Exploring Discrete Analogs of Hamiltonian Monte Carlo

Guangyao Zhou
Vicarious AI
tczhouguangyao@gmail.com

Abstract

Hamiltonian Monte Carlo (HMC) has emerged as a powerful way to sample from continuous distributions. However, a fundamental limitation of HMC is that it can't be applied to discrete distributions. In this paper, we propose the momentum sampler as a general framework for exploring the idea of building discrete analogs of HMC. The momentum sampler is a novel family of Markov Chain Monte Carlo (MCMC) algorithms that extends arbitrary single-site Metropolis Hastings (MH) samplers for discrete distributions with an auxiliary momentum process. It connects and generalizes a variety of existing methods, and points to a more general framework where we can mix stochastic evolution of discrete variables with deterministic evolution of continuous variables. The auxiliary momentum process can be integrated exactly, resulting in a rejection-free sampling process. The various trade-offs involved in picking different proposal distributions and kinetic energies have clear physical interpretations, and are demonstrated by numerical experiments on a simple Potts model.

1 Introduction

Markov Chain Monte Carlo (MCMC) methods are widely used to sample from complex probability distributions. Recently, Hamiltonian Monte Carlo (HMC) [7] has emerged as a powerful MCMC method to sample from continuous distributions. A fundamental limitation of HMC is that it can't be applied to discrete distributions. There have been some recent efforts [9, 8, 11, 2] in applying HMC to discrete distributions. However, they often support only a limited family of distributions: [9] only works for binary distributions; [11], without approximation, only works for binary pairwise models; and [2] is specifically designed for phylogenetic trees. More importantly, there are no clear connections between the behavior of these methods and that of HMC: these methods are usually based on embedding the discrete state space into the continuum, and naively applying HMC (or its discontinuous [9, 6, 8] and probabilistic path [2] variants) on the continuous state space, while the keys to HMC's success lie in its ability to exploit gradient information, to make global proposals by traversing long isoprobability contours, and to make use of the "momentum behavior" (i.e. accelerate in high probability/low energy regions and decelerate in low probability/high energy regions).

In this paper, we explore the idea of building discrete analogs of HMC. We propose the momentum sampler as a general framework for this exploration. The momentum sampler is a novel family of MCMC algorithms that extends arbitrary single-site Metropolis Hastings (MH) samplers for discrete distributions with an auxiliary momentum process. The auxiliary momentum process can be integrated exactly, resulting in a rejection-free sampling process. The various trade-offs involved in picking different proposal distributions and kinetic energies have clear physical interpretations.

In single-site MH samplers (sometimes also referred to as Metropolis-within-Gibbs samplers), each proposal only changes the value of a single random variable. Most of the commonly used MCMC methods for discrete distributions, including the random walk Metropolis sampler and the Gibbs sampler, fall within this family. In fact, many embedding-based HMC methods for discrete

distributions are also closely connected to single-site MH samplers. In Section 4.3 of [8], the authors point out connections between their method and single-site MH samplers. In Section 3.1 of this paper, we show that the Gaussian binary HMC sampler proposed in [9] with travel time $T = \pi$ behaves almost the same as a systematic scan single-site random walk Metropolis sampler.

The momentum sampler has rich connections to existing methods. In Section 3, we prove the Gaussian/exponential binary HMC samplers proposed in [9] are special cases of the momentum sampler, and show that the momentum sampler further generalizes the existing discontinuous [9, 6] and probabilistic path [2] variants of HMC: instead of working with only a uniform proposal (as in [2]), it allows the use of arbitrary proposals while maintaining the desired behavior of HMC. It also points to a more general framework where we can mix stochastic evolution of discrete variables with deterministic evolution of continuous variables.

The rest of the paper is organized as follows. In Section 2, we describe the momentum sampler, prove that it samples from the correct distribution, and discuss the various trade-offs involved in picking different proposal distributions and kinetic energies. In Section 3, we establish connections between the momentum sampler and a variety of existing methods [9, 2, 6]. In Section 4, we present some numerical experiments on a simple 1D 6-state Potts model, to demonstrate the behavior of the momentum sampler with different choices of travel time, proposal distributions and kinetic energies. Finally, we give conclusions and future directions in Section 5.

2 The Momentum Sampler

2.1 The Algorithm

We want to sample from a target discrete distribution $\pi(x)$ on $\Omega = \prod_{i=1}^{N}\{1, \ldots, m_i\}$ (known up to a normalization constant), where $x = (x_1, \ldots, x_N)$ is a random vector, and $x_i \in \{1, \ldots, m_i\}, \forall i \in \{1, \ldots, N\}$. Use $Q_i, i = 1, \ldots, N$ to denote N single-site proposal distributions for a single-site MH sampler, where $Q_i(\tilde{x}|x) > 0$ only when $\tilde{x}_j = x_j, \forall j \neq i$. In other words, the proposal distribution Q_i only changes the value of the random variable x_i. We require each Q_i to be irreducible.

Formally, the momentum sampler operates on the expanded state space $\Sigma = \Omega \times \mathbb{T}^N \times \mathbb{R}^N$, where $\mathbb{T}^N = \mathbb{R}^N/\mathbb{Z}^N$ denotes the N-dimensional flat torus. We are going to identify $\mathbb{T}^N$ as the N-dimensional hypercube $[0, 1]^N$ with the 0's and 1's in different dimensions glued together. The auxiliary location variable $q \in \mathbb{T}^N$ and auxiliary momentum variable $p \in \mathbb{R}^N$ evolve according to the Hamiltonian dynamics, with a flat potential energy $U(q) = 0, \forall q \in \mathbb{T}^N$ and a kinetic energy $\sum_{i=1}^{N} K(p_i), p \in \mathbb{R}^N$, where $K : \mathbb{R} \to \mathbb{R}^+$ is some one-dimensional kinetic energy.

On a high level, the momentum sampler evolves according to the following dynamics: If $q \in (0, 1)^N$, x remains unchanged, and q and p evolve according to the Hamiltonian dynamics

$$\begin{aligned} \frac{\mathrm{d}q_i(t)}{\mathrm{d}t} &= K'(p_i), i = 1, \ldots, N \\ \frac{\mathrm{d}p(t)}{\mathrm{d}t} &= -\nabla U(q) = 0 \end{aligned}$$

If q hits the either 0 or 1 at site j (i.e. $q_j \in \{0, 1\}$), we propose a new $\tilde{x}$ using Q_j, calculate $\Delta E = \log \frac{\pi(x)Q_j(\tilde{x}|x)}{\pi(\tilde{x})Q_j(x|\tilde{x})}$, and modify x, q_j and p_j accordingly

$$\begin{aligned} \tilde{x} &\sim Q_j(\cdot|x) \\ (x, q_j, p_j) &\leftarrow \begin{cases} (\tilde{x}, 1 - q_j, \mathrm{sign}(p_j)K^{-1}(K(p_j) - \Delta E)) & \text{if } K(p_j) > \Delta E \\ (x, q_j, -p_j) & \text{if } K(p_j) \leqslant \Delta E \end{cases} \end{aligned}$$

Because of the flat potential energy U, we can exactly integrate the Hamiltonian dynamics for arbitrary kinetic energy K. How to choose the kinetic energy K for HMC is an active research area [5, 1]. In general, K needs to be differentiable and symmetric around 0, have a well-defined inverse on $\mathbb{R}^+$ (which we denote by K^{-1}, i.e. $K^{-1}(K(z)) = z, \forall z \in \mathbb{R}^+$), and we should be able to sample from $\eta(z) \propto e^{-K(z)}$. In this paper, we consider a simple family of kinetic energies $K_\beta(z) = |z|^\beta$ that's general enough to encompass all commonly used kinetic energies and clearly demonstrates the

Algorithm 1 The Momentum Sampler Algorithm

Input π, target discrete distribution; $Q_i, i = 1, \ldots, N$, single-site proposals; β, parameter for the kinetic energy $K_\beta(z) = |z|^\beta$; S, number of MCMC samples; T, travel time for each step

Output $x^{(1)}, x^{(2)}, \ldots, x^{(S)}$, MCMC samples from the momentum sampler

Initialize x (e.g. sample x_i uniformly from $\{1, \ldots, m_i\}$)
$q_i \leftarrow U([0, 1]), i = 1, \ldots, N$
for l **from** 1 **to** S **do**
    $p_i \leftarrow G(0, 1, \beta), i = 1, \ldots, N$
    $x, q, p \leftarrow$ TAKEMOMENTUMSTEP$(x, q, p, T|\pi, Q_i, i = 1, \ldots, N, \beta)$ ▷ See Algorithm 2
    $x^{(l)} \leftarrow x$
end for
return $x^{(1)}, x^{(2)}, \ldots, x^{(S)}$

Algorithm 2 The *TakeMomentumStep* Function

Parameters π, target discrete distribution; $Q_i, i = 1, \ldots, N$, single-site proposal distributions; β, parameter for the kinetic energy $K_\beta(z) = |z|^\beta$

Input $x^{(0)}$, current discrete state; $q^{(0)}$, current location; $p^{(0)}$, current momentum; T, travel time

Output x, next discrete state; q, next location; p, next momentum

function TAKEMOMENTUMSTEP$(x^{(0)}, q^{(0)}, p^{(0)}, T|\pi, Q_i, i = 1, \ldots, N, \beta)$
    $x \leftarrow x^{(0)}, q \leftarrow q^{(0)}, p \leftarrow p^{(0)}$
    $v_i \leftarrow K_\beta^{'}(p_i) = \text{sign}(p_i)\beta|p_i|^{\beta-1}, i = 1, \ldots, N$
    $t_i \leftarrow \frac{\text{sign}(v_i)+1-2q_i}{2v_i}, i = 1, \ldots, N$
    while $T > 0$ **do**
        $j \leftarrow \text{argmin}_i\{t_i, i = 1, \ldots, N\}$
        $\varepsilon = \min\{t_j, T\}$
        $q_i \leftarrow q_i + \varepsilon v_i, i = 1, \ldots, N$
        $T \leftarrow T - \varepsilon$
        if $\varepsilon = t_j$ **then**
            $t_i \leftarrow t_i - t_j, i = 1, \ldots, N$
            $\tilde{x} \sim Q_j(\cdot|x)$
            $\Delta E \leftarrow \log \frac{\pi(x)Q_j(\tilde{x}|x)}{\pi(\tilde{x})Q_j(x|\tilde{x})}$
            if $K_\beta(p_j) = |p_j|^\beta > \Delta E$ **then**
                $p_j \leftarrow \text{sign}(p_j)(K_\beta(p_j) - \Delta E)^{\frac{1}{\beta}}$
                $v_j \leftarrow K_\beta^{'}(p_j) = \text{sign}(p_j)\beta|p_j|^{\beta-1}$
                $q_j \leftarrow 1 - q_j$
                $x \leftarrow \tilde{x}$
            else
                $p_j \leftarrow -p_j$
                $v_j \leftarrow -v_j$
            end if
            $t_j \leftarrow \frac{\text{sign}(v_j)+1-2q_j}{2v_j}$
        end if
    end while
    $p \leftarrow -p$ ▷ Exists to ensure the *TakeMomentumStep* function is reversible
    return x, q, p
end function

underlying physical interpretations. Use $G(\mu, \alpha, \beta)$ to denote the generalized normal distribution with location μ, scale α and shape β, $K_\beta(z)$ corresponds to $G(0, 1, \beta)$, and $K_\beta^{-1}(z) = z^{\frac{1}{\beta}}, \forall z \in \mathbb{R}^+$.

A detailed description of the momentum sampler algorithm is given in Algorithm 1

2.2 The Momentum Sampler Samples from the Correct Distribution $\pi(x)$

In this section, we prove that the momentum sampler samples from the correct distribution $\pi(x)$. For this purpose, we show the momentum sampler preserves the joint invariant distribution $\varphi((x,q,p)) \propto \pi(x)e^{-\left[U(q)+\sum_{i=1}^{N} K_\beta(p_i)\right]}$, and establish its irreducibility and aperiodicity.

At each iteration, the resampling of p can be seen as a Gibbs step, where we sample p from the conditional distribution of p given x, q. This obviously preserves φ. So we only need to prove detailed balance of the *TakeMomentumStep* function w.r.t. φ. Formally, $\forall T > 0$, *TakeMomentumStep* defines a transition probability kernel R_T: $\forall (x,q,p) \in \Sigma$ and $B \subset \Sigma$ measurable

$$R_T((x,q,p),B) = \mathbb{P}\left(\text{TakeMomentumStep}(x,q,p,T|\pi,Q_i,i=1,\ldots,N,\beta) \in B\right)$$

For all $A \subset \Sigma$ measurable, $q \in \mathbb{T}^N$ and $p \in \mathbb{R}^N$, define $A(q,p) = \{x \in \Omega : (x,q,p) \in A\}$. We have

Theorem 1. ***(Detailed Balance)*** *The TakeMomentumStep function satisfies detailed balance w.r.t. the joint invariant distribution φ, i.e. for any measurable sets $A, B \subset \Sigma$,*

$$\int_{\mathbb{T}^N}\int_{\mathbb{R}^N}\sum_{x\in A(q,p)} R((x,q,p),B)\varphi((x,q,p))\mathrm{d}p\mathrm{d}q = \int_{\mathbb{T}^N}\int_{\mathbb{R}^N}\sum_{x\in B(q,p)} R((x,q,p),A)\varphi((x,q,p))\mathrm{d}p\mathrm{d}q$$

We defer the proof of this theorem to the supplementary materials. This theorem immediately implies that the momentum sampler preserves the joint invariant distribution φ.

In each iteration, since we can sample arbitrary values for the momentum variable p, we can affect the location variable q in arbitrary ways. This, coupled with the irreducibility of the proposal distributions $Q_i, i = 1,\ldots,N$, can typically guarantee the irreducibility of the momentum sampler. The potential problem of periodicity can be solved by randomly choosing the travel time T for each *TakeMomentumStep* call. The above argument establishes the irreducibility and aperiodicity of the momentum sampler, and proves the momentum sampler samples from the correct distribution $\pi(x)$.

2.3 Various Trade-offs

2.3.1 Trade-offs in Picking Single-site MH Proposal Distributions

To understand the trade-offs in picking different single-site MH proposal distributions, we look at the key step in Algorithm 2: $\Delta E \leftarrow \log \frac{\pi(x)Q_j(\tilde{x}|x)}{\pi(\tilde{x})Q_j(x|\tilde{x})}$. Note that $\frac{\pi(x)Q_j(\tilde{x}|x)}{\pi(\tilde{x})Q_j(x|\tilde{x})}$ also appears in MH acceptance probability. Intuitively, it measures how far the Q_i's are from detailed balance.

The main trade-offs in picking the Q_i's are between how far the Q_i's are from detailed balance and how much information goes into the auxiliary momentum process:

- On one extreme, we have the Gibbs sampler, which perfectly satisfies detailed balance. In this case, ΔE is always 0, no information goes into the auxiliary momentum process, and the momentum sampler is equivalent to the systematic scan Gibbs sampler.
- On the other extreme, we have the random walk Metropolis sampler, which is "furthest away" from detailed balance. In this case, $\Delta E = \log \frac{\pi(x)}{\pi(\tilde{x})}$, and arguably the largest amount of information goes into the auxiliary momentum process.
- In between the two extremes, we have various other choices. In many cases, we want the Q_i's to be "informed" to some degree. Similar problems have recently been considered in [10]. Thinking physically, these informed proposals can be seen as providing the analog of gradients information for discrete distributions. Depending on how far the Q_i's are from detailed balance, different amounts of information go into the auxiliary momentum process.

2.3.2 Trade-offs in Picking Kinetic Energies

We start with some illuminating calculations. Given $K_\beta(z) = |z|^\beta$, the velocity is given by $K'_\beta(z) = \mathrm{sign}(z)\beta|z|^{\beta-1}$. Regarding the magnitude of the velocity $v = |K'_\beta(z)|$ as a function of $k = K_\beta(z)$ gives us $v(k) = \beta k^{1-\frac{1}{\beta}}$. Taking the derivative of this function gives us $v'(k) = (\beta-1)k^{-\frac{1}{\beta}}$.

The main trade-offs in picking kinetic energies are between how the velocity is affected by changes in the kinetic energy and how much kinetic energy we typically have:

- One the one hand, to capture the momentum behavior of HMC, it's desirable to have larger velocity when the kinetic energy increases. This requires K_β to be convex, or more precisely $\beta \geq 1$. Furthermore, we want the changes in velocity to be sensitive to the changes in the kinetic energy. This typically points to having a larger β.
- On the other hand, we need to make use of the kinetic energy to overcome energy barriers ΔE, and intuitively, it's desirable to typically have large kinetic energy. Calculating the expected kinetic energy $\mathbb{E}_{p\sim G(0,1,\beta)}[K_\beta(p)] = \frac{2}{\beta}$ points to having a smaller β.

3 Connections to Existing Methods

3.1 Binary HMC Samplers are Special Cases of the Momentum Sampler

In this section, we prove that the Gaussian/exponential binary HMC samplers proposed in [5] are special cases of the momentum sampler. We use "visit site j" to denote hitting 0 at site j for the binary HMC samplers, and hitting 0 or 1 at site j for the momentum sampler. The key to proving this result is to understand that, in essence, the auxiliary momentum process in the momentum sampler and the embedding into the continuum in the binary HMC samplers are modeling the same two things: the time it takes to visit a site again (or more practically, at which site are we going to make the next proposal?), and how much total (potential and kinetic) energy we have at each site.

Formally, in order to establish the equivalences between the binary HMC samplers and certain momentum samplers, for a particular site j, we need to study:

1. The distribution on the initial time it takes to visit site j
2. The distribution on the initial total energy at site j
3. After we visit site j, the time it takes to visit site j again if we have total energy k at site j

For two samplers, if we can establish the equivalence of the joint distribution on the initial time and initial total energy, and the equivalence of the time it takes to visit this site again, we would be able to establish the equivalence of the corresponding samplers.

To simplify the proof, for the momentum sampler, rather than using the discrete state space $\Omega = \{1, 2\}^N$, we work with $\Omega = \{-1, 1\}^N$ instead. For the Gaussian binary HMC sampler, we have

Proposition 1. *The Gaussian binary HMC sampler is equivalent to the momentum sampler with deterministic proposals* $Q_i(\tilde{x}|x) = \begin{cases} 1 & \tilde{x}_i = -x_i, \tilde{x}_j = x_j, \forall j \neq i \\ 0 & \text{otherwise} \end{cases}, i = 1, \ldots, N$ *and* $\beta = 1$.

As a by-product of the above proposition, we also have

Proposition 2. *If, in addition to resampling p, we also resample q at every iteration, then the Gaussian binary HMC sampler with travel time $T = \pi$ is equivalent to a systematic scan single-site random walk Metropolis sampler, where the site-visitation order is a random permutation of $1, \ldots, N$ and is refreshed after every N site visits.*

For the original Gaussian binary HMC sampler with travel time $T = \pi$, since we don't resample q at each iteration, its behavior is not exactly the same as a systematic scan single-site random walk Metropolis sampler. Nevertheless, it's easy to see that these two samplers are closely related.

For the exponential binary HMC sampler, we have

Proposition 3. *The exponential binary HMC sampler is equivalent to the momentum sampler with deterministic proposal* $Q_i(\tilde{x}|x) = \begin{cases} 1 & \tilde{x}_i = -x_i, \tilde{x}_j = x_j, \forall j \neq i \\ 0 & \text{otherwise} \end{cases}, i = 1, \ldots, N$ *and* $\beta = \frac{2}{3}$.

We defer the proofs of these propositions to the supplementary materials. These results also explain the observations in [9] that the exponential binary HMC sampler is less efficient than the Gaussian binary HMC sampler: looking back at the trade-offs discussed in Section 2.3.2, it's easy to see that using $\beta = \frac{2}{3}$ induces undesirable momentum behavior (where we decelerate in high probability/low energy regions and accelerate in low probability/high energy regions), makes it easy for the process to get trapped in local energy minima, and makes the algorithm less efficient.

3.2 The Momentum Sampler Generalizes Discontinuous and Probabilistic Path HMCs

A foundational tool in applying HMC to discrete distributions is the discontinuous variant [9, 6] of HMC, which operates on piecewise continuous potential energies. The discontinuous HMC was first proposed in [9], where the authors considered a simple piecewise continuous potential energy arising from embedding binary distributions into the continuum. The Hamiltonian dynamics was integrated exactly across the discontinuities. This was later generalized in [6] to arbitrary piecewise continuous potential energies with discontinuities across affine boundaries, and a modified Leapfrog algorithm was introduced to account for such discontinuities.

In [2], the authors introduced a further generalization, the probabilistic path variant of HMC. In addition to its successful applications to the space of phylogenetic trees, an important contribution of probabilistic path HMC is the introduction of probabilistic paths to Hamiltonian dynamics, mixing the stochastic evolution of discrete variables with the deterministic evolution of continuous variables. However, the authors only considered a simple uniform proposal distribution for the evolution of the discrete variables. Moreover, a version of the probabilistic path HMC algorithm which allows for discontinuities in the potential energy was presented, but detailed balance, irreducibility and aperiodicity were only established for the case of continuous potential energy.

The momentum sampler generalizes discontinuous [9, 6] and probabilistic path [2] HMCs, by introducing probabilistic paths that can evolve according to arbitrary proposal distributions to Hamiltonian dynamics, while properly accounting for the corresponding discontinuities in the potential energy. Its detailed balance, irreducibility and aperiodicity are rigorously established. In the momentum sampler, we focus on a very simple auxiliary momentum process. However, it's easy to see that the essential idea is applicable to more general cases, in which we have complicated auxiliary state spaces and nontrivial potential energies. This points to a much more general framework where we can introduce interesting interactions between stochastic evolution of discrete variables and deterministic evolution of continuous variables.

4 Numerical Experiments on a 1D Potts Model

4.1 Experimental Setup

In this section, we carry out numerical experiments on a simple 1D 6-state Potts model of size $N = 400$ with periodic conditions, to demonstrate the behavior of the momentum sampler. More concretely, we consider the discrete distribution

$$\pi(x) \propto e^{J\sum_{i=1}^{400} \mathbb{1}_{\{x_i = x_{i+1}\}}}, \text{ where } x_i \in \{1, \dots, 6\}, i = 1, \dots, 400, \text{ and } x_{401} \text{ refers to } x_1$$

J controls the strength of correlations among neighboring locations. In our experiments, we use $J = 5$, which induces relatively strong correlations among neighboring locations.

For the momentum sampler, following [10], we consider 4 different proposals: RW, GB, LB1, LB2. For $j = 1, \dots, 400$, the four different proposals share the common form

$$Q_j(\tilde{x}|x) \propto g\left(\frac{\pi(\tilde{x})}{\pi(x)}\right) \rho_j(\tilde{x}|x), \text{ where } \rho_j(\tilde{x}|x) = \begin{cases} 1 & \text{if } \tilde{x}_j \neq x_j, \tilde{x}_i = x_i, i \neq j \\ 0 & \text{otherwise} \end{cases}$$

For RW, GB, LB1, LB2, $g(t)$ takes the form $1, t, \sqrt{t}, \frac{t}{1+t}$ respectively. We experiment with $\beta = \frac{2}{3}, 0.9, 0.95, 1, 1.05, 1.1, \frac{4}{3}$ and travel time $T = 1, 1.5, 2, 2.5, 3, 4, 5$.

For comparison, we also test four basic single-site MH samplers: random/systematic scan Gibbs samplers and random/systematic scan MH samplers with RW, GB, LB1, LB2 proposals (which we call "informed samplers"). For random scan samplers, at each iteration, we randomly pick a site j and make a proposal using Q_j, while for systematic scan samplers, we repeatedly pick a random order and visit all 400 sites in this order to make proposals. Note that our informed samplers are different from the samplers considered in [10]: in [10], they use proposal distributions of the form

$$Q(\tilde{x}|x) \propto g\left(\frac{\pi(\tilde{x})}{\pi(x)}\right) \mathbb{1}_{\left\{\tilde{x} \in \left\{y : \sum_{i=1}^{400} \mathbb{1}_{\{y_i \neq x_i\}} = 1\right\}\right\}}$$

which simultaneously pick a site to visit and propose a new state for this site; for our informed samplers, we rely on the random/systematic scan to pick a site j, and use Q_j to make the proposal.

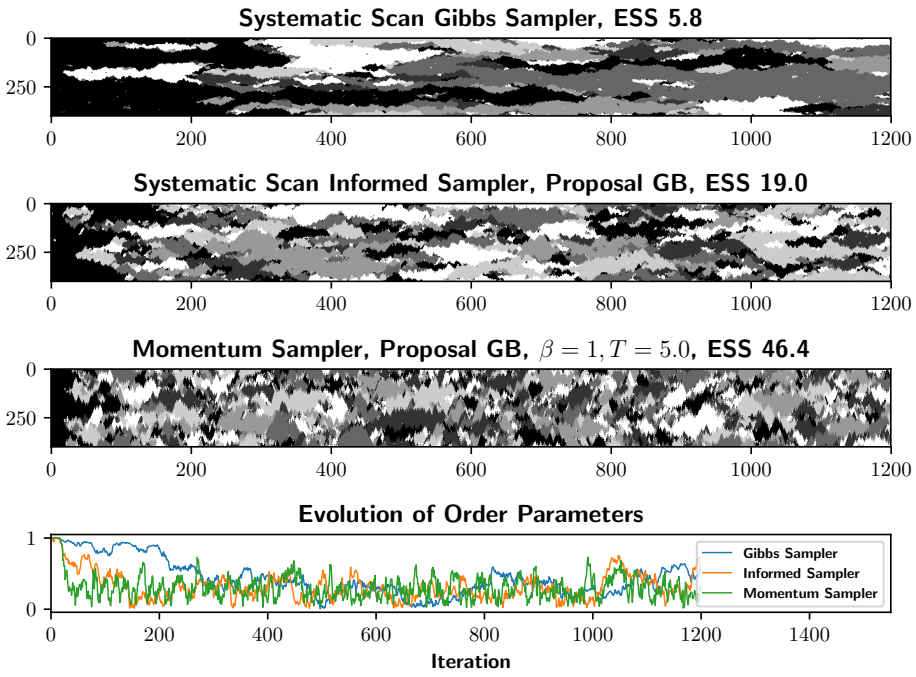

Figure 1: **Evolution of States and Order Parameters.** Visualizations of the evolution of states and order parameters for 3 best-performing samplers: the systematic scan Gibbs sampler, the systematic scan informed sampler with proposal GB, and the momentum sampler with proposal GB, $\beta = 1$, and travel time $T = 5$. The corresponding ESS's of the order parameters (larger is better) are shown.

For the Potts model, a useful summary statistic is the "order parameter". For a particular $x \in \Omega$, the order parameter is defined as $\frac{1}{400}\left|\sum_{j=1}^{400} e^{\frac{i2\pi x_j}{6}}\right|$. The order parameter ranges from 0 (totally disordered) to 1 (totally ordered). In our experiments, for each run of the sampler, we would also calculate the corresponding sequence of order parameters.

To make a fair comparison of different samplers, we run each sampler until they make approximately 2.4×10^6 site visits. For the momentum sampler with travel time T, this translates to approximately (exactly for $\beta = 1$) $6000/T$ samples. We use the effective sample size (ESS) (of the order parameters) per 2.4×10^6 site visits as our performance measure. The ESS is calculated using the *effective_sample_size* function in the *Arviz* python package [4]. To improve estimation accuracy, for a particular sampler, we use the order parameters from 50 different runs to estimate the ESS. In our experiments, we always initialize the system in a totally ordered state, where all x_i's are in state 1.

4.2 Results

Before going into the details of the experimental results, we want to emphasize that the goal of our experiments is **not** to argue for the efficiency of the momentum sampler, but rather to **demonstrate its behavior**. Since the momentum sampler is built on top of single-site MH samplers, it's not hard to imagine having more efficient cluster Monte Carlo algorithms [3]. Even if we only consider single-site MH algorithms, we can make use of the problem structure to design more efficient samplers: for our 1D Potts model, using a fixed sequential scan from 1 to 400 with GB can easily outperform all the above samplers. The momentum sampler provides a general framework to explore the idea of building discrete analogs of HMC, and the experiments here serve the purpose of demonstrating how its performance changes as we vary the travel time T, proposal distributions and β, and how the auxiliary momentum process affects the performances of the underlying single-site MH samplers.

Since we are working with a 1D model, we first visualize the evolution of the states and order parameters for three good-performing samplers: the systematic scan Gibbs sampler, the systematic scan informed sampler with proposal GB, and the momentum sampler with proposal GB, travel time $T = 5$ and $\beta = 1$. These are in fact the best-performing Gibbs/informed/momentum samplers.

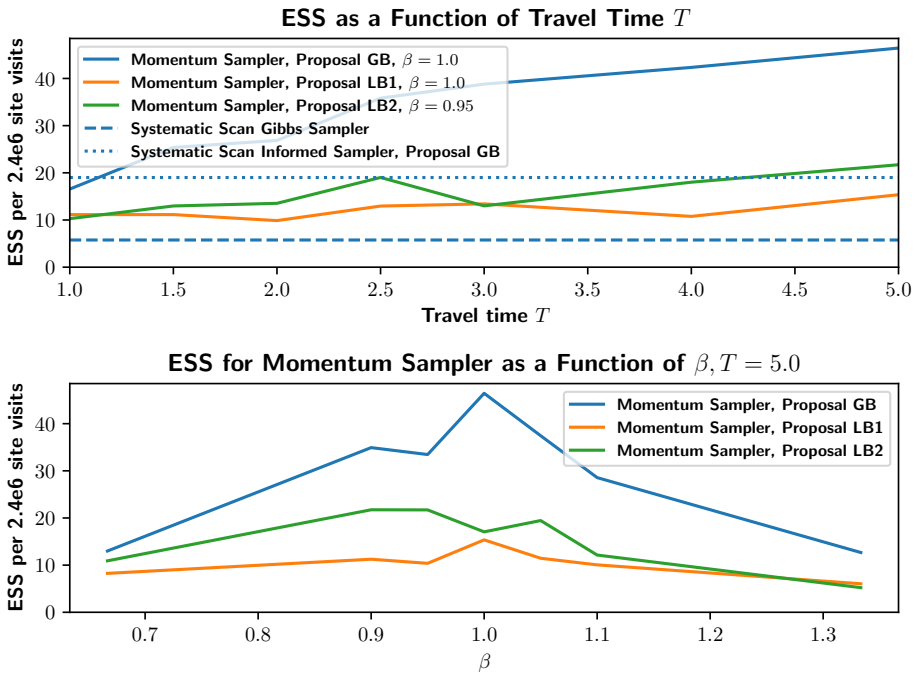

Figure 2: **ESS as a Function of travel time T, proposal distributions and β.** Upper figure: visualizations of how the ESS changes as a function of travel time T for the momentum sampler with proposals GB, LB1, LB2, and reference ESS's of systematic scan Gibbs sampler and systematic scan informed sampler with proposal GB. Lower figure: visualizations of how the ESS changes as a function of proposal distributions and β for the momentum sampler with proposals GB, LB1, LB2.

We can clearly see that the momentum sampler performs the best, improving upon the underlying single-site MH sampler (the informed sampler with proposal GB), while the Gibbs sampler performs the worst among the three different kinds of samplers.

To better understand the performance of the momentum sampler, we investigate how the ESS changes as a function of the travel time T, proposal distributions and β. The results are summarized in Figure 2. From the figure, we can see that for the momentum sampler with proposals GB, LB1 and LB2, ESS generally increases as T increases. This effect is particularly pronounced for GB, and is potentially a indicator of the momentum sampler's ability to make global proposals by traversing long isoprobability contours, much like HMC. We also see that the best performance for the momentum sampler generally happens at around $\beta = 1$, demonstrating the trade-offs discussed in Section 2.3.2.

5 Conclusions and Future Directions

In this paper, we propose the momentum sampler as a general framework for exploring the idea of building discrete analogs of HMC. It connects and generalizes a variety of existing methods, and points to a more general framework where we can mix stochastic evolution of discrete variables with deterministic evolution of continuous variables. Numerical experiments on a simple 1D Potts model demonstrate the value of the auxiliary momentum process for single-site MH samplers, and the various trade-offs involved in picking different proposal distributions and kinetic energies.

The 1D Potts model used in this paper only serves as a proof-of-concept, to demonstrate the behavior of the momentum sampler. Comprehensive numerical experiments on more complicated discrete distributions should be an immediate next step. In addition, working out the more general framework indicated by the momentum sampler in full generality and establish further connections with other existing methods remain an exciting future direction.

A Proof of Theorem 1

A.1 Proof of the Theorem

Theorem 1. ***(Detailed Balance)*** *The TakeMomentumStep function satisfies detailed balance w.r.t. the joint invariant distribution φ, i.e. for any measurable sets $A, B \subset \Sigma$,*

$$\int_{\mathbb{T}^N}\int_{\mathbb{R}^N}\sum_{x\in A(q,p)} R((x,q,p),B)\varphi((x,q,p))\mathrm{d}p\mathrm{d}q = \int_{\mathbb{T}^N}\int_{\mathbb{R}^N}\sum_{x\in B(q,p)} R((x,q,p),A)\varphi((x,q,p))\mathrm{d}p\mathrm{d}q$$

Proof. For notational simplicity, we use $s = (x, q, p)$ and $s' = (x', q', p')$ to denote two points in Σ.

Sequence of Proposals and Probabilistic Paths

If we start from $s \in \Sigma$, for a given travel time T, a concrete run of the *TakeMomentumStep* function would involve a finite sequence of random proposals. Assume the length of the sequence is M. The sequence of random proposals Y can be denoted as

$$Y = (y^{(0)}, y^{(1)}, \ldots, y^{(M-1)}), y^{(m)} \in \Omega, m = 0, \ldots, M-1$$

This sequence of proposals indicates that, for this particular run of *TakeMomentumStep*, we reach 0 or 1 at individual sites M times, and each time the system makes a proposal to go to the discrete state $y^{(m)} \in \Omega, m = 0, \cdots, M-1$ from the current discrete state.

If we fix Y, the *TakeMomentumStep* function in fact specifies a deterministic mapping, and would map s to a single point $s' \in \Sigma$. For each such sequence of proposals Y, we introduce an associated probabilistic path $\omega(s, T, Y)$, which contains all the information of the system going from s to s' in time T through the function *TakeMomentumStep*. Formally, $\omega(s, T, Y)$ is specified by

- The sequence of random proposals Y
$$Y = (y^{(0)}, y^{(1)}, \ldots, y^{(M-1)}), y^{(m)} \in \Omega, m = 0, \ldots, M-1$$
- The indices of the sites for the M site visitations $j^{(0)}, j^{(1)}, \ldots, j^{(M-1)} \in \{1, \ldots, N\}$
- The times of the M site visitations $0 \leqslant t^{(0)} < t^{(1)} < \ldots < t^{(M-1)} \leqslant T$
- The discrete states of the system at M site visitations $x = x^{(0)}, x^{(1)}, \ldots, x^{(M-1)} \in \Omega$
- Accept/reject decisions for the M site visitations $a^{(m)} = \mathbb{1}_{\{y^{(m)}=x^{(m+1)}\}}$, where $x^{(M)} = x'$
- The evolution of the location variables $q(t)$ and the momentum variables $p(t), 0 \leqslant t \leqslant T$. Note that we might have discontinuities in $p(t)$. We use $p(t^-)$ to denote the left limit and $p(t^+)$ to denote the right limit.

Countable Number of Probabilistic Paths

In order for a probabilistic path $\omega(s, T, Y)$ to be valid, the different components of $\omega(s, T, Y)$ have to interact with each other in a way as determined by the *TakeMomentumStep* function. For example, we should have $y_i^{(m)} = x_i^{(m)}, \forall i \neq j^{(m)}$ and

$$x^{(m+1)} = \begin{cases} y^{(m)} & \text{if } K_\beta(p(t^{(m)-})) > \log \frac{\pi(x^{(m)})Q_{j^{(m)}}(y^{(m)}|x^{(m)})}{\pi(y^{(m)})Q_{j^{(m)}}(x^{(m)}|y^{(m)})} \\ x^{(m)} & \text{otherwise} \end{cases}$$

For $s \in \Sigma$ and some given travel time T, we say a sequence of proposals Y is compatible with s, T and *TakeMomentumStep* if we can find a corresponding probabilistic path $\omega(s, T, Y)$ that's valid.

Not all sequences of proposals correspond to valid probabilistic paths. But even if we don't consider the compatibility of the sequence of proposals with s, T and *TakeMomentumStep*, the set of all possible such sequences has only a countable number of elements. This is because we only need to

look at sequences of finite length (because of the fixed travel time T), and all the individual proposals are on discrete state spaces with a finite number of states.

The above analysis indicates that for some starting point $s \in \Sigma$ and travel time T, running the *TakeMomentumStep* function would result in only a countable number of possible destinations s'. Furthermore, $\forall s, s' \in \Sigma$ for which $R_T(s, \{s'\}) > 0$, there are at most a countable number of probabilistic paths which bring s to s' in time T through *TakeMomentumStep*.

Formally, given some travel time T and a sequence of proposals Y, define

$$\mathcal{D}(T, Y) = \{s \in \Sigma : Y \text{ is compatible with } s, T \text{ and } \textit{TakeMomentumStep}\}$$

Use $\mathcal{T}_{T,Y} : \mathcal{D}(T, Y) \to \Sigma$ to denote the deterministic mapping defined by *TakeMomentumStep* for the given Y in time T (so that $\mathcal{D}(T, Y)$ represents the domain of the mapping $\mathcal{T}_{T,Y}$), and use

$$\mathcal{I}(T, Y) = \{s' \in \Sigma : \exists s \in \mathcal{S}(T, Y), s.t. \mathcal{T}_{T,Y}(s) = s'\}$$

to denote the image of the mapping $\mathcal{T}_{T,Y}$. For a given $x \in \Omega$, use

$$\mathcal{T}_{T,Y,x} : \{(q, p) \in \mathbb{T}^N \times \mathbb{R}^N : s = (x, q, p) \in \mathcal{D}(T, Y)\} \to \mathbb{T}^N \times \mathbb{R}^N$$

to denote the deterministic mapping induced by $\mathcal{T}_{T,Y}$ on $\mathbb{T}^N \times \mathbb{R}^N$. (i.e. $\forall s = (x, q, p) \in \mathcal{D}(T, Y)$, $\mathcal{T}_{T,Y,x}((q, p)) = (q', p')$, where $s' = (x', q', p') = \mathcal{T}_{T,Y}(s)$). Define

$$\Sigma(s, T) = \{s' = (x', q', p') \in \Sigma : R_T(s, \{s'\}) > 0\}$$

$\forall s, s' \in \Sigma$ for which $R_T(s, \{s'\}) > 0$, further define

$$\mathcal{P}(s, s', T) = \{Y \text{ a sequence of proposals: } s \in \mathcal{D}(T, Y) \text{ and } \mathcal{T}_{T,Y}(s) = s'\}$$

Then both $\Sigma(s, T)$ and $\mathcal{P}(s, s', T)$ have at most a countable number of elements.

Proof of Detailed Balance

For a given travel time T and a sequence of proposals Y, $\forall s \in \mathcal{D}(T, Y)$, we use $r_{T,Y}(s, s')$ to denote the probability of going from s to s' through the probabilistic path $\omega(s, T, Y)$. Since *TakeMomentumStep* defines a deterministic mapping $\mathcal{T}_{T,Y}$ for given T and Y, the only non-zero term is $r_{T,Y}(s, \mathcal{T}_{T,Y}(s))$. For all $s' \neq \mathcal{T}_{T,Y}(s)$, we have $r_{T,Y}(s, s') = 0$.

Using the above notation, $\forall s \in A$ and $B \subset \Sigma$ measurable, we can write $R_T(s, B)$ as

$$\begin{aligned} R_T(s, B) &= \sum_{s' \in B \cap \Sigma(s,T)} R_T(s, \{s'\}) \\ &= \sum_{s' \in B \cap \Sigma(s,T)} \sum_{Y \in \mathcal{P}(s,s',T)} r_{T,Y}(s, s') \\ &= \sum_{s' \in B \cap \Sigma(s,T)} \sum_{Y \in \mathcal{P}(s,s',T)} r_{T,Y}(s, \mathcal{T}_{T,Y}(s)) \end{aligned}$$

For a given travel time T, $\forall s, s' \in \Sigma$, if $R_T(s, \{s'\}) > 0$, then $\mathcal{P}(s, s', T) \neq \emptyset$. In Lemma 3, we prove that $\forall Y \in \mathcal{P}(s, s', T)$, the absolute value of the determinant of the Jacobian of $\mathcal{T}_{T,Y,x}$ is $|\det \mathcal{J}\mathcal{T}_{T,Y,x}| = 1$, for all $x \in \Omega$ where $\mathcal{T}_{T,Y,x}$. Furthermore, the deterministic mapping $\mathcal{T}_{T,Y}$ is reversible, and there exists a sequence of proposals $\tilde{Y} \in \mathcal{P}(s', s, T)$, s.t. $s = \mathcal{T}_{T,Y}^{-1}(s') = \mathcal{T}_{T,\tilde{Y}}(s')$.

In Lemma 4, we prove that

$$\varphi(s) r_{T,Y}(s, s') = \varphi(s) r_{T,Y}(s, \mathcal{T}_{T,Y}(s)) = \varphi(s') r_{T,\tilde{Y}}(s', \mathcal{T}_{T,\tilde{Y}}(s')) = \varphi(s') r_{T,\tilde{Y}}(s', s)$$

Using the above results, it's not hard to see that

$$
\begin{aligned}
&\int_{\mathbb{T}^N}\int_{\mathbb{R}^N}\sum_{x\in A(q,p)} R_T(s,B)\varphi(s)\mathrm{d}p\mathrm{d}q \\
=&\int_{\mathbb{T}^N}\int_{\mathbb{R}^N}\sum_{x\in A(q,p)}\sum_{s'\in B\cap\Sigma(s,T)}\sum_{Y\in\mathcal{P}(s,s',T)} r_{T,Y}(s,s')\varphi(s)\mathrm{d}p\mathrm{d}q \\
=&\int_{\mathbb{T}^N}\int_{\mathbb{R}^N}\sum_{x\in A(q,p)}\sum_{s'\in B\cap\Sigma(s,T)}\sum_{Y\in\mathcal{P}(s,s',T)} r_{T,\tilde{Y}}(s',s)\varphi(s')\mathrm{d}p\mathrm{d}q \\
\overset{\text{change of variables}}{=}&\int_{\mathbb{T}^N}\int_{\mathbb{R}^N}\sum_{x'\in B(q',p')}\sum_{s\in A\cap\Sigma(s',T)}\sum_{\tilde{Y}\in\mathcal{P}(s',s,T)} r_{T,\tilde{Y}}(s',s)\varphi(s')\frac{1}{|det\mathcal{J}\mathcal{T}_{T,Y,x}|}\mathrm{d}p'\mathrm{d}q' \\
=&\int_{\mathbb{T}^N}\int_{\mathbb{R}^N}\sum_{x'\in B(q',p')} R_T(s',A)\varphi(s')\mathrm{d}p'\mathrm{d}q'
\end{aligned}
$$

which proves the desired detailed balance property of *TakeMomentumStep* w.r.t. φ

$$
\int_{\mathbb{T}^N}\int_{\mathbb{R}^N}\sum_{x\in A(q,p)} R((x,q,p),B)\varphi((x,q,p))\mathrm{d}p\mathrm{d}q = \int_{\mathbb{T}^N}\int_{\mathbb{R}^N}\sum_{x\in B(q,p)} R((x,q,p),A)\varphi((x,q,p))\mathrm{d}p\mathrm{d}q
$$

□

A.2 Useful Lemmas

In this section, we prove a few useful lemmas to complete the proof of Theorem 1.

First, we prove two lemmas, similar to Lemma 1 and Lemma 2 in Section 5.1 of [6].

Lemma 1. ***(Refraction)*** *Let $\mathcal{T}:\mathbb{T}^N\times\mathbb{R}^N\to\mathbb{T}^N\times\mathbb{R}^N$ be a transformation in $\mathbb{T}^N$ that takes a unit mass located at $q=(q_1,\ldots,q_N)$ and moves it with constant velocity $v=(K'_\beta(p_1),\ldots,K'_\beta(p_N))$. Assume it reaches 0 or 1 at site j first. Subsequently q_j is changed to $1-q_j$, and p_j is changed to $sign(p_j)K_\beta^{-1}(K_\beta(p_j)-\Delta E)$ (where ΔE is a constant and satisfies $\Delta E<K_\beta(p_j)$). The move is carried on, with the velocity v_j changed to $K'_\beta(sign(p_j)K_\beta^{-1}(K_\beta(p_j)-\Delta E))$, for the total time period τ till it ends in location q' and momentum p', before it reaches 0 or 1 again at any sites. Then $\mathcal{T}$ is volume preserving, i.e. the absolute value of the determinant of its Jacobian $|det\mathcal{J}\mathcal{T}|=1$.*

Proof. Following the same argument as in the proof of Lemma 1 of [6], we have

$$
|\mathrm{det}\mathcal{J}\mathcal{T}| = \left|\mathrm{det}\begin{pmatrix} \frac{\partial q'_j}{\partial q_j} & \frac{\partial q'_j}{\partial p_j} \\ \frac{\partial p'_j}{\partial q_j} & \frac{\partial p'_j}{\partial p_j}\end{pmatrix}\right|
$$

If we define $t_j=\frac{\mathrm{sign}(v_j)+1-2q_j}{2K'_\beta(p_j)}=\frac{\mathrm{sign}(p_j)+1-2q_j}{2K'_\beta(p_j)}$, then

$$
\begin{aligned}
p'_j &= \mathrm{sign}(p_j)K_\beta^{-1}(K_\beta(p_j)-\Delta E) \\
q'_j &= \frac{1-\mathrm{sign}(p_j)}{2}+K'_\beta(p'_j)(\tau-t_j) \\
&= \frac{1-\mathrm{sign}(p_j)}{2}+K'_\beta(p'_j)\left(\tau-\frac{\mathrm{sign}(p_j)+1-2q_j}{2K'_\beta(p_j)}\right)
\end{aligned}
$$

This implies

$$
|\mathrm{det}\mathcal{J}\mathcal{T}| = \left|\mathrm{det}\begin{pmatrix} \frac{\partial q'_j}{\partial q_j} & \frac{\partial q'_j}{\partial p_j} \\ \frac{\partial p'_j}{\partial q_j} & \frac{\partial p'_j}{\partial p_j}\end{pmatrix}\right| = \left|\mathrm{det}\begin{pmatrix} \frac{\partial q'_j}{\partial q_j} & \frac{\partial q'_j}{\partial p_j} \\ 0 & \frac{\partial p'_j}{\partial p_j}\end{pmatrix}\right| = \left|\frac{\partial q'_j}{\partial q_j}\frac{\partial p'_j}{\partial p_j}\right| = \left|\frac{K'_\beta(p'_j)}{K'_\beta(p_j)}\frac{K'_\beta(p_j)}{K'_\beta(p'_j)}\right| = 1
$$

□

Lemma 2. ***(Reflection)*** *Let $\mathcal{T} : \mathbb{T}^N \times \mathbb{R}^N \to \mathbb{T}^N \times \mathbb{R}^N$ be a transformation in $\mathbb{T}^N$ that takes a unit mass located at $q = (q_1, \ldots, q_N)$ and moves it with constant velocity $v = (K'_\beta(p_1), \ldots, K'_\beta(p_N))$. Assume it reaches 0 or 1 at site j first. Subsequently p_j is changed to $-p_j$. The move is carried on, with the velocity v_j changed to $-v_j$, for the total time period τ till it ends in location q' and momentum p', before it reaches 0 or 1 at any sites again. Then $\mathcal{T}$ is volume preserving, i.e. the absolute value of the determinant of its Jacobian $|det\mathcal{J}\mathcal{T}| = 1$.*

Proof. Following the same argument as in the proof of Lemma 2 of [6], we have

$$|\text{det}\mathcal{J}\mathcal{T}| = \left|\text{det}\begin{pmatrix} \frac{\partial q'_j}{\partial q_j} & \frac{\partial q'_j}{\partial p_j} \\ \frac{\partial p'_j}{\partial q_j} & \frac{\partial p'_j}{\partial p_j} \end{pmatrix}\right|$$

If we define $t_j = \frac{\text{sign}(v_j)+1-2q_j}{2K'_\beta(p_j)} = \frac{\text{sign}(p_j)+1-2q_j}{2K'_\beta(p_j)}$, then

$$\begin{aligned} p'_j &= -p_j \\ q'_j &= \frac{1+\text{sign}(p_j)}{2} - K'_\beta(p_j)(\tau - t_j) \\ &= \frac{1+\text{sign}(p_j)}{2} - K'_\beta(p_j)\left(\tau - \frac{\text{sign}(p_j)+1-2q_j}{2K'_\beta(p_j)}\right) \\ &= 1 + \text{sign}(p_j) - K'_\beta(p_j)\tau - q_j \end{aligned}$$

This implies

$$|\text{det}\mathcal{J}\mathcal{T}| = \left|\text{det}\begin{pmatrix} \frac{\partial q'_j}{\partial q_j} & \frac{\partial q'_j}{\partial p_j} \\ \frac{\partial p'_j}{\partial q_j} & \frac{\partial p'_j}{\partial p_j} \end{pmatrix}\right| = \left|\text{det}\begin{pmatrix} -1 & \frac{\partial q'_j}{\partial p_j} \\ 0 & -1 \end{pmatrix}\right| = 1$$

□

Lemma 3. *Given travel time T, $\forall s, s' \in \Sigma$ for which $R_T(s, \{s'\}) > 0$, $\mathcal{P}(s, s', T) \neq \emptyset$. $\forall Y \in \mathcal{P}(s, s', T)$, the absolute value of the determinant of the Jacobian of $\mathcal{T}_{T,Y,x}$ is $|det\mathcal{J}\mathcal{T}_{T,Y,x}| = 1$, for all $x \in \Omega$ where $\mathcal{T}_{T,Y,x}$ is well-defined. Furthermore, the deterministic mapping $\mathcal{T}_{T,Y}$ is reversible, and there exists a sequence of proposals $\tilde{Y} \in \mathcal{P}(s', s, T)$, s.t. $s = \mathcal{T}^{-1}_{T,Y}(s') = \mathcal{T}_{T,\tilde{Y}}(s')$*

Proof. Given travel time T, $\forall s, s' \in \Sigma$, if $R_T(s, \{s'\}) > 0$, then by definition $\mathcal{P}(s, s', T) \neq \emptyset$. $\forall Y \in \mathcal{P}(s, s', Y)$, for some $x \in \Omega$, if the deterministic mapping $\mathcal{T}_{T,Y,x}$ is well-defined, then $\mathcal{T}_{T,Y,x}$ can be be written as the composition of a sequence of deterministic mappings

$$\mathcal{T}_{T,Y,x} = \mathcal{T}^{(0)}_{T,Y,x} \circ \mathcal{T}^{(1)}_{T,Y,x} \circ \cdots \circ \mathcal{T}^{(M-1)}_{T,Y,x}$$

where the mappings $\mathcal{T}^{(m)}_{T,Y,x}, m = 0, \ldots, M-1$ are either a refraction mapping as described in Lemma 1, or a reflection mapping as described in Lemma 2. Using Lemma 1 and Lemma 2, it's easy to see that the absolute value of the determinant of the Jacobian

$$|\text{det}\mathcal{J}\mathcal{T}_{T,Y,x}| = \prod_{m=0}^{M-1} |\text{det}\mathcal{J}\mathcal{T}^{(m)}_{T,Y,x}| = 1$$

$\forall Y \in \mathcal{P}(s, s', Y)$, define a new sequence of proposals $\tilde{Y} = (\tilde{y}^{(0)}, \tilde{y}^{(1)}, \ldots, \tilde{y}^{(M-1)})$ where

$$\tilde{y}^{(m)} = \begin{cases} x^{(M-m-1)} & \text{if } a^{(M-m-1)} = 1 (\text{i.e. } y^{(M-m-1)} = x^{(M-m)}) \\ y^{(M-m-1)} & \text{otherwise (i.e. } y^{(M-m-1)} \neq x^{(M-m)}, \text{which means } x^{(M-m-1)} = x^{(M-m)}) \end{cases}$$

We claim that $\tilde{Y} \in \mathcal{P}(s, s', T)$, and $\mathcal{T}_{T,\tilde{Y}}(s') = s$. To see $\tilde{Y}$ has these desired properties, we look at its corresponding probabilistic path $\omega(s', T, \tilde{Y})$. The corresponding discrete states of the system at M site visitations $\tilde{x}^{(m)}, m = 0, \ldots, M$ and the indices of the sites for the M site visitations

$\tilde{j}^{(m)}, m = 0, \ldots, M-1$ are given by simple reversals of the original sequence of discrete states $x^{(m)}, m = 0, \ldots, M$ and the original sequence of indices for visited sites $j^{(m)}, m = 0, \ldots, M-1$:

$$\begin{aligned} \tilde{j}^{(m)} &= j^{(M-m-1)}, m = 0, \ldots, M-1 \\ \tilde{x}^{(m)} &= x^{(M-m)}, m = 0, \ldots, M \end{aligned}$$

The corresponding sequence of accept/reject decisions $\tilde{a}^{(m)}, m = 0, \ldots, M-1$ is also a simple reversal of the original sequence of accept/reject decisions $a^{(m)}, m = 0, \ldots, M-1$

$$\tilde{a}^{(m)} = \mathbb{1}_{\{\tilde{y}^{(m)} = \tilde{x}^{(m+1)}\}} = \begin{cases} \mathbb{1}_{\{x^{(M-m-1)} = x^{(M-m-1)}\}} = 1 & \text{if } a^{(M-m-1)} = 1 \\ \mathbb{1}_{\{y^{(M-m-1)} = x^{(M-m-1)}\}} = 0 & \text{if } a^{(M-m-1)} = 0 \end{cases} = a^{(M-m-1)}$$

It's straightforward to verify that $\omega(s', T, \tilde{Y})$ is a valid probabilistic path that brings s' back to s in time T through *TakeMomentumStep*. In particular, note the importance of the momentum negating step $p \leftarrow -p$ in ensuring the existence of such a probabilistic path. This proves our claim.

□

Lemma 4. *$\forall s, s' \in \Sigma$ for which $R_T(s, \{s'\}) > 0$, for $Y \in \mathcal{P}(s, s', T)$, we have*

$$\varphi(s) r_{T,Y}(s, s') = \varphi(s) r_{T,Y}(s, \mathcal{T}_{T,Y}(s)) = \varphi(s') r_{T,\tilde{Y}}(s', \mathcal{T}_{T,\tilde{Y}}(s')) = \varphi(s') r_{T,\tilde{Y}}(s', s)$$

where $\tilde{Y}$ is defined as in Lemma 3.

Proof. We can directly calculate the transition probability $r_{T,Y}(s, s')$:

$$r_{T,Y}(s, s') = \prod_{m=0}^{M-1} Q_{j^{(m)}}(y^{(m)} | x^{(m)})$$

Correspondingly, we can also calculate the transition probability $r_{T,\tilde{Y}}(s', s)$:

$$r_{T,\tilde{Y}}(s', s) = \prod_{m=0}^{M-1} Q_{\tilde{j}^{(m)}}(\tilde{y}^{(m)} | \tilde{x}^{(m)})$$

Note that

$$\begin{aligned} r_{T,Y}(s, s') &= \prod_{m=0}^{M-1} Q_{j^{(m)}}^{a^{(m)}}(y^{(m)} | x^{(m)}) \prod_{m=0}^{M-1} Q_{j^{(m)}}^{1-a^{(m)}}(y^{(m)} | x^{(m)}) \\ &= \prod_{m:a^{(m)}=1} Q_{j^{(m)}}(y^{(m)} | x^{(m)}) \prod_{m:a^{(m)}=0} Q_{j^{(m)}}(y^{(m)} | x^{(m)}) \\ r_{T,\tilde{Y}}(s', s) &= \prod_{m=0}^{M-1} Q_{\tilde{j}^{(m)}}^{\tilde{a}^{(m)}}(\tilde{y}^{(m)} | \tilde{x}^{(m)}) \prod_{m=0}^{M-1} Q_{\tilde{j}^{(m)}}^{1-\tilde{a}^{(m)}}(\tilde{y}^{(m)} | \tilde{x}^{(m)}) \\ &= \prod_{m:\tilde{a}^{(m)}=1} Q_{\tilde{j}^{(m)}}(\tilde{y}^{(m)} | \tilde{x}^{(m)}) \prod_{m:\tilde{a}^{(m)}=0} Q_{\tilde{j}^{(m)}}(\tilde{y}^{(m)} | \tilde{x}^{(m)}) \\ &= \prod_{m:a^{(M-m-1)}=1} Q_{j^{(M-m-1)}}(x^{(M-m-1)} | y^{(M-m-1)}) \\ &\times \prod_{m:a^{(M-m-1)}=0} Q_{j^{(M-m-1)}}(y^{(M-m-1)} | x^{(M-m)}) \\ &= \prod_{m:a^{(M-m-1)}=1} Q_{j^{(M-m-1)}}(x^{(M-m-1)} | y^{(M-m-1)}) \\ &\times \prod_{m:a^{(M-m-1)}=0} Q_{j^{(M-m-1)}}(y^{(M-m-1)} | x^{(M-m-1)}) \\ &= \prod_{m:a^{(m)}=1} Q_{j^{(m)}}(x^{(m)} | y^{(m)}) \prod_{m:a^{(m)}=0} Q_{j^{(m)}}(y^{(m)} | x^{(m)}) \end{aligned}$$

By following the probabilistic path $\omega(s, T, Y)$, it's not hard to see that

$$\sum_{j=1}^{N}[K_\beta(p'_j) - K_\beta(p_j)] = -\sum_{m:a^{(m)}=1} \log \frac{\pi(x^{(m)})Q_{j^{(m)}}(y^{(m)}|x^{(m)})}{\pi(y^{(m)})Q_{j^{(m)}}(x^{(m)}|y^{(m)})}$$

Using the above equations, it's easy to see that

$$\begin{aligned}
& \frac{\varphi(s)r_{T,Y}(s,s')}{\varphi(s')r_{T,\tilde{Y}}(s',s)} \\
= & \frac{\pi(x)e^{-\sum_{j=1}^{N} K_\beta(p_j)}r_Y(s,s')}{\pi(x')e^{-\sum_{j=1}^{N} K_\beta(p'_j)}r_{\tilde{Y}}(s',s)} \\
= & \frac{\pi(x)}{\pi(x')} \prod_{m:a^{(m)}=1} \frac{\pi(y^{(m)})Q_{j^{(m)}}(x^{(m)}|y^{(m)})}{\pi(x^{(m)})Q_{j^{(m)}}(y^{(m)}|x^{(m)})} \frac{Q_{j^{(m)}}(y^{(m)}|x^{(m)})}{Q_{j^{(m)}}(x^{(m)}|y^{(m)})} \\
= & 1
\end{aligned}$$

□

B Binary HMC Samplers are Special Cases of the Momentum Sampler

As we've already mentioned in the main text, in order to show that the binary HMC samplers are special cases of the momentum sampler, we need to study:

1. For site j, the distribution on the initial time it takes to visit site j, which we denote by $t_j^{(0)}$.
   - As shown in Algorithm 2, in the momentum sampler,

     $$t_j^{(0)} = \frac{\text{sign}(v_j^{(0)}) + 1 - 2q_j^{(0)}}{2v_j^{(0)}}$$

     where $v_j^{(0)} = K'_\beta(p_j^{(0)}) = \text{sign}(p_j^{(0)})\beta|p_j^{(0)}|^{\beta-1}$ is the velocity at site j, and

     $$q_j^{(0)} \sim U([0,1]), p_j^{(0)} \sim G(0,1,\beta)$$

   - For the Gaussian binary HMC sampler,

     $$t_j^{(0)} = \begin{cases} -\arctan\left(\frac{q_j^{(0)}}{p_j^{(0)}}\right) & \text{if } \frac{q_j^{(0)}}{p_j^{(0)}} \leqslant 0 \\ \pi - \arctan\left(\frac{q_j^{(0)}}{p_j^{(0)}}\right) & \text{if } \frac{q_j^{(0)}}{p_j^{(0)}} > 0 \end{cases}$$

     where $q_j^{(0)}, p_j^{(0)} \sim N(0,1)$.
   - For the exponential binary HMC sampler,

     $$t_j^{(0)} = p_j^{(0)} + \sqrt{(p_j^{(0)})^2 + 2q_j^{(0)}}$$

     where $q_j^{(0)} \sim \exp(1), p_j^{(0)} \sim N(0,1)$.
2. For site j, the distribution on the initial total energy, which we denote by $k_j^{(0)}$.
   - For the momentum sampler, $k_j^{(0)} = K_\beta(p_j^{(0)})$, where $p_j^{(0)} \sim G(0,1,\beta)$.
   - For the Gaussian binary HMC sampler,

     $$k_j^{(0)} = \frac{1}{2}(q_j^{(0)})^2 + \frac{1}{2}(p_j^{(0)})^2$$

     where $q_j^{(0)}, p_j^{(0)} \sim N(0,1)$.

- For the exponential binary HMC sampler,

$$k_j^{(0)} = q_j^{(0)} + \frac{1}{2}(p_j^{(0)})^2$$

where $q_j^{(0)} \sim \exp(1), p_j^{(0)} \sim N(0,1)$.

3. For site j, after we reach 0 or 1, if we have total energy k, the time it takes to hit a boundary again at this site. We denote this time by $t_j(k)$.
   - For the momentum sampler, $t_j(k) = \frac{1}{\beta k^{1-\frac{1}{\beta}}}$
   - For the Gaussian binary HMC, $t_j(k) = \pi$
   - For the exponential binary HMC, $t_j(k) = 2\sqrt{2k}$

Since different dimensions are independent of each other, we only need to look at one particular dimension j. We can prove the corresponding propositions if we can establish suitable equivalence concerning the joint distribution on $(t_j^{(0)}, k_j^{(0)})$, and the function $t_j(k)$.

B.1 Proof of Proposition 1

In order to prove Proposition 1, we first prove a lemma

Lemma 5. *Assume $q, p \sim N(0,1)$ are two independent standard normal random variables. Then $\frac{q}{p}$ and q^2+p^2 are independent. Furthermore, $\arctan\left(\frac{q}{p}\right)$ follows the uniform distribution $U\left(\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]\right)$, and $\frac{q^2+p^2}{2}$ follows the exponential distribution $\exp(1)$.*

Proof. We calculate the characteristic function of the random vector $\left(\frac{q}{p}, q^2+p^2\right)$:

$$\begin{aligned}
& \mathbb{E}_{q,p\sim N(0,1)}\left[e^{i\left[t_1\frac{q}{p}+t_2(q^2+p^2)\right]}\right] \\
= & \frac{1}{2\pi}\int_{\mathbb{R}^2} e^{it_1\frac{q}{p}+it_2(q^2+p^2)} e^{-\frac{q^2+p^2}{2}} \mathrm{d}q\mathrm{d}p \\
= & \frac{1}{2\pi}\int_0^{+\infty}\int_0^{2\pi} e^{it_1\tan\theta} e^{it_2 r^2} e^{-\frac{r^2}{2}} r\mathrm{d}r\mathrm{d}\theta \\
= & \left[\int_0^{2\pi} e^{it_1\tan\theta}\frac{1}{2\pi}\mathrm{d}\theta\right]\left[\int_0^{+\infty} e^{it_2r^2-\frac{r^2}{2}} r\mathrm{d}r\right] \\
= & \left[\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} e^{it_1\tan\theta}\frac{1}{\pi}\mathrm{d}\theta\right]\left[\int_0^{+\infty} e^{it_2x}\frac{1}{2}e^{-2x}\mathrm{d}x\right] \\
= & \left[\int_{-\infty}^{+\infty} e^{it_1x}\frac{1}{\pi(1+x^2)}\mathrm{d}x\right]\left[\int_0^{+\infty} e^{it_2x}\frac{1}{2}e^{-2x}\mathrm{d}x\right] \\
= & \mathbb{E}_{x\sim\text{Cauchy}(0,1)}[e^{it_1x}]\mathbb{E}_{x\sim\exp(2)}[e^{it_2x}]
\end{aligned}$$

This calculation implies that $\frac{q}{p}$ and q^2+p^2 are independent, and that $\frac{q}{p} \sim \text{Cauchy}(0,1)$, $q^2+p^2 \sim \exp(2)$. Since the cumulative distribution function (CDF) of $\text{Cauchy}(0,1)$ is given by

$$\frac{1}{\pi}\arctan(x) + \frac{1}{2}$$

we have $\frac{1}{\pi}\arctan\left(\frac{q}{p}\right) + \frac{1}{2} \sim U([0,1])$, which implies that $\arctan\left(\frac{q}{p}\right) \sim U\left(\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]\right)$. From $q^2+p^2 \sim \exp(2)$, it's easy to deduce that $\frac{q^2+p^2}{2} \sim \exp(1)$. □

Proposition 1. *The Gaussian binary HMC sampler is equivalent to the momentum sampler with deterministic proposals* $Q_i(\tilde{x}|x) = \begin{cases} 1 & \tilde{x}_i = -x_i, \tilde{x}_j = x_j, \forall j \neq i \\ 0 & \text{otherwise} \end{cases}$ *, $i = 1, \ldots, N$ and $\beta = 1$.*

Proof. For the Gaussian binary HMC sampler, using Lemma 5 and the expressions we derived in Section B, given a dimension j, it's easy to see that $t_j^{(0)}$ and $k_j^{(0)}$ are independent, and that $t_j^{(0)} \sim U([0,\pi])$, $k_j^{(0)} \sim \exp(1)$. For the momentum sampler with $\beta = 1$, it's easy to see that we also have $t_j^{(0)}$ and $k_j^{(0)}$ are independent, and that $t_j^{(0)} \sim U([0,1])$, $k_j^{(0)} \sim \exp(1)$. This implies that the random vector $\left(\frac{t_j^{(0)}}{\pi}, k_j^{(0)}\right)$ from the Gaussian binary HMC sampler has the same joint distribution as the random vector $(t_j^{(0)}, k_j^{(0)})$ from the momentum sampler with $\beta = 1$.

For the Gaussian binary HMC sampler, $t_j(k) = \pi$, which is a constant function and is independent of the value of k. For the momentum sampler with $\beta = 1$, it's easy to see that $t_j(k) = 1$, which is also a constant function. This implies that $\forall k$, $\frac{t_j(k)}{\pi}$ for the Gaussian binary HMC sampler is equivalent to $t_j(k)$ for the momentum sampler with $\beta = 1$.

The above equivalences imply that the Gaussian binary HMC has exactly the same behavior as the momentum sampler with $\beta = 1$. In fact, the Gaussian binary HMC sampler behaves like scaling the time of the momentum sampler with $\beta = 1$ by π. □

B.2 Proof of Proposition 2

Proposition 2. *If, in addition to resampling p, we also resample q at every iteration, then the Gaussian binary HMC sampler with travel time $T = \pi$ is equivalent to a systematic scan single-site random walk Metropolis sampler, where the site-visitation order is a random permutation of $1, \ldots, N$ and is refreshed after every N site visits.*

Proof. We've already established that for the Gaussian binary HMC sampler, $t_j^{(0)}$ and $k_j^{(0)}$ are independent, and that $t_j^{(0)} \sim U([0,\pi])$, $k_j^{(0)} \sim \exp(1)$. If we adopt the travel time $T = \pi$, in the Gaussian binary HMC sampler, we would make exactly N site visitations. The order of the site visitations are determined by ranking N independent samples from $U([0,\pi])$, which is equivalent to doing a random permutation of $1, \ldots, N$.

When we come to site j, we are going to look at the energy difference, which we denote as $\log \frac{\pi(x)}{\pi(\tilde{x})}$, where $\tilde{x}_j = -x_j$ and $\tilde{x}_i = x_i, \forall i \neq j$. We would compare the energy difference with $k_j^{(0)} \sim \exp(1)$. The probability for us to cross the boundary and accept the proposal $\tilde{x}$ is

$$\mathbb{P}_{X \sim \exp(1)}\left(X > \log \frac{\pi(x)}{\pi(\tilde{x})}\right) = \min\left\{1, \frac{\pi(\tilde{x})}{\pi(x)}\right\}$$

which is exactly the same as the acceptance probability in regular random walk Metropolis algorithm.

Note that this equivalence between the probability to cross the boundary and the acceptance probability in regular random walk Metropolis algorithm has already been pointed out in Appendix A in [9], but the independence between $t_j^{(0)}$ and $k_j^{(0)}$, which greatly simplifies the analysis of the behaviors of the Gaussian binary HMC sampler, wasn't mentioned in [9].

Since with $T = \pi$, we are only going to visit each site once, sampling all $k_j^{(0)}$'s beforehand and comparing the energy differences with the $k_j^{(0)}$'s to see whether we cross are equivalent to making flip proposals and using the Metropolis acceptance probability to determine whether we accept or not.

As a result, it's easy to see that with travel time $T = \pi$, if we also resample q at each iteration, the Gaussian binary HMC sampler is exactly equivalent to a systematic scan single-site random walk Metropolis algorithm, where the site-visitation order is a random permutation of $1, \ldots, N$ and is refreshed after every N site visitations. □

B.3 Proof of Proposition 3

Proposition 3. *The exponential binary HMC sampler is equivalent to the momentum sampler with deterministic proposal* $Q_i(\tilde{x}|x) = \begin{cases} 1 & \tilde{x}_i = -x_i, \tilde{x}_j = x_j, \forall j \neq i \\ 0 & \text{otherwise} \end{cases}$ *, $i = 1, \ldots, N$ and $\beta = \frac{2}{3}$.*

Proof. Using the expressions we derived in Section B, we can see that, at a given site j,

- For the exponential binary HMC sampler, the joint distribution of the random vector $(t_j^{(0)}, k_j^{(0)})$ is the same as the random vector $\left(p+\sqrt{p^2+2q}, q+\frac{1}{2}p^2\right)$, where $q \sim \exp(1), p \sim N(0,1)$ are independent. For a given total energy level k, $t_j(k) = 2\sqrt{2k}$.
- For the momentum sampler with $\beta = \frac{2}{3}$, the joint distribution of the random vector $(t_j^{(0)}, k_j^{(0)})$ is the same as the random vector $\left(\frac{3}{2}q|p|^{\frac{1}{3}}, |p|^{\frac{2}{3}}\right)$, where $q \sim U([0,1]), p \sim G\left(0,1,\frac{2}{3}\right)$ are independent. For a given total energy level k, $t_j(k) = \frac{3}{2}\sqrt{k}$.

In order to establish the equivalence between these two samplers, we calculate the characteristic functions of two random vectors. We first calculate the characteristic function of the random vector $\left(p+\sqrt{p^2+2q}, q+\frac{1}{2}p^2\right)$, where $q \sim \exp(1), p \sim N(0,1)$ are independent:

$$
\begin{aligned}
& \mathbb{E}_{q\sim\exp(1),p\sim N(0,1)}\left[e^{i\left[t_1\left(p+\sqrt{p^2+2q}\right)+t_2\left(q+\frac{1}{2}p^2\right)\right]}\right] \\
= & \frac{1}{\sqrt{2\pi}}\int_0^{+\infty}\int_{\mathbb{R}} e^{it_1\left(p+\sqrt{p^2+2q}\right)+it_2\left(q+\frac{p^2}{2}\right)}e^{-q}e^{-\frac{p^2}{2}}\mathrm{d}p\mathrm{d}q \\
= & \frac{1}{2\sqrt{2\pi}}\int_{\mathbb{R}^2} e^{it_1\left(p+\sqrt{p^2+2|q|}\right)+it_2\left(|q|+\frac{p^2}{2}\right)}e^{-|q|}e^{-\frac{p^2}{2}}\mathrm{d}p\mathrm{d}q \\
\overset{p=r\cos\theta,q=\mathrm{sign}(\sin\theta)\frac{r^2\sin^2\theta}{2}}{=} & \frac{1}{2\sqrt{2\pi}}\int_0^{+\infty}\int_0^{2\pi} e^{it_1 r(1+\cos\theta)+it_2\frac{r^2}{2}}e^{-\frac{r^2}{2}}r^2\sin\theta\mathrm{d}\theta\mathrm{d}r
\end{aligned}
$$

Next we calculate the characteristic function of the random vector $\left(2\sqrt{2}q|p|^{\frac{1}{3}}, |p|^{\frac{2}{3}}\right)$, where $q \sim U([0,1]), p \sim G\left(0,1,\frac{2}{3}\right)$ are independent:

$$
\begin{aligned}
& \mathbb{E}_{q\sim U([0,1]),p\sim G\left(0,1,\frac{2}{3}\right)}\left[e^{i\left(t_1 2\sqrt{2}q|p|^{\frac{1}{3}}+t_2|p|^{\frac{2}{3}}\right)}\right] \\
= & \frac{\frac{2}{3}}{2\Gamma\left(\frac{3}{2}\right)}\int_0^1\int_{\mathbb{R}} e^{it_1 2\sqrt{2}q|p|^{\frac{1}{3}}+it_2|p|^{\frac{2}{3}}}e^{-|p|^{\frac{2}{3}}}\mathrm{d}p\mathrm{d}q \\
= & \frac{2}{3\sqrt{\pi}}\int_0^1\int_{\mathbb{R}} e^{it_1 2\sqrt{2}q|p|^{\frac{1}{3}}+it_2|p|^{\frac{2}{3}}}e^{-|p|^{\frac{2}{3}}}\mathrm{d}p\mathrm{d}q \\
= & \frac{4}{3\sqrt{\pi}}\int_0^1\int_0^{+\infty} e^{it_1 2\sqrt{2}qp^{\frac{1}{3}}+it_2 p^{\frac{2}{3}}}e^{-p^{\frac{2}{3}}}\mathrm{d}p\mathrm{d}q \\
\overset{q=\frac{1+\cos\theta}{2},p=\frac{r^3}{2^{\frac{3}{2}}}}{=} & \frac{4}{3\sqrt{\pi}}\int_0^{\pi}\int_0^{+\infty} e^{it_1 r(1+\cos\theta)+it_2\frac{r^2}{2}}e^{-\frac{r^2}{2}}\frac{3}{2^{\frac{5}{2}}}r^2\sin\theta\mathrm{d}r\mathrm{d}\theta \\
= & \frac{1}{\sqrt{2\pi}}\int_0^{+\infty}\left[\int_0^{\pi} e^{it_1 r(1+\cos\theta)}\sin\theta\mathrm{d}\theta\right]e^{it_2\frac{r^2}{2}-\frac{r^2}{2}}r^2\mathrm{d}r \\
= & \frac{1}{2\sqrt{2\pi}}\int_0^{+\infty}\left[\int_0^{2\pi} e^{it_1 r(1+\cos\theta)}\sin\theta\mathrm{d}\theta\right]e^{it_2\frac{r^2}{2}-\frac{r^2}{2}}r^2\mathrm{d}r \\
= & \frac{1}{2\sqrt{2\pi}}\int_0^{+\infty}\int_0^{2\pi} e^{it_1 r(1+\cos\theta)+it_2\frac{r^2}{2}}e^{-\frac{r^2}{2}}r^2\sin\theta\mathrm{d}\theta\mathrm{d}r
\end{aligned}
$$

The above calculations indicate that the joint distribution of $(t_j^{(0)}, k_j^{(0)})$ for the exponential binary HMC sampler is equivalent to the joint distribution of $\left(\frac{4\sqrt{2}}{3}t_j^{(0)}, k_j^{(0)}\right)$ for the momentum sampler with $\beta = \frac{2}{3}$. Furthermore, if we multiply the $t_j(k)$ function of the momentum sampler with $\beta = \frac{2}{3}$ by $\frac{4\sqrt{2}}{3}$, we get the function $2\sqrt{2k}$, which is exactly the $t_j(k)$ function for the exponential binary HMC sampler.

The above equivalences imply that the exponential binary HMC has exactly the same behavior as the momentum sampler with $\beta = \frac{2}{3}$. In fact, the exponential binary HMC sampler behaves like scaling the time of the momentum sampler with $\beta = \frac{2}{3}$ by $\frac{3}{4\sqrt{2}}$. □

References

[1] Michael Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv:1701.02434*, July 2018.

[2] Vu Dinh, Arman Bilge, Cheng Zhang, and Frederick A Matsen, IV. Probabilistic path hamiltonian monte carlo. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 1009–1018, Sydney, NSW, Australia, 2017. JMLR.org.

[3] N Ito and G A Kohring. Cluster VS single-spin algorithms– which are more efficient? *International Journal of Modern Physics C*, 05(01):1–14, February 1994.

[4] Ravin Kumar, Colin Carroll, Ari Hartikainen, and Osvaldo Martin. ArviZ a unified library for exploratory analysis of bayesian models in python. *The Journal of Open Source Software*, 4(33):1143, January 2019.

[5] Samuel Livingstone, Michael F Faulkner, and Gareth O Roberts. Kinetic energy choice in hamiltonian/hybrid monte carlo. *arXiv:1706.02649*, November 2018.

[6] Hadi Mohasel Afshar and Justin Domke. Reflection, refraction, and hamiltonian monte carlo. In C Cortes, N D Lawrence, D D Lee, M Sugiyama, and R Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3007–3015. Curran Associates, Inc., 2015.

[7] Radford M. Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 54:113–162, 2010.

[8] Akihiko Nishimura, David Dunson, and Jianfeng Lu. Discontinuous hamiltonian monte carlo for discrete parameters and discontinuous likelihoods. *arXiv:1705.08510*, August 2018.

[9] Ari Pakman and Liam Paninski. Auxiliary-variable exact hamiltonian monte carlo samplers for binary distributions. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2490–2498. Curran Associates, Inc., 2013.

[10] Giacomo Zanella. Informed proposals for local MCMC in discrete spaces. *Journal of the American Statistical Association*, pages 1–27, March 2019.

[11] Yichuan Zhang, Zoubin Ghahramani, Amos J Storkey, and Charles A Sutton. Continuous relaxations for discrete hamiltonian monte carlo. In F Pereira, C J C Burges, L Bottou, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 3194–3202. Curran Associates, Inc., 2012.