&\leq \frac{1}{2}\sqrt{K}\|\sigma(\mathbf{x}) - \sigma(\mathbf{y})\|
\end{aligned}$$

Denote the energy of the initial phasepoints in each trajectory (q^1_0, p_0) and (q^2_0, p_0) as $\mathcal{E}^1_0$ and $\mathcal{E}^1_0$ and let $\mathcal{E}^1_i := \mathbf{x}_i$ and $\mathcal{E}^1_j := \mathbf{y}_j$; note that for some $i_0 \in \{1, \ldots, K\}$ we have $\mathcal{E}(\mathbf{t}^c_{i_0}) = \mathcal{E}^c_0$ for $c = 1, 2$, i.e. i_0 represents the initial time-index which is shared between the two. As the leapfrog integrator is of order two (Hairer et al., 2006; Bou-Rabee et al., 2020), for any sufficiently small step size $\varepsilon = T/L$, we have

$$|\mathcal{E}^c_0 - \mathcal{E}(\mathbf{t}^c_i))| \leq C_2(q^c_0, p_0)\, t_i\, \varepsilon^2 \leq C_2(q^c_0, p_0)\, T\, \varepsilon^2 \tag{27}$$

for $c = 1, 2$, where t_i denotes the corresponding integration time for the i-th phasepoint from the first phasepoint. Denote the energy differences as $\Delta^1_i = \mathcal{E}(\mathbf{t}^1_i) - \mathcal{E}^1_0$ and $\Delta^1_j = \mathcal{E}(\mathbf{t}^2_j) - \mathcal{E}^2_0$ and observe that

$$\sigma(\mathbf{x}) = \sigma([\Delta^1_1, \ldots, \Delta^1_K]) \quad \sigma(\mathbf{y}) = \sigma([\Delta^2_1, \ldots, \Delta^2_K])$$

Using the fact that the softmax function is 1-Lipschitz (Gao and Pavel, 2018) and applying (27), we have

$$\begin{aligned}
\|\sigma(\mathbf{x}) - \sigma(\mathbf{y})\| &= \left\|\sigma([\Delta^1_1, \ldots, \Delta^1_K]) - \sigma([\Delta^2_1, \ldots, \Delta^2_K])\right\| \\
&\leq \left\|[\Delta^1_1, \ldots, \Delta^1_K] - [\Delta^2_1, \ldots, \Delta^2_K]\right\| \\
&\leq \sqrt{\sum_{k=1}^{K} C_2(q^1_0, p_0) C_2(q^2_0, p_0)\, T^2\, \varepsilon^4} \\
&= \sqrt{K C_2(q^1_0, p_0) C_2(q^2_0, p_0)}\, T\, \varepsilon^2
\end{aligned}$$

Substituting back into our bound on $\mathrm{D_{TV}}(\boldsymbol{\mu}, \boldsymbol{\nu})$,

$$\mathrm{D_{TV}}(\boldsymbol{\mu}, \boldsymbol{\nu}) \leq \frac{1}{2} K \sqrt{C_2(q^1_0, p_0) C_2(q^2_0, p_0)}\, T\, \varepsilon^2$$

Since T is fixed, $\varepsilon = T/L$ and $K = L + 1$, we have

$$\begin{aligned}
\mathrm{D_{TV}}(\boldsymbol{\mu}, \boldsymbol{\nu}) &\leq \frac{1}{2}\sqrt{C_2(q^1_0, p_0) C_2(q^2_0, p_0)}\, T^3\, \frac{L+1}{L^2} \\
&\leq \sqrt{C_2(q^1_0, p_0) C_2(q^2_0, p_0)}\, T^3\, L^{-1} \\
&\leq \sqrt{C_2(q^1_0, p_0) C_2(q^2_0, p_0)}\, T^2\, \varepsilon.
\end{aligned} \tag{28}$$

Finally, note that the upper-bound decreases in with ε and T, hence for any given $\delta > 0$, there exists $\varepsilon_0 > 0$, $L_0 \in \mathbb{N}$ such that $\mathrm{D_{TV}}(\boldsymbol{\mu}, \boldsymbol{\nu}) \leq \delta$ for all $\varepsilon \in (0, \varepsilon_0)$ and $L \in \mathbb{N}$ satisfying $\varepsilon L < \varepsilon_0 L_0 = T$. □

C.5 Proof of Lemma 4.3

Similarly to in Appendix C.3 we only consider Condition 1 with $m = 1$ as the case of $m > 1$ follows similarly.

To prove Lemma 4.3 we first restate a more detailed version of the lemma, which we then prove.

Lemma C.2. *Suppose that the potential U satisfies Assumptions 1 and 2. For a maximal coupling γ^*,*

there exists a trajectory length $T > 0$ and a step size $\varepsilon_2 > 0$ such that for any $\varepsilon \in (0, \min\{\varepsilon_1, \varepsilon_2\}]$ and any $t \in [-T, T] \setminus \{0\}$ with $l := t/\varepsilon \in \mathbb{Z}$, there exists $\rho_2 \in (0,1)$ satisfying

$$\mathbb{E}_{(l_1,l_2)\sim\gamma^*}\left[\left\|\hat{\Phi}^{\circ}_{\varepsilon,l_1}(q^1,p) - \hat{\Phi}^{\circ}_{\varepsilon,l_2}(q^2,p)\right\|\right] \leq \rho\|q^1 - q^2\| \tag{29}$$

for all $(q^1, q^2) \in S \times S$, where $\bar{\mathcal{K}}^*_{\varepsilon,l}$ is the coupled kernel in Algorithm 2 with (i) shared momentum, (ii) shared forward and backward simulation steps and (iii) $(i,j) \sim \gamma^*$ for intra-trajectory sampling.

Proof. We first decompose γ^* into its "diagonal" and "non-diagonal" components

$$\gamma^* = \omega J^{\parallel} + (1-\omega)J^{\nparallel}$$

where $1 - \omega = \mathbb{P}(i \neq j)$ and $J^{\nparallel}$ is defined to be the residual with normalization. Thus we have

$$\begin{aligned} A_2 &:= \omega\mathbb{E}_{J^{\parallel}}\left[\left\|\hat{\Phi}^{\circ}_{\varepsilon,l_i}(q^1,p) - \hat{\Phi}^{\circ}_{\varepsilon,l_j}(q^2,p)\right\|\right] \\ &\quad + (1-\omega)\mathbb{E}_{J^{\nparallel}}\left[\left\|\hat{\Phi}^{\circ}_{\varepsilon,l_i}(q^1,p) - \hat{\Phi}^{\circ}_{\varepsilon,l_j}(q^2,p)\right\|\right] \\ &\leq \omega\tilde{\rho}\|q^1 - q^2\| \\ &\quad + (1-\omega)\mathbb{E}_{J^{\nparallel}}\left[\left\|\hat{\Phi}^{\circ}_{\varepsilon,l_i}(q^1,p) - \hat{\Phi}^{\circ}_{\varepsilon,l_j}(q^2,p)\right\|\right] \end{aligned} \tag{30}$$

for $T > 0$, $\varepsilon \in (0, \varepsilon_1]$ and $\tilde{\rho} \in (0,1)$ in Proposition C.1. As $\mathbb{E}_{J^{\nparallel}}\left[\left\|\hat{\Phi}^{\circ}_{\varepsilon,l_i}(q^1,p) - \hat{\Phi}^{\circ}_{\varepsilon,l_j}(q^2,p)\right\|\right]$ is finite, by Proposition 4.2, the limit of the upper bound goes to $\tilde{\rho}\|q^1 - q^2\|$ as $\varepsilon \to 0$. In other words, for any $\rho \in (\tilde{\rho}, 1)$, there exists a step size $\varepsilon_2 > 0$ such that for any $\varepsilon \in (0, \min\{\varepsilon_1, \varepsilon_2\}]$,

$$A_2 \leq \rho\|q^1 - q^2\|$$

which is exactly what we wanted to prove. □

D Additional Experimental Details

D.1 Target distributions

We follow the pre-processing steps in Heng and Jacob (2019) for the German credit dataset (Asuncion and Newman, 2007) and the Finnish pine saplings dataset (Møller et al., 1998) used in logistic regression and log-Gaussian Cox point process respectively.

Bayesian logistic regression We combine features in the German credit dataset with all of their standardized pairwise interactions, resulting in a design matrix in $\mathbb{R}^{300\times 1,000}$. Denoting an Exponential distribution with rate λ as $\mathcal{E}\text{xp}(\lambda)$, the Bayesian logistic regression follows the following generative process: $s^2 \sim \mathcal{E}\text{xp}(\lambda), a \sim \mathcal{N}(0, s^2), b \sim \mathcal{N}_{300}$, where the variance $s^2 \in \mathbb{R}$, the intercept $a \in \mathbb{R}$ and the coefficients $b \in \mathbb{R}^{300}$, giving a total dimension $d = 302$.

Log-Gaussian Cox point process Firstly, the plot of the forest is discretized into an $n \times n$ grid. For $i \in \{1, \ldots, n\}^2$, the number of points in each grid cell $y_i \in \mathbb{N}$ is assumed to be conditionally independent given a latent intensity variable Λ_i and follows a Poisson distribution with mean $a\Lambda_i$, where $a = n^{-2}$ is the area of each cell. We denote the logarithm of Λ as X and put a Gaussian process prior with mean $\mu \in \mathbb{R}$ and exponential covariance function $\Sigma_{i,j} = s^2 \exp(-|i-j|/(nb))$ on it, where s^2, b and μ are hyperparameters. The generative process of the number of grid cell points follows $X \sim \mathcal{GP}(\mu, \Sigma)$, $\forall\, i \in \{1, \ldots, n\}^2 : \Lambda_i = \exp(X_i)$, $y_i \sim \mathcal{P}oisson(a\Lambda_i)$. Following (Møller et al., 1998), we use a dataset of 126 Scot pine saplings in a natural forest in Finland, and adapt the parameters $s^2 = 1.91$, $b = 1/33$ and $\mu = \log(126) - s^2/2$.

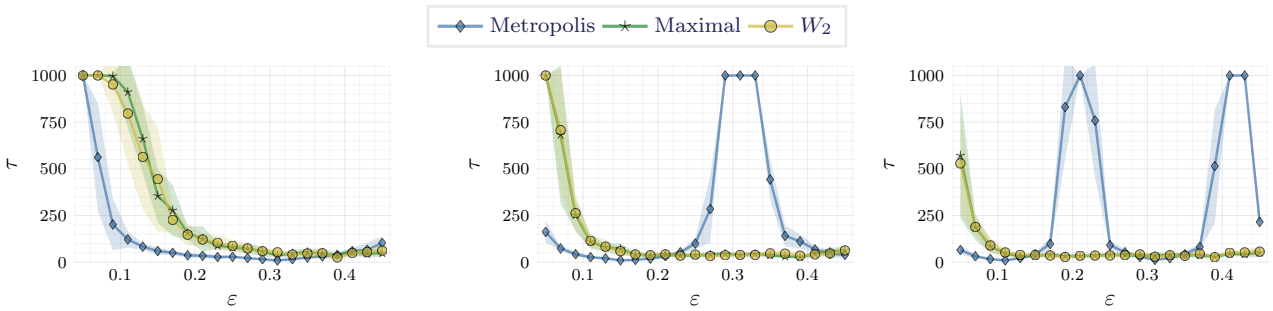
E Additional Experimental Results

E.1 Robustness: meeting time with more parameter sweeps

Figure 3, 4 and 5 provide a wider range of parameter sweep under the same experimental setup as Section 5.1.

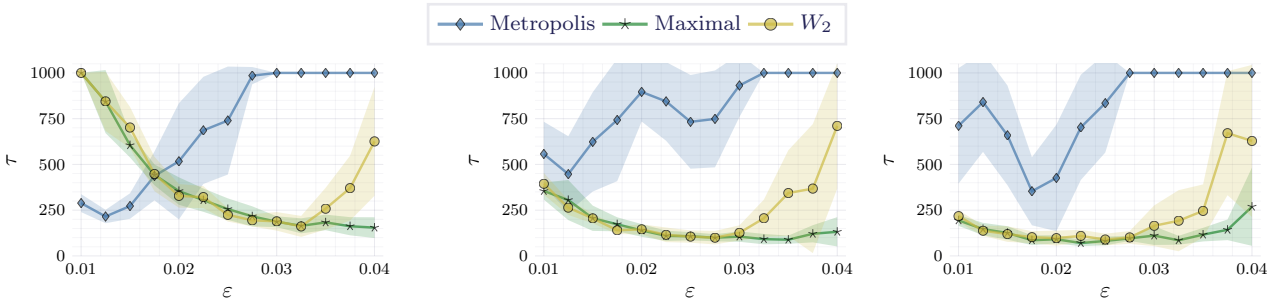
E.2 Toy examples

We first study how proposed methods behave on multimodal distributions. Specifically, we want to know if the coupled chains can meet in a short time given the target is multi-modal. We consider a mixture of Gaussians on $\mathbb{R}^2$ with three components $\mathcal{N}([-1,-1], 0.25^2 I)$, $\mathcal{N}([0,0], 0.25^2 I)$, $\mathcal{N}([1,1], 0.25^2 I)$ weighted by 0.25, 0.4 and 0.35 respectively. We initialise chains from $\mathcal{U}([0,1]^2)$, covering two of the modes. We simulate $R = 500$ pairs of chains and check if they meet within 100 iterations. Denoting the number of chains which meet as N_τ, we report $i_\tau = N_\tau/R$ as a measure of efficiency in meeting. Regarding the choice of ε, L, it is known that HMC is sensitive to the total trajectory length εL in multi-modal distributions: it requires the Hamiltonian simulation long enough to allow jumps between modes. Therefore, starting with $(\varepsilon, L) = (0.1, 10)$, we consider two ways of increasing εL: sweeping $\varepsilon \in \{0.1, 0.15, \ldots, 0.3\}$ and sweeping $L \in \{10, 15, \ldots, 30\}$, equivalently providing a range of total lengths between 1 and 3. While both means increase the trajectory length, the first approach doesn't introduce additional computation but might lead to larger simulation errors, which may then affect the overall performance. Figure 6 provides i_τ under such changes of total trajectory lengths for all methods. First, by increasing εL, our proposed methods overall improve the meeting efficiency, which is not the case



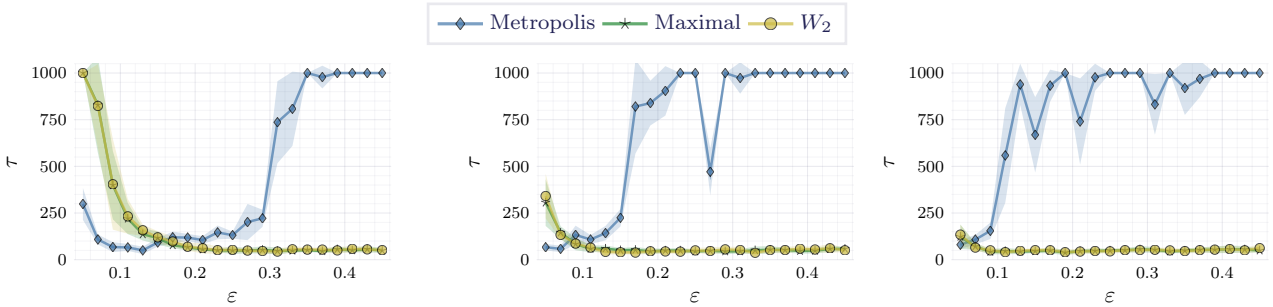
(a) $L = 5$ (b) $L = 10$ (c) $L = 15$

Figure 3: Averaged meeting time $\bar{\tau}$ with different ϵ and L for 1,000D Gaussian.



(a) $L = 10$ (b) $L = 20$ (c) $L = 30$

Figure 4: Averaged meeting time $\bar{\tau}$ with different ϵ and L for logistic regression.



(a) $L = 10$ (b) $L = 20$ (c) $L = 30$

Figure 5: Averaged meeting time $\bar{\tau}$ with different ϵ and L for log-Gaussian Cox point process.

for coupled Metropolis HMC. This can be explained by the following: for coupled Metropolis HMC, meetings can only happen if two chains are proposed to the same mode. However, for coupled multinomial HMC, as long as the trajectories explore common modes, there is a chance for meeting. Especially with W_2-coupling, this chance is further increased by utlizing the actual distances between pairs to find coupling, making it the best in the figure. Second, regarding the two ways of increasing εL, for our proposed methods, increasing L appears to be better as we expected. That said, the gap is relatively small – coupled multinomial HMC tends to be robust against large ε, which is practically useful as it allows the use of a smaller amount of computation comparing to increasing L. Note that we do not claim or indicate our methods improve the mixing in multimodal distributions, which by itself is an important and unsolved issue for HMC. Second, to examine the proposed methods on highly non-convex distributions, we consider a banana-shaped distribution on $\mathbb{R}^2$, of which the potential is given by the Rosenbrock function $U(x_1, x_2) = (1 - x_1)^2 + 10(x_2 - x_1^2)^2$ $(x_1, x_2 \in \mathbb{R})$. As it is done in (Heng and Jacob, 2019), we also take this chance to study the effect of other methods for

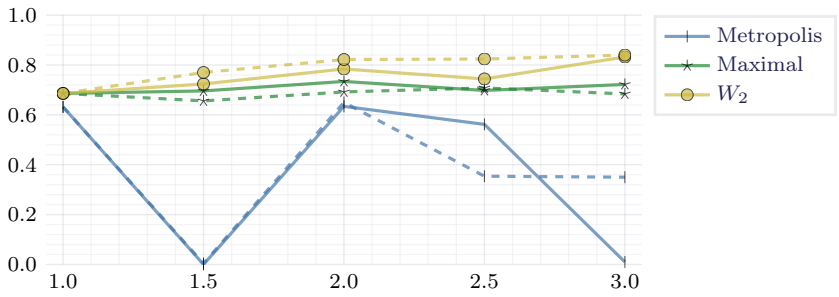


Figure 6: Meeting efficiency on the mixture of Gaussians target with the total trajectory length εL increasing. Solid lines are from increasing ε and dashed ones from increasing L.

Momentum	Metropolis	Maximal	W_2
Shared	136.6 ± 95.8	112.4 ± 74.9	103.8 ± 76.5
Contractive	$\mathbf{39.7 \pm 18.9}$	$\mathbf{81.3 \pm 56.3}$	$\mathbf{77.2 \pm 48.1}$

Table 2: Effect of different momentum coupling methods on meeting time for the Banana target.

coupling the initial momentums rather than simply sharing them. Specifically, we consider the contractive coupling from (Bou-Rabee et al., 2020), in which the initial momentums P^1, P^2 are sampled based on the current positions Q^1, Q^2 as follow

$$P^1 \sim \mathcal{N}(0, I),$$
$$P^2 = \begin{cases} P^1 + \kappa\Delta & \text{with prob. } \frac{\mathcal{N}(\bar{\Delta}^\top P^1 + \kappa|\Delta|;0,1)}{\mathcal{N}(\bar{\Delta}^\top P^1;0,1)} \\ P^1 - 2(\bar{\Delta}^\top P^1)\bar{\Delta} & \text{otherwise} \end{cases}$$

where $\kappa > 0$ is a tuning parameter, $\Delta = Q^1 - Q^2$ is the difference in position space and $\bar{\Delta}$ is the corresponding normalised difference. With initial states sampled from $\mathcal{U}([0,1]^2)$, we simulated $R = 500$ pairs of coupled chains with $(\varepsilon, L) = (1/50, 50)$ for maximally 500 iterations with two momentum coupling methods: shared momentum and contractive coupling with $\kappa = 1$. We summarise means and standard deviations of τ from R runs in Table 2.

First of all, all method with two momentum coupling methods can meet within 150 iterations in such high non-convex setup, except approximate W_2-coupling with contractive momentum. Also, it can be seen that our methods can also benefit from contractive coupling, even though it is derived as a maximal coupling (Thorisson, 2000) for Metropolis HMC. This is the reason why coupled Metropolis HMC is largely improved by it. That is to say, contractive coupling is an orthogonal method of ours rather than a replacement. Note that the table should not be used to compare coupled multinomial HMC against coupled Metropolis HMC in terms of meeting time because they have different optimal parameters for meeting in this target.