Accelerated Markov Chain Monte Carlo Algorithms on Discrete States

Bohan Zhou BHZHOU@UCSB.EDU
Department of Mathematics
University of California
Santa Barbara, CA 93106, USA.

Shu Liu SHULIU@MATH.UCLA.EDU
Department of Mathematics
University of California
Los Angeles, CA 90095, USA.

Xinzhe Zuo ZXZ@MATH.UCLA.EDU
Department of Mathematics
University of California
Los Angeles, CA 90095, USA.

Wuchen Li WUCHEN@MAILBOX.SC.EDU
Department of Mathematics
University of South Carolina
Columbia, SC 29208, USA.

Abstract

We propose a class of discrete state sampling algorithms based on Nesterov's accelerated gradient method, which extends the classical Metropolis-Hastings (MH) algorithm. The evolution of the discrete states probability distribution governed by MH can be interpreted as a gradient descent direction of the Kullback–Leibler (KL) divergence, via a mobility function and a score function. Specifically, this gradient is defined on a probability simplex equipped with a discrete Wasserstein-2 metric with a mobility function. This motivates us to study a momentum-based acceleration framework using damped Hamiltonian flows on the simplex set, whose stationary distribution matches the discrete target distribution. Furthermore, we design an interacting particle system to approximate the proposed accelerated sampling dynamics. The extension of the algorithm with a general choice of potentials and mobilities is also discussed. In particular, we choose the accelerated gradient flow of the relative Fisher information, demonstrating the advantages of the algorithm in estimating discrete score functions without requiring the normalizing constant and keeping positive probabilities. Numerical examples, including sampling on a Gaussian mixture supported on lattices or a distribution on a hypercube, demonstrate the effectiveness of the proposed discrete-state sampling algorithm.

Keywords: Markov chain Monte Carlo methods; Metropolis-Hastings algorithms; Nesterov acceleration methods; Wasserstein metrics on graphs; discrete score functions; mobility functions.

AMS MSC: 65C05, 60J22, 82M31, 49Q22, 65K10

1 Introduction

Markov Chain Monte Carlo (MCMC) are algorithms to generate samples from a given probability distribution over a sampling space. A key advantage of MCMC is that it only requires knowledge of the target distribution up to a constant, making it well-suited for distributions with intractable normalizing constants—such as those arising in high-dimensional sampling spaces or in Bayesian posterior inference problems (Mengersen and Tweedie, 1996). This advantage has led to widespread applications of MCMC algorithm across various scientific domains, including Bayesian inference (Robert and Casella, 2004), computational physics and chemistry (e.g., simulating Ising models; see Metropolis et al. (1953); Hastings (1970); Landau and Binder (2015)), financial mathematics (e.g., risk modeling and option pricing; Jäckel (2002); Kou and Wang (2004); Jacquier et al. (2004)), and machine learning (e.g., energy-based models; Teh et al. (2004); Hinton et al. (2006); Salakhutdinov and Hinton (2009)).

In MCMC methods, the core idea is to construct a Markov chain with designed transition probabilities so that its stationary distribution is the target distribution π. Samples are then generated by running the chain for a sufficient number of steps. A well-known class of MCMC methods is the Langevin Monte Carlo (LMC) algorithm (Roberts and Tweedie, 1996a,b), which is based on the overdamped Langevin dynamics. LMC is designed to sample from a continuous sampling space by conducting gradient descent steps in the direction $\nabla \log \pi$, perturbed by an additive white noise. It is known that the evolution of the probability distribution p under LMC is a gradient flow of the Kullback-Leibler (KL) divergence on the space of probability distributions with the Wasserstein-2 metric (Villani, 2009). Here the Wasserstein-2 gradient of the KL divergence represents the score function $\nabla \log \frac{p}{\pi}$. Classical MCMC algorithms frequently encounter challenges such as slow convergence and poor mixing. This has motivated the demand for accelerated sampling algorithms. In recent decades, the nesterov accelerated gradient (NAG) method (Nesterov, 1983; Su et al., 2016; Wibisono, 2018), a refinement of momentum-based optimization, has been shown to improve convergence speed of gradient descent algorithms. This naturally raises the question of whether a nesterov-type acceleration can be formulated for sampling, leveraging the Wasserstein gradient flow structure. While this question has been investigated for continuous-space sampling from different approaches (Cheng et al., 2018; Taghvaei and Mehta, 2019; Ma et al., 2021; Wang and Li, 2022; Li et al., 2022; Zuo et al., 2024), the crucial sampling problems are defined over discrete spaces, as seen in statistical physics (Landau and Binder, 2015), combinatorics (Jerrum and Sinclair, 1996), and discrete probabilistic graphical models (Jordan et al., 1999). A natural question arises:

How can we design a Nesterov accelerated sampling algorithm on discrete state spaces? What is the accelerated dynamics with discrete score functions?

In this paper, we mainly focus on sampling π on a finite *discrete* space $V = \{1, 2, \dots, n\}$. The Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970) serves as a foundational framework for sampling on discrete space. Given a candidate kernel (q_{ij}), leveraging the acceptance ratio, the MH algorithm designs a *time-homogeneous* Markov chain, with a transition rate matrix Q (Q_{ij} depends on (q_{ij})) that satisfies the *detailed-balance* condition $\pi_j Q_{ji} = \pi_i Q_{ij}$.

	Discrete-time		Continuous-time
MC	$\mathbb{P}(X^{(k+1)} = j \mid X^{(k)} = i) = P_{ij}$		$\mathbb{P}(X(t+h) = j \mid X(t) = i) \approx \delta_{ij} + Q_{ij}h$
	$\downarrow$		
FM	$p^{(k+1)} = p^{(k)}P$ in (4)	$\longleftrightarrow$	$\dot{p}(t) = pQ$ in (3)
			$\downarrow$
GF	$p^{(k+1)} = p^{(k)}(I_n + Q\Delta t)$ in (4)		$\dot{p}(t) = -\nabla_p D_\phi(p\\|\pi)\mathbb{K}(p)$ in (10)
			$\downarrow$
HF	$p^{(k+1)} = p^{(k)}(I_n + \bar{Q}^r_\psi \Delta t)$ in (22), (23)	$\longleftarrow$	(1) is an instance of general framework (18)

Table 1: The conceptual flow in this work. “MC”,“FM”,“GF”,“HF” are abbreviations for Markov chain, forward master equation, gradient flow and Hamiltonian flow, respectively. We start with the Markov chain and end up with the damped Hamiltonian dynamics originated from Nesterov’s accelerated gradient method.

Analogous to the LMC algorithm on the continuous state space, which can be regarded as a gradient flow on the probability space, recent work (Maas, 2011; Erbar and Maas, 2012; Karatzas et al., 2021) has shown that the reversible Markov chain described in the forward master equation can be interpreted as the gradient flow of the KL divergence with respect to the discrete Wasserstein metric, induced from a mobility function θ. This allows us to view the MCMC sampling as a first-order optimization algorithm on the probability simplex.

In this paper, we propose a momentum-based acceleration framework for discrete sampling. Our work is inspired by the interplay between the Wasserstein gradient flow structure of the MH algorithm and the NAG method. The key ingredients of our approach are summarized in Table 1. In particular, we consider a *damped Hamiltonian flow*, driven by the relative Fisher information:

$$\mathrm{I}(p\|\pi) = \frac{1}{4}\sum_{i,j=1}^{n} \pi_i Q_{ij}\theta_{ij}(p)\left(\log\frac{p_i}{\pi_i} - \log\frac{p_j}{\pi_j}\right)^2 .$$

The dynamics evolve according to

$$\begin{cases} \dfrac{\mathrm{d}p_i(t)}{\mathrm{d}t} + \displaystyle\sum_{j\neq i} \pi_i Q_{ij}\theta_{ij}(p(t))(\psi_j(t) - \psi_i(t)) = 0, \\ \dfrac{\mathrm{d}\psi_i(t)}{\mathrm{d}t} + \gamma(t)\psi_i(t) + \dfrac{1}{2}\displaystyle\sum_{j\neq i} \pi_i Q_{ij}\frac{\partial\theta_{ij}(p(t))}{\partial p_i}(\psi_i(t) - \psi_j(t))^2 + \frac{\partial \mathrm{I}(p(t)\|\pi)}{\partial p_i} = 0\,. \end{cases} \tag{1}$$

The first equation constitutes a forward master equation that describes the evolution of the probability function p. The second equation, of Hamilton–Jacobi type, governs the dynamics of the momentum variable ψ, incorporating both dissipative effects through the term $\gamma(t)\psi_i$ and information-theoretic feedback via $\frac{\partial \mathrm{I}(p\|\pi)}{\partial p_i}$. Here, $\gamma(t)$ is a user-specified damping parameter. (θ_{ij}) is a mobility weight matrix that may depend on $p(t)$. A particularly important choice is the *logarithmic mean* or *KL divergence mean*, given by

$$\theta_{ij}(p) = \frac{\frac{p_i}{\pi_i} - \frac{p_j}{\pi_j}}{\log \frac{p_i}{\pi_i} - \log \frac{p_j}{\pi_j}} = \frac{p_i}{\pi_i} \cdot \frac{1 - \frac{\pi_i}{\pi_j}\frac{p_j}{p_i}}{\log\left(\frac{\pi_j}{\pi_i}\frac{p_i}{p_j}\right)}. \tag{2}$$

The proposed accelerated MCMC (**aMCMC**) dynamics in (1) can be shown to guarantee the non-increasing behavior of its associated Hamiltonian, defined as $\mathcal{H}(p,\psi) = \frac{1}{2}\psi\mathbb{K}(p)\psi^\top + \mathrm{I}(p\|\pi)$, and the strict positivity of $p(t)$. For numerical implementation, we employ a staggered scheme to discretize (1) in time. Additionally, a particle swarm jumping process is introduced, with transition rates derived from a positive-negative decomposition of the gradient term $\psi_j - \psi_i$. We refer the reader to Section 5 for a detailed discussion of the algorithmic design.

Numerical experiments presented in Section 6 demonstrate that both the discrete time scheme and the particle swarm sampling-type algorithm achieve faster convergence to the target distribution π in L^2 norm, compared with the MH algorithm. Notably, the incorporation of score functions in the aMCMC algorithm yields a higher empirical accuracy, achieving an error of order $\mathcal{O}(\frac{1}{M})$, where M denotes the number of swarm particles. This marks a substantial improvement over the $\mathcal{O}(\frac{1}{\sqrt{M}})$ accuracy of the MH algorithm.

This paper is organized as follows: Section 2.1 reviews the classical MCMC with the MH update. Section 2.2 revisits the gradient flow formulation of the forward master equation, while Section 2.3 discusses gradient and accelerated flows in Euclidean space $\mathbb{R}^d$. Building on both perspectives, we introduce a class of aMCMC sampling algorithms on Markov chains in Section 3, with particular instances listed in Table 2. The properties of proposed dynamics are discussed in Section 4. Section 5 describes the numerical scheme and some practical techniques including initialization and restart. Section 6 presents our numerical results.

Methods	$\theta(\frac{p_i}{\pi_i}, \frac{p_j}{\pi_j})$	potential $\mathcal{U}(p)$	w/o Z	strict positivity
`Chi-squared`	1	$\frac{1}{2}\sum_{i=1}^n \frac{(p_i-\pi_i)^2}{\pi_i}$	No	No
`KL`	log-mean (2)	$\sum_{i=1}^n p_i \log \frac{p_i}{\pi_i}$.	Yes	No
`log-Fisher`	log-mean (2)	$\frac{1}{4}\sum_{i,j=1}^n \omega_{ij}(\log \frac{\pi_j}{\pi_i}\frac{p_i}{p_j})(\frac{p_i}{\pi_i} - \frac{p_j}{\pi_j})$	Yes	Yes
`con-Fisher`	θ_{ij}	$\frac{1}{4}\sum_{i,j=1}^n \omega_{ij}\theta_{ij}(\log \frac{\pi_j}{\pi_i}\frac{p_i}{p_j})^2$	Yes	Yes

Table 2: Examples of aMCMC dynamics. The 2nd column is the mobility weight matrix $(\theta_{ij}(p))$. The 3rd column is the potential function $\mathcal{U}(p)$. The 4th column is referred to if this method requires to know the normalizing constant Z of the target probability in prior. The 5th column is referred to if this method can ensure p to stay strictly positive.

Notations

The probability simplex $\mathbb{P}(V)$ supported on $V = \{1, 2, \ldots, n\}$ is defined as

$$\mathbb{P}(V) = \left\{ p = [p_1, \ldots, p_n] \in \mathbb{R}^n \;\middle|\; p_i \geqslant 0 \text{ for any } i, \text{ and } \sum_{i=1}^n p_i = 1 \right\}.$$

We adopt the convention that vectors are represented as row vectors. Denote $T_p\mathbb{P}(V)$ as the tangent space of $\mathbb{P}$ at p, then $T_p\mathbb{P}(V) = \{v = [v_1, \ldots, v_n] \in \mathbb{R}^n \mid v_1 + \ldots + v_n = 0\}$. The target probability is denoted by $\pi = [\pi_1, \ldots, \pi_n]$, typically used as the stationary probability of the proposed dynamics governing the state variables $p(t) = [p_1(t), \ldots, p_n(t)]$. For simplicity, we assume that π is strictly positive ($\pi_i > 0$ for all i). The transition probability matrix is denoted by $P = (P_{ij})$, where $P_{ij} = \mathbb{P}(X^{(k+1)} = j \mid X^{(k)} = i)$ represents the probability of transitioning from node i to node j. We adopt the row-sum-one convention, meaning $\sum_{j=1}^n P_{ij} = 1$ for each i. Additionally, we define the associated transition rate matrix as $Q = (Q_{ij})$, where $Q_{ij} \geqslant 0$ for non-diagonal entries, and $\sum_{j=1}^n Q_{ij} = 0$ for each i. Consequently, $Q_{ii} = -\sum_{j \neq i} Q_{ij}$. Let $\bar{A}^r$ denote the row-sum-zero projection of any matrix A such that $\begin{cases} \bar{A}^r_{ij} = A_{ij}, & \text{for } i \neq j; \\ \bar{A}^r_{ii} = -\sum\limits_{j \neq i} A_{ij}. \end{cases}$ Thus for any transition rate matrix Q, $\bar{Q}^r = Q$. We occasionally use the notation $\bar{Q}^r$ in place of Q to highlight that Q is a row-sum-zero matrix. The stationary probability function π can be defined to satisfy that either $\pi P = \pi$ or $\pi Q = 0$.

Given a Q-matrix, it induces a directed weighted graph $G = (V, E, \omega)$ by: 1) $e_{ij} \in E$ if $Q_{ij} \neq 0$; 2) $\omega_{ij} = \pi_i Q_{ij}$. Under this choice, ω is row-sum-zero as well. Furthermore, if Q-matrix is reversible w.r.t its stationary probability π, i.e, $\pi_i Q_{ij} = \pi_j Q_{ji}$, then we say the associated Markov chain is *reversible* and we obtain an undirected graph G and a symmetric matrix ω.

The identity matrix in $\mathbb{R}^{n\times n}$ is denoted by I_n. A row vector of size n (a matrix of size $m \times n$) whose entries are ones is denoted by $\mathbb{1}_n$ ($\mathbb{1}_{m\times n}$, respectively). Similarly, $0_{n\times n}$ denotes the $n \times n$ zero matrix. $\mathbb{R}_+, \mathbb{R}_{\geqslant 0}$ are positive and nonnegative real numbers respectively. The orthogonal complement of the span$\{\mathbb{1}_n\}$ is $\mathbb{1}_n^\perp = \{v \in \mathbb{R}^n \mid \mathbb{1}_n \cdot v = 0\}$. For any square matrix $A \in \mathbb{R}^{n\times n}$, we denote $\text{Ker}(A) = \{x = [x_1, \ldots, x_n] \mid xA = 0\}$.

2 Preliminaries

2.1 Metropolis-Hastings as the Forward Master Equation

Consider a *continuous-time reversible* Markov chain on a finite state space V. Let Q be a transition-rate matrix of the Markov chain, and $p(0) = [p_1(0), \ldots, p_n(0)]$ be a probability function on V. The evolution dynamics of the probability function $p(t)$ with an initial data $p(0)$ is described by the forward master equation:

$$\frac{\mathrm{d}}{\mathrm{d}t} p_i = \sum_{j=1}^n p_j Q_{ji} = \sum_{j \neq i} p_j Q_{ji} - p_i Q_{ij}, \tag{3}$$

whose stationary probability function is denoted by $\pi = [\pi_1, \ldots, \pi_n]$. The forward master equation can be represented in a matrix form as $\frac{\mathrm{d}}{\mathrm{d}t} p = pQ$, by using $Q_{ii} = -\sum_{j \neq i} Q_{ij}$. From the fundamental theorem of Markov chains, there is a unique, strictly positive stationary distribution π if we further assume the Markov chain is ergodic.

One can discretize (3) using the forward difference scheme. Given a step size $\Delta t > 0$, the discrete-time update satisfies

$$p^{(k+1)} = p^{(k)} P, \quad P = I_n + Q\Delta t \in \mathbb{R}^{n\times n}, \tag{4}$$

where I_n is the identity matrix. When Δt is chosen sufficiently small, the matrix P is a valid transition probability matrix so that (4) is a discrete-time jump process.

In a sampling algorithm, Q must be designed to ensure that the probability distribution $p(t)$ converge to the given target distribution π. The Metropolis-Hastings algorithm proposes a generic way to construct a *discrete-time* Markov-chain on V, that is ergodic (i.e., irreducible and aperiodic) and stationary with respect to π. Given a user-specified conditional density $q_{ij} = \mathbb{P}(Y = j \mid X = i)$, also known as the *candidate kernel*, MH designs an acceptance-rejection matrix:

$$A_{ij} := A(X = i, Y = j) = \begin{cases} \min\left\{\dfrac{\pi_j q_{ji}}{\pi_i q_{ij}}, 1\right\}, & \pi_i q_{ij} > 0; \\ 1, & \pi_i q_{ij} = 0. \end{cases}$$

A key advantage is that the acceptance-rejection probability depends on the ratio of $\frac{\pi_j}{\pi_i}$, eliminating the need to know the normalizing constant for π (Ottobre, 2016). In this work we use the standard random walk MH scheme for (q_{ij}) as a benchmark for comparisons. The Metropolis-Hastings algorithm is presented in Algorithm 1.

```
Algorithm 1: Metropolis-Hastings algorithms
  Input  : Initial data $X^{(0)}$, target distribution $\pi$, candidate kernel $q$, total
           iterations $N$.
  Output: The Markov chain $(X^{(i)})_{i=1}^{N}$.
  for $i \leftarrow 0$ to $N$ do
      Propose a candidate $Y^{(i)} \sim q(y \mid X^{(i)})$;
      Generate $u^{(i)} \sim$ uniform$[0, 1]$;
      Set $X^{(i+1)} = Y^{(i)}$ if $u \leqslant A(X^{(i)}, Y^{(i)})$; otherwise $X^{(i+1)} = X^{(i)}$.
  end
```

The acceptance-rejection matrix induces a transition rate matrix for (3), satisfying the detailed balance condition:

$$Q_{ij}^{\mathrm{MH}} := q_{ij} A_{ij} = \min\left\{\frac{\pi_j}{\pi_i} q_{ji}, q_{ij}\right\}. \tag{5}$$

In summary, the MH algorithm yields a reversible and ergodic Markov chain on a discrete state space V. Starting from an arbitrary initial distribution $p(0)$, the time evolution of the probability distribution $p(t)$ under the MH dynamics is governed by (3), alternatively, the jump process $p^{(k)}$ arising from MH can be described by (4), specifically using the transition rate matrix Q^{MH}.

One advantage of Q^{MH} is its time-homogeneity, which simplifies implementation in numerical experiments. Additionally, Q^{MH} is dependent of the normalizing constant Z for π. However, it may suffer from *pseudo-convergence* in multimodal distribution (see Geyer, 2011, Chapter 1.11.2). In practice, MH sampling can proceed either by running a single long chain or by averaging over multiple shorter chains.

Schnakenberg (1976) shows that given any transition rate probability matrix Q and the target distribution π, there is a weighted directed graph representation $G = (V, E, \omega)$ of the forward master equation, where weight matrix ω is given by $(\omega_{ij}) = (\pi_i Q_{ij})$. In the specific case where $Q = Q^{\mathrm{MH}}$, the detailed balance condition implies that $\omega_{ij} = \omega_{ji}$, rendering G an undirected graph that permits self-loops but excludes multiple edges. Our primary focus in this work is sampling on the corresponding undirected graph $G = (V, E, \omega)$ with $Q = Q^{\mathrm{MH}}$.

2.2 Reversible Markov Chain as a Wasserstein Gradient Flow

A series of seminal works (Maas, 2011; Erbar and Maas, 2012; Mielke, 2013; Karatzas et al., 2021) interpret the reversible Markov chain on a discrete state space as the “gradient flow” of the entropy functional over the probability manifold $\mathbb{P}(V)$, by introducing a graphical version of the Wasserstein metric, and exploiting the geodesic convexity of the entropy w.r.t this metric. In this subsection, we provide a brief review with a focus on the Wasserstein gradient flow and postpone the discussion on convexity to Section 4.4 and Appendix C.

We refer interested readers to (Otto, 2001; Ambrosio et al., 2008; Villani, 2009) for Wasserstein gradient flows (WGF) on continuous state spaces; and to (Maas, 2011; Chow et al., 2012) on discrete state spaces. In these studies, the probability manifold is endowed with a Riemannian metric known as the Wasserstein metric. The so-called *Otto calculus* (Otto, 2001) is then used to introduce a gradient flow interpretation on the probability manifold for a family of evolution equations.

We first recall approaches in (Mielke, 2013; Gao et al., 2024) to treat the gradient structure w.r.t the Riemannian metric. In optimization, it is equivalent to apply a preconditioning matrix—commonly known as *Onsager's response* matrix $\mathbb{K}$—onto the flat gradient. Given a convex function $f : \mathbb{R}_{\geqslant 0} \mapsto \mathbb{R}$ that satisfies $f(1) = 0$, $f(0) = \lim_{x \to 0+} f(x)$, and $f''(x) > 0$ for $x > 0$, we define the f-divergence on $\mathbb{P}(V)$ as

$$\mathrm{D}_f(p \| \pi) = \sum_{i=1}^{n} f(\frac{p_i}{\pi_i})\pi_i. \tag{6}$$

By Jensen's inequality, $\mathrm{D}_f(p\|\pi) \geqslant f(1) = 0$. A common choice of f is $f(x) = x \log x$. This yields the Kullback–Leibler divergence $\mathrm{D}_{\mathrm{KL}}(p\|\pi) = \sum_{i=1}^n p_i \log(\frac{p_i}{\pi_i})$.

Due to the detailed balance $\omega_{ij} = \pi_i Q_{ij} = \pi_j Q_{ji} = \omega_{ji}$, we can rewrite the forward master equation (3) as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}p_i &= \sum_{j\neq i} p_j Q_{ji} - p_i Q_{ij} = \sum_{j \neq i} \frac{p_j}{\pi_j}\pi_j Q_{ji} - \frac{p_i}{\pi_i}\pi_i Q_{ij} = \sum_{j\neq i}\omega_{ij}(\frac{p_j}{\pi_j} - \frac{p_i}{\pi_i}) \\ &= \sum_{j \neq i} \omega_{ij}\theta_{ij}(p)(f'(\frac{p_j}{\pi_j}) - f'(\frac{p_i}{\pi_i}))\,,\end{aligned} \tag{7}$$

where we introduce a mobility function $\theta(x, y)$ (also known as an activation function , see Gao et al. (2024)) as

$$\theta(x,y) = \begin{cases} \dfrac{x-y}{f'(x) - f'(y)} & x \neq y; \\ \dfrac{1}{f''(x)} & x = y, \end{cases} \tag{8}$$

and use the abbreviation $\theta_{ij}(p) = \theta(\frac{p_i}{\pi_i}, \frac{p_j}{\pi_j})$.

Now we define the Onsager's response matrix $\mathbb{K}(p)$ as

$$\mathbb{K}_{ij}(p) = -\omega_{ij}\theta_{ij}(p) \qquad \text{and} \qquad \mathbb{K}_{ii}(p) = -\sum_{j\neq i}\mathbb{K}_{ij}(p)\,. \tag{9}$$

$\mathbb{K}(p)$ is row-sum-zero and symmetric. One can verify that $\mathbb{K}(p)$ is positive semi-definite with $\mathrm{Ker}(\mathbb{K}(p)) = \mathrm{span}(\mathbb{1}_n)$. We refer the readers to Appendix A for detailed discussion. In fact, $\mathbb{K}(p)$ is the graph Laplacian of a weighted undirected graph with a weight matrix $(\omega_{ij}\theta_{ij})$.

Since $\partial_{p_j}\mathrm{D}_f(p\|\pi) = f'(\frac{p_j}{\pi_j})$, (7) can be rewritten as $\frac{\mathrm{d}}{\mathrm{d}t}p_i = -\sum_{j=1}^{n}\mathbb{K}_{ij}(p)\partial_{p_j}\mathrm{D}(p\|\pi)$. Denote by $\nabla_p\mathrm{D}(p\|\pi) := [\partial_{p_1}\mathrm{D}(p\|\pi), \dots, \partial_{p_n}\mathrm{D}(p\|\pi)]$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}p = -\nabla_p\mathrm{D}(p\|\pi)\mathbb{K}(p), \qquad \text{or equivalently} \quad \frac{\mathrm{d}}{\mathrm{d}t}p\mathbb{K}^{\dagger}(p) = -\nabla_p\mathrm{D}(p\|\pi), \tag{10}$$

where $\mathbb{K}^{\dagger}(p)$ is the Moore-Penrose inverse of $\mathbb{K}(p)$. Thus (10) is employed with a preconditioning matrix $\mathbb{K}(p)$ on the gradient $\nabla_p\mathrm{D}(p\|\pi)$ from the view of optimization.

Consider a curve $p(t)$ on $\mathbb{P}(V)$ parametrized by a time variable t, we denote its tangent vector by $\dot{p}(t)$. Define the *graphical Wasserstein* metric tensor $g_W(\cdot,\cdot)$ as

$$g_W(\dot{p}(t),\dot{p}(t)) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j\neq i}\omega_{ij}\theta_{ij}(p(t))(\psi_i(t)-\psi_j(t))^2 = \psi(t)\mathbb{K}(p(t))(\psi(t))^{\top}, \tag{11}$$

where $\psi(t) = [\psi_1(t), \dots, \psi_n(t)]$ denotes the momentum (cotangent) vector associated with the tangent vector $\dot{p}(t)$. We have the following graphical continuity equation:

$$\dot{p}_i(t) + \sum_{j\neq i}\omega_{ij}\theta_{ij}(p(t))(\psi_j(t)-\psi_i(t)) = 0, \qquad \text{that is,} \qquad \dot{p} - \psi\mathbb{K}(p) = 0. \tag{12}$$

Thus the inner product (11) is expressed as

$$g_W(\dot{p}(t),\dot{p}(t)) = \dot{p}(t)\mathbb{K}^{\dagger}(p)\dot{p}(t)^{\top}. \tag{13}$$

One can interpret $g_W(\cdot,\cdot)$ as a Riemannian metric on $\mathbb{P}(V)$ with a metric tensor $\mathbb{K}^{\dagger}(p)$

It is worth noting that $\mathrm{Ker}(\mathbb{K}(p)^{\dagger}) = \mathrm{Ker}(\mathbb{K}(p)) = \mathrm{span}(\mathbb{1}_n)$. One can verify that $\mathbb{K}(p)$ is positive definite on $\mathbb{1}_n^{\perp}$, and so is $\mathbb{K}(p)^{\dagger}$. Since the tangent space $T_p\mathbb{P}(V)$ at p corresponds to $\mathbb{1}_n^{\perp}$, $\mathbb{K}(p)^{\dagger}$ is positive definite on $T_p\mathbb{P}(V)$ for all $p \in \mathbb{P}(V)$. Consequently, one can treat $g_W(\cdot,\cdot)$ as a Riemannian metric on $\mathbb{P}(V)$. Such a metric also determines the geodesic and the corresponding distance function on $(\mathbb{P}(V), g_W)$. One can introduce the *kinetic energy* on the graph from the Riemannian metric by

$$\frac{1}{2}g_W(\dot{p}(t),\dot{p}(t)) = \frac{1}{2}\psi(t)\mathbb{K}(p(t))\psi(t)^{\top} = \frac{1}{4}\sum_{i=1}^{n}\sum_{j\neq i}\omega_{ij}\theta_{ij}(p(t))(\psi_i(t)-\psi_j(t))^2.$$

Furthermore, analogous to the Benamou-Brenier theorem (Benamou and Brenier, 2000) in continuous space, the graphical version of the Wasserstein metric between two probabilities p_0 and p_1 on the probability manifold $\mathbb{P}(V)$ is defined as follows:

$$W_2^2(p_0, p_1) = \inf_{p,\psi} \left\{ \frac{1}{2} \int_0^1 \sum_{i=1}^{n} \sum_{j \neq i}^{n} \omega_{ij} \theta_{ij}(p(t))(\psi_i(t) - \psi_j(t))^2 \mathrm{d}t \right\}, \tag{14}$$

where the infimum is taken over all sufficiently regular curves $p : [0,1] \mapsto \mathbb{P}(V)$ and $\psi : [0,1] \mapsto \mathbb{R}^n$ that satisfy the graphical continuity equation (12) with boundary conditions $p(0) = p_0$, $p(1) = p_1$. And the minimizer $p(t)$ is the geodesic from p_0 to p_1 on the Riemannian manifold $(\mathbb{P}(V), g_W)$. Consequently, (10) is the *Wasserstein gradient flow* of $\mathrm{D}_f(\cdot\|\pi)$ on $(\mathbb{P}(V), g_W)$:

$$\frac{\mathrm{d}p}{\mathrm{d}t} = -\mathrm{grad}\mathrm{D}_f(p\|\pi) = -\nabla_p \mathrm{D}_f(p\|\pi)\mathbb{K}(p).$$

The detailed derivation for a detailed derivation of the Wasserstein gradient $\mathrm{grad}\mathrm{D}_f(p\|\pi)$ can be founded in Appendix A.

2.3 Gradient Flow and Accelerated Flow in $\mathbb{R}^n$

In this subsection, we review some basic concepts and results from classical optimization in the Euclidean space $\mathbb{R}^d$. This perspective will help clarify how an accelerated framework can emerge naturally from gradient flows.

Let $U : \mathbb{R}^d \to \mathbb{R}$ be a $\mathcal{C}^1$ convex function. The classical gradient descent method for finding the global minimum of $U(x)$ consists of iteratively moving in the negative gradient direction at a small step size Δt

$$x^{(k+1)} = x^{(k)} - \Delta t \nabla U(x^{(k)}).$$

Convergence of the gradient descent algorithm can be guaranteed when $U(x)$ is L-smooth and $0 < \Delta t < 1/L$. Then one has $U(x^{(k)}) - U(x^*) = \mathcal{O}(k^{-1})$, where $U(x^*)$ is the minimum value. If $U(x)$ is further assumed to be λ-strongly convex, then with the choice of $\Delta t = 1/L$, one has linear convergence: $U(x^{(k)}) - U(x^*) = \mathcal{O}\big((1 - \lambda/L)^k\big)$.

The gradient flow can be viewed as the limit of the gradient descent as $\Delta t \to 0$. We consider instead the continuous trajectory $x(t)$ which follows the first-order ODE:

$$\dot{x}(t) = -\nabla U(x(t))$$

for $t > 0$ with the initial condition $x(0) = x_0$. It is well known that the convergence rate of gradient flow is $\mathcal{O}(t^{-1})$ for convex functions. If f is λ-strongly convex for some $\lambda > 0$, then gradient flow converges at a rate of $\mathcal{O}(e^{-\lambda t})$.

The seminal paper Nesterov (1983) proposed the following iteration scheme to minimize a convex function $U(x)$ with some initial guess $x_0 = y_0$:

$$\begin{cases} x^{(k)} = y^{(k-1)} - \Delta t \nabla U(y^{(k-1)}), \\ y^{(k)} = x^{(k)} + \dfrac{k-1}{k+2}(x^{(k)} - x^{(k-1)}) \,. \end{cases} \tag{15}$$

With the choice of $0 < \Delta t \leqslant 1/L$, this scheme can be shown to converge at a much faster rate than the gradient descent. Specifically, $U(x^{(k)}) - U(x^*) = \mathcal{O}(\frac{1}{\Delta t k^2})$. As a result, Nesterov's method is also referred as the *accelerated gradient* method.

For λ-strongly convex and L-smooth functions, Nesterov's iterations are

$$\begin{cases} x^{(k)} = y^{(k-1)} - \Delta t \nabla U(y^{(k-1)}), \\ y^{(k)} = x^{(k)} + \dfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}(x^{(k)} - x^{(k-1)}) \,. \end{cases} \tag{16}$$

Here $\kappa = \frac{\lambda}{L}$ is the condition number. With the choice of $\Delta t = 1/L$, one can show that $U(x^{(k)}) - U(x^*) = \mathcal{O}\big(e^{-k/\sqrt{\kappa}}\big)$.

In contrast to the gradient flow, the continuous version of the Nesterov acceleration is a second order ODE, first proposed and studied by Su et al. (2016):

$$\ddot{x} + \gamma(t)\dot{x} + \nabla U(x) = 0 \,, \tag{17}$$

with initial condition $x(0) = x_0$, $\dot{x}(0) = 0$. Here $\gamma(t)$ is a user-specified damping parameter that can be time-homogeneous or time-inhomogeneous. When $U(x)$ is convex, the original Nesterov's iterations (15) corresponds to the choice $\gamma(t) = \frac{3}{t}$. When $U(x)$ is further assumed to be λ-strongly convex, (16) corresponds to the choice $\gamma(t) = 2\sqrt{\lambda}$.

An important observation by Maddison et al. (2018) is that the second-order ODE (17) can be formulated as

$$\begin{bmatrix} \dot{x} \\ \dot{v} \end{bmatrix} + \begin{bmatrix} 0 \\ \gamma(t)v \end{bmatrix} - \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} \begin{bmatrix} \nabla_x \mathcal{H}(x,v) \\ \nabla_v \mathcal{H}(x,v) \end{bmatrix} = 0 \,,$$

where x is the state variable and v is the momentum variable. $\mathcal{H}(x,v) \triangleq \|v\|^2/2 + U(x)$ is the Hamiltonian function. This formulation lifts the second-order ODE to a system of ODEs, and shows that the accelerated gradient flow can be viewed as a damped Hamiltonian flow.

Zhang and Sra (2018) proposed the Riemannian version of Nesterov's method for geodesically smooth and strongly convex problems. Borrowing ideas from Su et al. (2016), Alimisis et al. (2020) generalized Nesterov's ODE to a Riemannian manifold $(\mathcal{M}, g)$:

$$\ddot{X} + \gamma(t)\dot{X} + \mathrm{grad}_{\mathcal{M}} U(X) = 0 \,,$$

where $X \in \mathcal{M}$ and $\mathrm{grad}_{\mathcal{M}} U(X)$ is the Riemannian gradient of U at X. In particular, when U is geodesically convex, Alimisis et al. (2020) proved an $\mathcal{O}(t^{-2})$ convergence rate, matching the convergence rate in Euclidean space. The choice of the damping parameter $\gamma(t)$ also depends on the lower bound of the sectional curvature of the manifold $\mathcal{M}$.

3 Nesterov Acceleration on Probability Manifolds

Inspired by previous works on the Euclidean space (Su et al., 2016; Maddison et al., 2018) and on the probability manifold (Wang and Li, 2022; Chen et al., 2025), for a target probability π and an irreducible Q-matrix that is reversible w.r.t π, we consider a graph $G = (V, E, \omega)$ with a weight matrix $(\omega_{ij}) = (\pi_i Q_{ij})$ and any activation function θ in the form of (8). The Onsager's response matrix $\mathbb{K}(p)$ defined in (9) induces a Riemannian metric (11) or (13) on the probability manifold $\mathbb{P}(V)$. Within this framework, we construct a damped Hamiltonian dynamics for a pair of state and momentum variables $(p(t), \psi(t))$,

$$\begin{bmatrix} \dot{p}(t) \\ \dot{\psi}(t) \end{bmatrix} = \begin{bmatrix} 0 \\ -\gamma(t)\psi(t) \end{bmatrix} + \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} \begin{bmatrix} \partial_p \mathcal{H}(p(t), \psi(t)) \\ \partial_\psi \mathcal{H}(p(t), \psi(t)) \end{bmatrix}, \tag{18}$$

where $\gamma(t)$ is a user-specified positive damping parameter. The Hamiltonian $\mathcal{H}(p, \psi)$ is defined as follows

$$\mathcal{H}(p, \psi) = \frac{1}{2}\psi \mathbb{K}(p)\psi^\top + \mathcal{U}(p) = \frac{1}{4}\sum_{i=1}^{n}\sum_{j \neq i} \omega_{ij}\theta_{ij}(p)(\psi_i - \psi_j)^2 + \mathcal{U}(p), \tag{19}$$

where $\mathcal{U} : \mathbb{P}(V) \to \mathbb{R}$ is a function, whose unique critical point is the target probability π. The first term in (19) can be interpreted as the graphical analogue of kinetic energy. The second term can be viewed as the potential energy in the classical Hamiltonian system.

Recall that $\mathbb{K}$ is symmetric and positive semi-definite, $\mathcal{H}(p(t), \psi(t))$ is nonincreasing:

$$\begin{aligned} \frac{\mathrm{d}\mathcal{H}(t)}{\mathrm{d}t} &= \partial_p \mathcal{H} \cdot \frac{\mathrm{d}p(t)}{\mathrm{d}t} + \partial_\psi \mathcal{H} \cdot \frac{\mathrm{d}\psi(t)}{\mathrm{d}t} = \partial_p \mathcal{H} \cdot \partial_\psi \mathcal{H} + \partial_\psi \mathcal{H} \cdot (-\gamma(t)\psi(t) - \partial_p \mathcal{H}) \\ &= -\gamma(t)\partial_\psi \mathcal{H} \cdot \psi(t) = -\gamma(t)(\psi(t)\mathbb{K}(p)) \cdot \psi(t) \leqslant 0. \end{aligned} \tag{20}$$

Given any initial probability function $p(0)$, for any t, $p(t) \in \mathbb{P}(V)$:

$$\dot{p}(t)\mathbb{1}_n^\top = \partial_\psi \mathcal{H}\mathbb{1}_n^\top = \psi(t)\mathbb{K}(p(t))\mathbb{1}_n^\top = 0\,.$$

We expand (18) as

$$\begin{cases} \dfrac{\mathrm{d}p_i}{\mathrm{d}t} + \displaystyle\sum_{j \neq i} \omega_{ij}\theta_{ij}(p)(\psi_j - \psi_i) = 0, & \text{(21a)} \\ \dfrac{\mathrm{d}\psi_i}{\mathrm{d}t} + \gamma(t)\psi_i + \dfrac{1}{2}\displaystyle\sum_{j \neq i} \omega_{ij}\frac{\partial \theta_{ij}(p)}{\partial p_i}(\psi_i - \psi_j)^2 + \frac{\partial \mathcal{U}(p)}{\partial p_i} = 0, & \text{(21b)} \end{cases}$$

where (21a) can be written compactly as $\frac{\mathrm{d}}{\mathrm{d}t}p = \psi\mathbb{K}(p)$ while the second equation is in general nonlinear. For any stationary point (p^*, ψ^*) to (21), from (21a) it is necessary that $\psi_i^* = \psi_j^* = c$ for some constant c. Consequently, $\frac{\partial U(p^*)}{\partial p_i} = 0$ from (21b). Given that π is the unique critical point to $\mathcal{U}(p)$, the state variables $p(t)$ driven by our proposed Hamiltonian system must converge to π.

While the choice of activation function θ can be made independently from the potential $\mathcal{U}$, we typically connect these via a suitable function f. In this paper, we primarily focus on the activation function $\theta_{ij}(p) = \dfrac{\frac{p_i}{\pi_i} - \frac{p_j}{\pi_j}}{f'(\frac{p_i}{\pi_i}) - f'(\frac{p_j}{\pi_j})}$ motivated by (Maas, 2011), and the corresponding potential $\mathcal{U}(p)$ either as the f-divergence, defined by $\mathrm{D}_f(p\|\pi) = \sum_{i=1}^n f(\frac{p_i}{\pi_i})\pi_i$ in (6), or the relative Fisher information, given by $\mathrm{I}(p\|\pi) = \frac{1}{4}\sum_{i,j=1}^n \omega_{ij}\theta_{ij}(p)(f'(\frac{p_i}{\pi_i}) - f'(\frac{p_j}{\pi_j}))^2$. For the former potential, from the perspectives outlined in Section 2.2 and Section 2.3, the corresponding forward master equation (3) can be interpreted as the gradient flow of the

f-divergence in the probability manifold equipped with the graphical Wasserstein metric (14). Consequently the damped Hamiltonian dynamics (18) driven with f-divergence as its potential, can be regarded as the Nesterov accelerated flow in the probability manifold.

To construct a sampling algorithm (a "jump process" of particles on V), we rewrite the equation of the state variables p in the form of the forward master equation (3), while leaving the momentum variables ψ as a parameter in the transition rate matrix Q. By splitting the positive part $(\psi_i - \psi_j)_+ := \max\{\psi_i - \psi_j, 0\}$ and the negative part $(\psi_i - \psi_j)_- := -\min\{\psi_i - \psi_j, 0\}$, recall $(\omega_{ij}), (\theta_{ij}(p))$ are symmetric, we have

$$\begin{aligned}
\frac{\mathrm{d}p_i}{\mathrm{d}t} &= \sum_{j\neq i} \omega_{ij}\theta_{ij}(p)(\psi_i - \psi_j) = \sum_{j\neq i} \omega_{ij}\theta_{ij}(p)\left[(\psi_i - \psi_j)_+ - (\psi_i - \psi_j)_-\right] \\
&= -\sum_{j\neq i} \omega_{ij}\theta_{ij}(p)(\psi_i - \psi_j)_- + \sum_{j\neq i} \omega_{ji}\theta_{ji}(p)(\psi_i - \psi_j)_+ \\
&= -\left[\sum_{j\neq i} \frac{\omega_{ij}\theta_{ij}(p)(\psi_i - \psi_j)_-}{p_i}\right] p_i + \sum_{j\neq i} \frac{\omega_{ji}\theta_{ji}(p)(\psi_i - \psi_j)_+}{p_j} p_j \\
&= p\left[\frac{\omega_{1i}\theta_{1i}(p)(\psi_i - \psi_1)_+}{p_1} \quad \cdots \quad -\left[\sum_{j\neq i} \frac{\omega_{ij}\theta_{ij}(p)(\psi_i - \psi_j)_-}{p_i}\right] \quad \cdots \quad \frac{\omega_{ni}\theta_{ni}(p)(\psi_i - \psi_n)_+}{p_n}\right]^\top .
\end{aligned}$$

Thus (21a) is in the form of the forward master equation

$$\frac{\mathrm{d}}{dt} p = p\bar{Q}^r_\psi, \tag{22}$$

where the notation $\bar{Q}^r_\psi$ is to resemble the Q-matrix used in Section 2.1, given by

$$\begin{bmatrix}
-\sum_{j\neq 1} \frac{\omega_{1j}\theta_{1j}(p)(\psi_1 - \psi_j)_-}{p_1} & \frac{\omega_{12}\theta_{12}(p)(\psi_2 - \psi_1)_+}{p_1} & \cdots & \frac{\omega_{1n}\theta_{1n}(p)(\psi_n - \psi_1)_+}{p_1} \\
\frac{\omega_{21}\theta_{21}(p)(\psi_1 - \psi_2)_+}{p_2} & -\sum_{j\neq 2} \frac{\omega_{2j}\theta_{2j}(p)(\psi_2 - \psi_j)_-}{p_2} & \cdots & \frac{\omega_{2n}\theta_{2n}(p)(\psi_n - \psi_2)_+}{p_2} \\
\vdots & \cdots & \ddots & \vdots \\
\frac{\omega_{n1}\theta_{n1}(p)(\psi_1 - \psi_n)_+}{p_n} & \frac{\omega_{n2}\theta_{n2}(p)(\psi_2 - \psi_n)_+}{p_n} & \cdots & -\sum_{j\neq n} \frac{\omega_{nj}\theta_{nj}(p)(\psi_n - \psi_j)_-}{p_n}.
\end{bmatrix} \tag{23}$$

Here, $\bar{Q}^r_\psi$ is row-sum-zero, and it depends on both $p(t)$ and $\psi(t)$, thus no longer time-homogeneous. Additionally, if $p_i(t) = 0$, then the i-th row is undefined. In order to use $P = I_n + \bar{Q}^r_\psi \Delta t$ as the transition probability matrix in the jump process, it is crucial to ensure the strict positivity of $p_i(t)$ for any i and any $t > 0$.

3.1 χ^2 Divergence and Uniform Activiation

First, we choose $f(x) = \frac{1}{2}|x-1|^2$ in (8) and (6). Consequently, $\theta_{ij} = 1$ and $\mathcal{U}(p) = \mathrm{D}_f(p\|\pi) = \frac{1}{2}\sum_{i=1}^n \frac{(p_i - \pi_i)^2}{\pi_i}$. This function is called the χ^2 divergence. It is clear that

$\mathcal{U}(p) \geqslant 0$ and $\mathcal{U}(\pi) = 0$. Now (21) reads

$$\begin{cases} \dfrac{\mathrm{d}p_i}{\mathrm{d}t} = \psi_i \sum\limits_{j\neq i} \omega_{ij} - \sum\limits_{j\neq i} \psi_j \omega_{ji} \\ \dfrac{\mathrm{d}\psi_i}{\mathrm{d}t} + \gamma(t)\psi_i + \dfrac{p_i}{\pi_i} - 1 = 0. \end{cases} \tag{24}$$

Denote by $[p, \psi] = [p_1, \ldots, p_n, \psi_1, \ldots, \psi_n]$, we rewrite (24) into a matrix equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}[p, \psi] = [p, \psi]L + [0_{1\times n}, \mathbb{1}_{1\times n}],$$

where the matrix L is given by

$$L \triangleq \begin{bmatrix} 0_{n\times n} & -\mathrm{diag}(\frac{1}{\pi}) \\ K & -\gamma(t)I_n \end{bmatrix}, \tag{25}$$

where $\mathrm{diag}(\frac{1}{\pi})$ represents a diagonal matrix with diagonal entries $(\frac{1}{\pi_1}, \ldots, \frac{1}{\pi_n})$. Choosing $\theta_{ij} = 1$ results in a constant matrix K, which corresponds to a flat metric of $\mathbb{P}(V)$. We will refer this method as the `Chi-squared` method.

3.2 KL Divergence and Logarithmic Mean

We pick $f(x) = x \log x$ in (8) and (6), which results in $\theta_{ij}(p) = \frac{\frac{p_j}{\pi_j} - \frac{p_i}{\pi_i}}{\log \frac{\pi_i p_j}{\pi_j p_i}}$ and $\mathcal{U}(p) = \mathrm{D}_f(p\|\pi) = \sum_{i=1}^n p_i \log \frac{p_i}{\pi_i}$. This functional is known as the *Kullback-Leibler* (KL) divergence. It is clear that $\mathcal{U}(p) \geqslant 0$ and $\mathcal{U}(p) = 0$ if and only if $p = \pi$. We can compute

$$\frac{\partial \theta_{ij}(p)}{\partial p_i} = \frac{\frac{\log \frac{p_i}{\pi_i} - \log \frac{p_j}{\pi_j}}{\pi_i} - (\frac{p_i}{\pi_i} - \frac{p_j}{\pi_j})\frac{\pi_i}{p_i}\frac{1}{\pi_i}}{(\log \frac{p_i}{\pi_i} - \log \frac{p_j}{\pi_j})^2} = \frac{\log\left(\frac{\pi_j p_i}{\pi_i p_j}\right) - 1 + \left(\frac{\pi_j p_i}{\pi_i p_j}\right)^{-1}}{\log^2\left(\frac{\pi_j p_i}{\pi_i p_j}\right)} \frac{1}{\pi_i}.$$

Thus, (21) becomes

$$\begin{cases} \dfrac{\mathrm{d}p_i}{\mathrm{d}t} = \sum\limits_{j\neq i} \pi_i Q_{ij} \dfrac{\frac{p_i}{\pi_i} - \frac{p_j}{\pi_j}}{\log \frac{\pi_j p_i}{\pi_i p_j}} (\psi_i - \psi_j), \\ \dfrac{\mathrm{d}\psi_i}{\mathrm{d}t} + \gamma(t)\psi_i + \log \dfrac{p_i}{\pi_i} + \dfrac{1}{2} \sum\limits_{j\neq i} Q_{ij} \dfrac{\log\left(\frac{\pi_j p_i}{\pi_i p_j}\right) - 1 + \left(\frac{\pi_j p_i}{\pi_i p_j}\right)^{-1}}{\log^2\left(\frac{\pi_j p_i}{\pi_i p_j}\right)} (\psi_i - \psi_j)^2 = 0. \end{cases}$$

We will refer this method as the `KL` method in subsequent sections. The choice of θ and $\mathcal{U}(p)$ in the `KL` method is natural, since it can be regarded as the Nesterov accelerated flow for the KL divergence in the probability manifold $(\mathbb{P}(V), g_W)$. However, the positivity of density $p(t)$ might not be preserved under the functional $\mathcal{U}(p) = \sum p_i \log \frac{p_i}{\pi_i}$. One has to select a careful numerical scheme in order to do the jump process proposed in the beginning of this section. Therefore we propose the relative Fisher information in the following subsections to tackle this issue.

3.3 The Relative Fisher Information and Logarithmic Mean

We still select $f(x) = x\log x$ for $\theta(x,y) = \frac{x-y}{f'(x)-f'(y)}$ just as in Section 3.2 and choose the relative Fisher information $\mathrm{I}(p\|\pi)$ as the potential energy $\mathcal{U}(p)$, given by

$$\mathrm{I}(p\|\pi) = \frac{1}{4}\sum_{i,j=1}^{n}\omega_{ij}\theta_{ij}(p)(f'(\frac{p_i}{\pi_i}) - f'(\frac{p_j}{\pi_j}))^2 = \frac{1}{4}\sum_{i=1}^{n}\sum_{j\neq i}\omega_{ij}(\log\frac{p_i}{\pi_i} - \log\frac{p_j}{\pi_j})(\frac{p_i}{\pi_i} - \frac{p_j}{\pi_j}).$$

Notably, $\mathrm{I}(p\|\pi) \geqslant 0$, with equality if and only if $\frac{p_i}{\pi_i} = \frac{1}{Z}$ for all i. The following lemma shows that the relative Fisher information also has a unique critical point.

Lemma 1 *Given a strictly positive target probability π and an irreducible transition-rate matrix Q, then $p > 0$ is a critical point of $\mathrm{I}(p\|\pi)$ in $\mathbb{P}(V)$, i.e., $\nabla_p \mathrm{I}(p\|\pi) = 0$, if and only if $p = \pi$.*

Note that $\frac{\partial}{\partial p_i}\mathrm{I}(p\|\pi_i) = \frac{1}{2\pi_i}\sum_{j\neq i}\omega_{ij}[\log\left(\frac{\pi_j p_i}{\pi_i p_j}\right) + 1 - \left(\frac{\pi_j p_i}{\pi_i p_j}\right)^{-1}]$. (21) takes the form:

$$\begin{cases} \dfrac{\mathrm{d}p_i}{\mathrm{d}t} = \displaystyle\sum_{j\neq i}\pi_i Q_{ij}\frac{\frac{p_j}{\pi_j} - \frac{p_i}{\pi_i}}{\log\frac{\pi_i p_j}{\pi_j p_i}}(\psi_i - \psi_j), \\ \dfrac{\mathrm{d}\psi_i}{\mathrm{d}t} + \gamma(t)\psi_i + \dfrac{1}{2}\displaystyle\sum_{j\neq i}Q_{ij}[\log\left(\frac{\pi_j p_i}{\pi_i p_j}\right) + 1 - \left(\frac{\pi_j p_i}{\pi_i p_j}\right)^{-1}] \\ + \dfrac{1}{2}\displaystyle\sum_{j\neq i}Q_{ij}\frac{\log\left(\frac{\pi_j p_i}{\pi_i p_j}\right) - 1 + \left(\frac{\pi_j p_i}{\pi_i p_j}\right)^{-1}}{\log^2\left(\frac{\pi_j p_i}{\pi_i p_j}\right)}(\psi_i - \psi_j)^2 = 0. \end{cases}$$

We will refer to this approach as the `log-Fisher` method. One advantage of this method is that it preserves the positivity of $p(t)$ along the damped Hamiltonian flow, whereas the `KL` method does not.

3.4 The Relative Fisher Information and Constant Activation

Now we consider a constant mobility weight matrix (θ_{ij}) that does not depend on $p(t)$. Selecting $\mathcal{U}(p) = \frac{1}{4}\sum_{i,j=1}^{n}\omega_{ij}\theta_{ij}(\log\frac{p_i}{\pi_i} - \log\frac{p_j}{\pi_j})^2$, (21) reads

$$\begin{cases} \dfrac{\mathrm{d}p_i}{\mathrm{d}t} = \displaystyle\sum_{j\neq i}\pi_i Q_{ij}\theta_{ij}(\psi_i - \psi_j), \\ \dfrac{\mathrm{d}\psi_i}{\mathrm{d}t} + \gamma(t)\psi_i + \dfrac{\pi_i}{p_i}\displaystyle\sum_{j\neq i}Q_{ij}\theta_{ij}\log\left(\frac{\pi_j p_i}{\pi_i p_j}\right) = 0. \end{cases}$$

We refer this as the `con-Fisher` method. This potential $\mathcal{U}(p)$ inherits the advantages discussed in the Section 3.3 such as the unique critical point and preserving strict positivity, while simplifying ψ-equation (21b). Additionally, the Hessian of $\mathcal{U}(p)$ takes a simpler form compared to the one in `log-Fisher`. A detailed analysis will be presented in Section 4.4.

4 Properties of aMCMC

In this section we discuss key properties of aMCMC and compare advantages and disadvantages across different methods.

4.1 Normalizing Constant

To design a damped Hamiltonian system (18) that does not require the normalizing constant Z, it is necessary to verify the following conditions:

- The Onsager's response matrix $\mathbb{K}(p)$ defined in (9) is independent of Z.
- π is the unique critical point to $\mathcal{U}(p)$ in $\mathbb{P}(V)$.
- The damping parameter $\gamma(t)$ does not rely on the knowledge of Z.

In `Chi-Squared` and `con-Fisher` methods, the associated Onsager response matrix $K_{ij} = -\omega_{ij}\theta_{ij}$ is constant, where parameters are commonly selected as $\omega_{ij} = \pi_i Q^{\text{MH}}_{ij}$ and $\theta_{ij} = 1$. Though Q^{MH} itself is independent of Z, the resulting K depends on Z. In contrast, `KL` and `log-Fisher` methods take the Onsager response matrix in the form of $K_{ij}(p) = -\pi_i Q^{\text{MH}}_{ij} \frac{\frac{p_i}{\pi_i} - \frac{p_j}{\pi_j}}{\log \frac{\pi_j}{\pi_i}\frac{p_i}{p_j}} = -Q^{\text{MH}}_{ij} \frac{p_i - p_j \frac{\pi_i}{\pi_j}}{\log \frac{\pi_j}{\pi_i}\frac{p_i}{p_j}}$, which does not depend on Z.

4.2 Positivity

The χ^2 divergence $\frac{1}{2}\sum_{i=1}^n \frac{(p_i-\pi_i)^2}{\pi_i}$ and the KL divergence $\sum_{i=1}^n p_i \log \frac{p_i}{\pi_i}$ remain well-defined even if some state $p_i = 0$. However, neither divergence explicitly prevents $p_i(t)$ from approaching zero during evolution. In the following theorem, we propose a general condition on the function $\mathcal{U}(p)$ to ensure $p_i(t)$ away from zero. The potentials in `log-Fisher` and `con-Fisher` are two examples.

Theorem 2 *Given a strictly positive target probability π and an irreducible transition-rate matrix Q, let $(p(t), \psi(t))$ be the solution to the damped Hamiltonian dynamics* (21) *on a graph $G = (V, E, \omega)$, with the potential function $\mathcal{U}(p)$ in the form of*

$$\mathcal{U}(p) = \frac{1}{4}\sum_{i,j=1}^{n} F_{\frac{p_i}{\pi_i}}(\frac{p_j}{\pi_j})\omega_{ij},$$

where $F_{r_0}(r) : \mathbb{R}_+ \mapsto \mathbb{R}$ is a family of functions with a parameter $r_0 \in \mathbb{R}_+$, satisfying that

1) $F_{r_0}(r) \geqslant 0$ with the equality if and only if $r = r_0$.

2) $\lim_{r\to 0+} F_{r_0}(r) = \infty$.

3) $F_{r_0}(r)$ is strictly decreasing on $(0, r_0]$.

4) Fix $r \in (0, r_0]$, if $r_1 > r_0$, then $F_{r_1}(r) > F_{r_0}(r)$.

Suppose $\mathcal{H}(p(0), \psi(0))$ is bounded and $p(0) > 0$, then there exists a positive constant $\varepsilon > 0$, such that for any t and for any i, $p_i(t) > \varepsilon$.

Example 1 *In Section 3.3,* $\mathcal{U}(p) = \frac{1}{4}\sum_{i,j=1}^{n}(\log\frac{p_i}{\pi_i} - \log\frac{p_j}{\pi_j})(\frac{p_i}{\pi_i} - \frac{p_j}{\pi_j})\omega_{ij}$ *corresponds to* $F_{r_i}(r_j) = (\log r_i - \log r_j)(r_i - r_j)$ *and* $r_i = \frac{p_i}{\pi_i}$.

Example 2 *In Section 3.4,* $\mathcal{U}(p) = \frac{1}{4}\sum_{i,j=1}^{n}(\log\frac{p_i}{\pi_i} - \log\frac{p_j}{\pi_j})^2\theta_{ij}\omega_{ij}$ *chooses* $F_{r_i}(r_j) = (\log r_i - \log r_j)^2\theta_{ij}$ *for some positive constant matrix* (θ_{ij}) *and* $r_i = \frac{p_i}{\pi_i}$.

One can verify conditions in Theorem 2 are satisfied in both examples. The proof can be found in Appendix D.

4.3 Damping Parameter and Convergence Analysis

With the Onsager's response matrix $\mathbb{K}(p)$ in (9), our proposed dynamics (21) can be expressed as

$$\begin{cases} \dot{p}(t) = \psi(t)\mathbb{K}(p(t))\,, \\ \dot{\psi}(t) = -\dfrac{1}{2}\nabla_p\langle\psi(t), \psi(t)\mathbb{K}(p(t))\rangle - \gamma(t)\psi(t) - \nabla_p\mathcal{U}(p(t))\,. \end{cases}$$

Recall the Hamiltonian is $\mathcal{H}(p,\psi) = \frac{1}{2}\psi\mathbb{K}(p)\psi^\top + \mathcal{U}(p)$.

In this subsection, we establish the exponential convergence of the state variables for an explicit chosen damping parameter, by employing the Lyapunov analysis. Our derivation relies on the assumptions that the Onsager's response matrix is constant (i.e, independent of $p(t)$), and the potential function $\mathcal{U}(p)$ is geodescially λ-strongly convex. Let the Onsager's response matrix $\mathbb{K}(p) = K$ for some positive semi-definite K. The corresponding Riemannian metric defined in (11) or (13) is flat. A function $\mathcal{U}$ on $\mathbb{P}(V)$ equipped with such a flat metric is geodesically λ-strongly convex for some constant $\lambda > 0$, if

$$\mathcal{U}(\hat{p}) \geqslant \mathcal{U}(p) + (\hat{p} - p)\nabla_p\mathcal{U}(p)^\top + \frac{\lambda}{2}\|\hat{p} - p\|_{K^\dagger}^2. \tag{26}$$

Now, we define a Lyapunov function by

$$\mathcal{L}(t) = \frac{1}{2}e^{\sqrt{\lambda}t}\|\sqrt{\lambda}(p(t) - \pi) + \psi K\|_{K^\dagger}^2 + e^{\sqrt{\lambda}t}(\mathcal{U}(p(t)) - \mathcal{U}(\pi))\,. \tag{27}$$

Theorem 3 *Consider the probability manifold* $\mathbb{P}(V)$ *on a graph* $G = (V, E, \omega)$ *equipped with a flat metric* (11)*or* (13)*, induced from a constant Onsager's response matrix* (9)*. Given a function* $\mathcal{U}$ *that is geodesically* λ*-strongly convex for* $\lambda > 0$*, select a time-homogeneous damping parameter* $\gamma(t) = 2\sqrt{\lambda}$*, and let* $(p(t), \psi(t))$ *be the solution to the proposed damped Hamiltonian dynamics* (18)*. Then the Lyapunov function* (27) *is non-increasing. Furthermore,* $\mathcal{U}(p(t)) - \mathcal{U}(\pi) \leqslant \mathcal{O}(e^{-\sqrt{\lambda}t})$.

Proof

$$
\begin{aligned}
e^{-\sqrt{\lambda}t}\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}t} &= \frac{1}{2}\sqrt{\lambda}\left[\lambda\|p-\pi\|_{K^\dagger}^2 + \|\psi\|_K^2 + 2\sqrt{\lambda}\psi(p-\pi)^\top\right] + \sqrt{\lambda}(\mathcal{U}(p)-\mathcal{U}(\pi)) \\
&\quad + \nabla_p\psi K\mathcal{U}(p)^\top - (\sqrt{\lambda}\psi K + \nabla_p\mathcal{U}(p)K)K^\dagger(\sqrt{\lambda}(p-\pi)+\psi K) \\
&= \frac{1}{2}\lambda^{\frac{3}{2}}\|p-\pi\|_{K^\dagger}^2 + \frac{\sqrt{\lambda}}{2}\|\psi\|_K^2 + \lambda\psi(p-\pi)^\top + \sqrt{\lambda}(\mathcal{U}(p)-\mathcal{U}(\pi)) + \psi K\nabla_p\mathcal{U}(p)^\top \\
&\quad - \lambda\psi(p-\pi)^\top - \sqrt{\lambda}\|\psi\|_K^2 - \sqrt{\lambda}(p-\pi)\nabla_p\mathcal{U}(p)^\top - \psi K\nabla_p\mathcal{U}(p)^\top \\
&= -\frac{1}{2}\sqrt{\lambda}\|\psi\|_K^2 + \sqrt{\lambda}\left[\mathcal{U}(p)-\mathcal{U}(\pi) + (\pi-p)\nabla_p\mathcal{U}(p)^\top + \frac{\lambda}{2}\|\pi-p\|_{K^\dagger}^2\right]. \qquad (28)
\end{aligned}
$$

Observe that the first term in (28) is non-positive. The second term in (28) is also non-positive by the geodesically λ-strongly convex assumption on $\mathcal{U}$. Thus, we conclude that $\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}t} \leqslant 0$, which implies

$$
\mathcal{U}(p(t)) - \mathcal{U}(\pi) \leqslant e^{-\sqrt{\lambda}t}\mathcal{L}(t) \leqslant e^{-\sqrt{\lambda}t}\mathcal{L}(0) = \mathcal{O}(e^{-\sqrt{\lambda}t}).
$$

■

4.4 Geodesic Convexity, Hessian and Eigenvalues

Previous studies design damping parameter $\gamma(t)$ based on the minimum eigenvalue of the Hessian matrix of the objective function (Nesterov, 1983, 2004) (in the Euclidean space) and sectional curvature (Alimisis et al., 2020) (in the Riemannian manifold). When the Onsager's response matrix $\mathbb{K}(p)$ depends on p, analyzing the sectional curvature within the probability manifold $\mathbb{P}(V)$ equipped with the graphical Wasserstein metric tensor g_W becomes important (Li, 2025). We study the optimal $\gamma(t)$ in the flat-metric cases when $\mathbb{K}(p)$ is constant. Appendix C reviews the notions of geodesic λ-strong convexity. Thus it suffices to verify condition (26) or to establish that $\mathrm{D}^2\mathcal{U} \geqslant \lambda K^\dagger$ for $\lambda > 0$. Once such a positive λ is obtained, the damping parameter can be chosen as $\gamma(t) = 2\sqrt{\lambda}$, ensuring the convergence result by Theorem 3.

4.4.1 Chi-squared method

To check if $\mathcal{U}(p) = \frac{1}{2}\sum_{i=1}^n \frac{(p_i-\pi_i)^2}{\pi_i}$ is geodesically λ-strongly convex, we need to find a positive constant $\lambda > 0$ such that

$$
\begin{aligned}
\mathcal{U}(\hat{p}) - \mathcal{U}(p) - (\hat{p}-p)(\nabla_p\mathcal{U}(p))^\top &= \sum_{i=1}^n(\hat{p}_i-p_i)\frac{\hat{p}_i+p_i-2\pi_i}{2\pi_i} - \sum_{i=1}^n(\hat{p}_i-p_i)\frac{p_i-\pi_i}{\pi_i} \\
=\sum_{i=1}^n(\hat{p}_i-p_i)\frac{\hat{p}_i-p_i}{2\pi_i} = (\hat{p}-p)\mathrm{diag}(\frac{1}{2\pi})(\hat{p}-p)^\top &\geqslant \frac{\lambda}{2}(\hat{p}-p)K^\dagger(\hat{p}-p)^\top,
\end{aligned}
$$

where $K = -\omega = -\mathrm{diag}(\pi)Q$. A sufficient condition is that $0 < \lambda \leqslant \frac{\min\frac{1}{\pi_i}}{\lambda_{\max}((-\omega)^\dagger)}$. Given that π is assumed to be strictly positive and ω is the weight matrix associated to a reversible and irreducible Markov chain, we know λ exists.

Given a transition rate matrix Q, for instance Q^{MH} in MH, we aim to compare the convergence behavior of the classical MCMC method and its accelerated counterpart driven by the χ^2 divergence. This comparison is facilitated via spectral analysis. In particular, when a constant damping parameter is applied, the matrix L defined in (25) is a constant matrix, whose eigenvalues can be computed directly. We derive an explicit relationship between the spectrum of L and those of Q in Theorem 4, which enables a quantitative comparison of their respective convergence rates in terms of the spectral gap in Theorem 5. Background material on the spectral properties of general irreducible and reversible transition rate matrices is provided in Appendix B.

Lemma 4 *Let $\gamma(t) = d$ for some $d > 0$.*

- *If μ is a real eigenvalue of L, then $\alpha = \mu(d + \mu)$ is a real eigenvalue of Q.*
- *If α is a real eigenvalue of Q, then there exist (possibly complex) eigenvalues μ of L such that $\mu(d + \mu) = \alpha$.*

Proof Let A, B, C, D be square matrices, we have

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(AD - BC).$$

Take $A = 0_{n\times n}$, $B = -\mathrm{diag}(\frac{1}{\pi})$, $C = -\mathrm{diag}(\pi)Q$, and $D = -dI_n$, then

$$\det(L - \mu I_{2n}) = \det((A - \mu I_n)(D - \mu I_n) - BC) = \det(\mu(d + \mu)I_n - Q). \tag{29}$$

Obviously, if μ is a real eigenvalue of L, then $\alpha = \mu(d + \mu)$ is a real eigenvalue of Q. On the other hand, if α is a real eigenvalue of Q, there exists (possibly complex) μ such that $\mu(d + \mu) = \alpha$, and

$$0 = \det(\alpha I_n - Q) = \det(\mu(d + \mu)I_n - Q) = \det(L - \mu I_{2n}).$$

■

In particular, 0 is an eigenvalue of both Q and L.

Theorem 5 *Let α_* be the largest negative eigenvalue (i.e., the spectral gap) of Q,*

$$\alpha_* = \max\{\alpha < 0 \mid (Q - \alpha I) \text{ is not injective }\}.$$

If $|\alpha_*| < 1$, then there exists damping parameter $\gamma(t) = d \in [2\sqrt{|\alpha_*|}, |\alpha_*| + 1)$, such that the largest negative eigenvalue μ_* of L satisfies $\mu_* < \alpha_*$.

Proof Let α be an eigenvalue of Q, then from (29) there exists $\mu_1(\alpha) = \frac{-d+\sqrt{d^2+4\alpha}}{2}$, $\mu_2(\alpha) = \frac{-d-\sqrt{d^2+4\alpha}}{2}$ which are (possibly complex) eigenvalues of L. In particular, $\mu_1(0) = 0$ and $\mu_2(0) = -d$. Let $\alpha_{n-1} \leqslant \alpha_{n-2} \leqslant \cdots \leqslant \alpha_1 < 0 = \alpha_0$ denote the ordered eigenvalues of Q. Then $\alpha_* = \alpha_1$. Fix $\alpha \in [\alpha_{n-1}, \alpha_1]$. If $d \geqslant 2\sqrt{|\alpha|}$, then $\mu_2(\alpha) \leqslant -\frac{d}{2} \leqslant \mu_1(\alpha) < 0$. If $d < 2\sqrt{|\alpha|}$, $\mathrm{Re}(\mu_1(\alpha)) = \mathrm{Re}(\mu_2(\alpha)) = \frac{-d}{2}$. Since $|\alpha_*| < 1$, we can pick $d \in [2\sqrt{|\alpha_*|}, |\alpha_*|+1)$. We consider two cases:

- If there exists some k such that $d \in [2\sqrt{|\alpha_k|}, 2\sqrt{|\alpha_{k+1}|})$, then

$$\mu_1(\alpha_1) = \max\{\mu_1(\alpha_1), \mu_2(\alpha_1), \cdots, \mu_1(\alpha_k), \mu_2(\alpha_k)\} \geqslant -\frac{d}{2},$$

  while $\mathrm{Re}(\mu_1(\alpha_j)) = \mathrm{Re}(\mu_2(\alpha_j)) = -\frac{d}{2}$ for $j \geqslant k+1$. Thus $\mu_* = \mu_1(\alpha_*) = \frac{-d+\sqrt{d^2+4\alpha_*}}{2}$.

- If $2\sqrt{|\alpha_{n-1}|} \leqslant d < |\alpha_*| + 1$, then $\mu_* = \mu_1(\alpha_*) = \frac{-d+\sqrt{d^2+4\alpha_*}}{2}$.

In either case, note that $\mu_* = \frac{-d+\sqrt{d^2+4\alpha_*}}{2}$ is increasing w.r.t d. Thus $d < |\alpha_*| + 1$ implies that $\mu_* < \alpha_*$. ■

Remark 6 *The optimal choice is in the form of $d = 2\sqrt{|\alpha_*|}$, which resembles the choice of the damping parameter in the classical Nesterov's acceleration* (17)*, and also our choice of $\gamma(t)$ in Theorem 3. It is worth noting that α_* refers to the largest negative eigenvalue of Q in* (3)*, while λ in the Nesterov's method as well as Theorem 3 refers to the λ-strong convexity of the function, satisfying $0 < \lambda \leqslant \frac{\min \frac{1}{\pi_i}}{\lambda_{\max}((-\omega)^\dagger)}$.*

With the optimal damping parameter, $\mu_* = -\sqrt{|\alpha_*|}$ is the optimal convergence rate of `Chi-squared` *method in terms of spectral analysis.*

4.4.2 con-Fisher method

To check if $\mathcal{U}(p) = \mathrm{I}(p\|\pi) = \frac{1}{4}\sum_{i=1}^n \sum_{j\neq i} \omega_{ij}\theta_{ij}(\log\frac{p_i}{\pi_i} - \log\frac{p_j}{\pi_j})^2$ is geodesically λ-strongly convex, we need to find a positive constant $\lambda > 0$ such that

$$\begin{aligned}
&\mathrm{I}(\hat{p}\|\pi) - \mathrm{I}(p\|\pi) - (\hat{p}-p)\nabla_p \mathrm{I}(p\|\pi) \\
=&\frac{1}{4}\sum_{i=1}^n\sum_{j\neq i}\omega_{ij}\theta_{ij}[\log\frac{\pi_j^2 p_i \hat{p}_i}{\pi_i^2 p_j \hat{p}_j}\log\frac{\hat{p}_i p_j}{p_i \hat{p}_j}] - \sum_{i=1}^n(\hat{p}_i - p_i)\frac{\pi_i}{p_i}\sum_{j\neq i} Q_{ij}\theta_{ij}\log\frac{\pi_j p_i}{\pi_i p_j} \\
=&\sum_{i=1}^n\sum_{j\neq i}\omega_{ij}\theta_{ij}\left[\frac{1}{4}\log\left(\frac{\hat{p}_i p_j}{p_i\hat{p}_j}\right)^2 + \frac{1}{2}\log\left(\frac{\hat{p}_i p_j}{p_i \hat{p}_j}\right)\log\frac{\pi_j p_i}{\pi_i p_j} - (\frac{\hat{p}_i}{p_i} - 1)\log\frac{\pi_j p_i}{\pi_i p_j}\right] \\
\geqslant&\frac{\lambda}{2}(\hat{p}-p)K^\dagger(\hat{p}-p)^\top,
\end{aligned}$$

where $K = (K_{ij})$ and $K_{ij} = -\omega_{ij}\theta_{ij}$. It seems nontrivial to find a sufficient condition at this moment. Thus we compute the Hessian $\mathrm{D}^2\mathcal{U}$ instead. Appendix C provides detailed derivations of the explicit form of the Hessian and its related properties.

In theory, to find λ, one may study the Rayleigh quotient problem

$$\min_{\psi \mathbb{1}_n^\top = 0} \frac{\psi K \mathrm{D}^2\mathcal{U}(p) K^\top \psi^\top}{\psi K \psi^\top}. \tag{30}$$

In computation, (30) can be reformulated as an eigenvalue problem. See Appendix C for details.

Lemma 7 *Given a strictly positive target probability π and an irreducible transition-rate matrix Q, we consider the probability manifold $\mathbb{P}(V)$ on a graph $G = (V, E, \omega)$, with the potential $\mathcal{U}(p) = \mathrm{I}(p\|\pi) = \frac{1}{4}\sum_{i=1}^{n}\sum_{j\neq i}\omega_{ij}\theta_{ij}(\log\frac{p_i}{\pi_i} - \log\frac{p_j}{\pi_j})^2$. Let K be the associated Onsager response matrix. Then there exists a positive constant λ such that $\mathrm{D}^2\mathcal{U}(p)|_{p=\pi} \geqslant \lambda K^{\dagger}$.*

Consequently, we may use this λ computed from `con-Fisher` at the $p = \pi$ to construct the asymptotical limit of the damping parameter $\gamma(t) = 2\sqrt{\lambda}$ in the `log-Fisher` method.

5 Numerical Schemes

In this section, we outline numerical schemes employed in subsequent numerical experiments for the ODE integrator of (21) and the jump process associated with (22), (23) in aMCMC.

5.1 ODE Integrator

The classical Hamiltonian Monte Carlo (HMC), requires specialized discretization schemes, such as the symplectic Euler scheme for volume-preserving or the Leapfrog scheme for quantity-conserving (see Neal, 2011). For our proposed damped Hamiltonian dynamics (21), we adopt the staggered scheme with splitting methods. Specifically, given a prescribed damping parameter $\gamma(t)$, we can express (21) as:

$$\begin{cases} \dfrac{\mathrm{d}p(t)}{\mathrm{d}t} = A(p(t), \psi(t)), \\ \dfrac{\mathrm{d}\psi(t)}{\mathrm{d}t} = B(p(t), \psi(t), \gamma(t)), \end{cases}$$

where $A(p(t), \psi(t)) = \partial_\psi \mathcal{H}(p(t), \psi(t))$ and $B(p(t), \psi(t), \gamma(t)) = -\gamma(t)\psi(t) - \partial_p \mathcal{H}(p(t), \psi(t))$. The associated staggered scheme is given by

$$\begin{cases} \dfrac{p^{(k+1)} - p^{(k)}}{\Delta t} = A(p^{(k)}, \psi^{(k)}), & \text{(31a)} \\ \dfrac{\psi^{(k+1)} - \psi^{(k)}}{\Delta t} = B(p^{(k+1)}, \psi^{(k)}, \gamma^{(k)}). & \text{(31b)} \end{cases}$$

When $\mathbb{K}$ does not depend on p, the Hamiltonian $\mathcal{H}(p, \psi)$ is separable, (31) is indeed the symplectic Euler scheme. Note that (31) involves designing an interacting particle system, which requires simultaneously integrating the coupled system of all state variables and momentum variables. In contrast, the ODE solver of the forward master equation (3) in the MH can integrate each state variable independently.

Since the Hamiltonian $\mathcal{H}(p, \psi)$ is proved to decay according to (20), it is desirable to use a numerical scheme that ensures this decay property. In our experiments, we select a sufficiently small step size to numerically approximate this property.

5.2 Jump Process

For the forward master equation in (21a), an equivalent form (22) is proposed for the jump process of state variables. Consequently, the transition probability matrix in the classical

MCMC with the MH update and in aMCMC can be expressed in the form of $P = I_n + Q\Delta t$, where MH takes $Q = Q^{\mathrm{MH}}$ (5) and aMCMC takes $Q = \bar{Q}^r_\psi$. Note that the construction of $\bar{Q}^r_\psi$ defined in (23) requires $p_i(t)$ to be strictly positive. Theorem 2 guarantees the strict positivity of $p(t)$ in `log-Fisher` and `con-Fisher` methods. In numerical experiments, we will use *restart* techniques, detailed in Section 6, to guarantee positivity. Additionally, the step size Δt must be small enough to ensure every entry in P is nonnegative.

The Metropolis-Hastings update is described in Algorithm 2. The transition probability matrix $P^{\mathrm{MH}} = I_n + Q^{\mathrm{MH}}\Delta t$ is time-homogeneous. In Matlab, we use the `randsample` function to simulate transitions of particles from one state to another. And we use the `histcounts` function to count the number of particles in each state.

Algorithm 2: The Metropolis-Hastings sampling in Matlab

```
Input: Initial distribution $\rho^{(0)}$, transition rate matrix $Q^{\mathrm{MH}}$, total states $n$,
       sampling size $M$, step size $\Delta t$ and total iterations $N$.
Output: Terminal distribution $\rho^{(N)}$
Initialize the transition probability $P^{\mathrm{MH}} \leftarrow I_n + Q^{\mathrm{MH}}\Delta t$;
Initialize bin $\leftarrow$ histcounts(randsample($n, M$, true, $\rho^{(0)}, n$));
// randsample: for $M$ particles, draw with replacement into $n$ bins
   with probability $\rho^{(0)}$.
// bin(i): the number of particles in state i counted by histcounts
for iter $\leftarrow 0$ to $N$ do
    for state $\leftarrow 0$ to $n$ do in parallel
        tmp(state, :) $\leftarrow$ histcounts(randsample($n$, bin(state), true, $P^{\mathrm{MH}}$(state, :
         )), $n$);
        // tmp(state,i): the number of particles that jump from
           previous state to state i.
    end
    bin($i$) $\leftarrow \sum_j$ tmp($j, i$) ;                         // $i = 1, 2, \ldots, n$
end
$\rho^{(N)} \leftarrow \frac{1}{M}$bin.
```

For aMCMC, there is no corresponding jump process for the momentum variables. Thus, we design a jump process to solve (31a) while we employ the ODE integrator on (31b).

5.3 Initialization and Restart

For the classical MCMC with the MH update, we initialize the state distribution $p(0)$ by a uniform distribution. While it is well known that MCMC methods generally require a *burn-in* period—where the first few iterations are discarded to mitigate the effects of a poor starting point (Geyer, 2011)—we choose to include all iterations in our experiments in Section 6. For aMCMC, in addition to initializing the state variables with a uniform distribution, we must also initialize the momentum variables. We observe the following relationship

- $\theta_{ij} = 1$ in `Chi-squared` and `con-Fisher` methods. Under the choice $\psi_j = -\frac{p_j}{\pi_j}$, (21a) becomes

$$\frac{\mathrm{d}p_i}{\mathrm{d}t} = \psi_i \sum_{j \neq i} \omega_{ij} - \sum_{j \neq i} \psi_j \omega_{ji} = -\frac{p_i}{\pi_i} \sum_{j \neq i} \pi_i Q_{ij} + \sum_{j \neq i} \frac{p_j}{\pi_j} \pi_j Q_{ji} = \sum_{j \neq i} p_j Q_{ji} - p_i Q_{ij},$$

  which is precisely the forward master equation for the classical MCMC. Under such a choice of initial momentum variables, the state variables in (21a) via `Chi-squared` and `con-Fisher` methods will evolve the same way as the classical MCMC at the start of the experiment.

- $\theta_{ij} = \frac{\frac{p_j}{\pi_j} - \frac{p_i}{\pi_i}}{\log \frac{\pi_i p_j}{\pi_j p_i}}$ in `KL` and `log-Fisher` methods. Under the choice $\psi_j = -\log \frac{p_j}{\pi_j}$, (21a) becomes

$$\frac{\mathrm{d}p_i}{\mathrm{d}t} = \sum_{j \neq i} (-\log \frac{p_i}{\pi_i} + \log \frac{p_j}{\pi_j}) \frac{\frac{p_i}{\pi_i} - \frac{p_j}{\pi_j}}{\log \frac{p_i}{\pi_i} - \log \frac{p_j}{\pi_j}} \pi_i Q_{ij} = \sum_{j \neq i} \frac{p_j}{\pi_j} \pi_j Q_{ji} - \frac{p_i}{\pi_i} \pi_i Q_{ij},$$

  which recovers the forward master equation for the classical MCMC. Thus, the state variables in (21a) via `Chi-squared` and `con-Fisher` methods will evolve the same way as the classical MCMC at the start of the experiment.

As a result, we can use the classical MCMC as a *warm-start* in aMCMC. The state variables in the last-step of the warm-start can be used to construct an initialization of the momentum variables, by $\psi_j(0) = -\frac{p_j(0)}{\pi_j}$ for `Chi-squared` and `con-Fisher` methods, or by $\psi_j(0) = -\log \frac{p_j(0)}{\pi_j}$ for `KL` and `log-Fisher` methods.

Given that the target distribution π is strictly positive and the Markov chain is reversible and irreducible, Theorem 2 ensures strict positivity of the state variables $p(t)$ for both the `log-Fisher` and `con-Fisher` methods. It suggests that when $p_i(t)$ is close to zero for some state i, there is sufficiently large momentum to pull $p_i(t)$ away from zero. Nonetheless, in numerical experiments, accumulation of truncation errors could lead to instances where the state variables become zero or even negative.

For the ease of our discussion, denote by $p^{(k)}_{\text{i,ode}}$ the numerical value of $p_i(t)$ in (21) at the k-th iteration using Euler scheme. Denote by $p^{(k)}_{\text{i,jump}}$ the empirical probability density of state i at the k-th iteration, which is calculated by the ratio between the number of particles in state i at the k-th iteration and the total number of particles M. We first opt for adaptive step sizes (see line 11-14 in Algorithm 3) to prevent any state variables $p^{(k)}_{\text{i,ode/jump}}$ from becoming negative. In practice, this is usually sufficient when solving $p^{(k)}_{\text{i,ode}}$ in (21) with `log-Fisher` and `con-Fisher` methods using Euler scheme. However, due to the random sampling following the transition probability matrix $P = I_n + \bar{Q}^r_\psi \Delta t$, a small sample size M may cause the empirical particle density $p^{(k)}_{\text{i,jump}}$ to become zero even when $p^{(k)}_{\text{i,ode}} > 0$.

Second, if $p^{(k)}_{\text{i,ode/jump}} = 0$ after some iterations, we employ a simple *restart* mechanism by the MH update that works for any aMCMC method, described below.

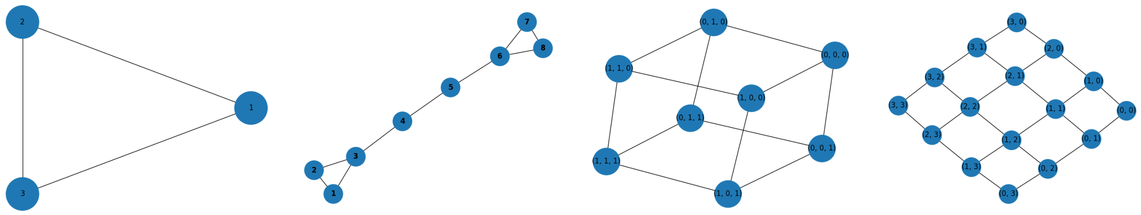
Restart (ODE solver): if $p_{\text{i,ode}}^{(k)} = 0$ for some state i, for all state j we set $\psi_{\text{j,ode}}^{(k)} = -\frac{p_{\text{j,ode}}^{(k)}}{\pi_j}$ in `Chi-squared` and `con-Fisher` methods, or $\psi_{\text{j,ode}}^{(k)} = -\log \frac{p_{\text{j,ode}}^{(k)}}{\pi_j}$ in `KL` and `log-Fisher` methods. This choice ensures the $(k+1)$-th iteration on the state variables effectively runs *one* iteration of the MH update, thereby pulling $p_{\text{i,ode}}$ away from 0.

Restart (jump process): if $p_{\text{i,jump}}^{(k)} = 0$, i.e., the number of particles at state i is zero after random sampling at the k-th iteration, then we add *one* particle to state i. Note that this also increases the total number of particles by *one*. The updated value of $p_{\text{i,jump}}^{(k)}$ is then set to $(M_{\text{old}}+1)^{-1}$, where M_{old} is the total amount of particles before adding the new particle. Then we set $\psi_{\text{j,jump}}^{(k)} = -\frac{p_{\text{j,jump}}^{(k)}}{\pi_j}$ in `Chi-squared` and `con-Fisher` methods, or $\psi_{\text{j,jump}}^{(k)} = -\log \frac{p_{\text{j,jump}}^{(k)}}{\pi_j}$ in `KL` and `log-Fisher` methods. In our experiments, the `log-Fisher` method barely requires restart.

Multiple iterations of Metropolis–Hastings restarts or the addition of multiple particles can be performed in experiments to ensure the strict positivity of state variables, albeit at the cost of deviating from the proposed damped Hamiltonian flow.

6 Numerical Examples

In this section, we compare the proposed aMCMC with the classical Metropolis-Hastings algorithm. Our code is available at `https://github.com/silentmovie/AMCMC`.



(a) Circular graph. (b) Two-loop graph. (c) Hypercube. (d) Lattice.

Figure 1: Graph configurations employed in the numerical experiments.

All experiments are performed on a variety of connected and irreducible graphs, as illustrated in Figure 1. Given a graph $G = (V, E)$, we employ the simple random walk as the candidate kernel, defined by: $q_{ij} = \begin{cases} \dfrac{1}{\deg(i)} & e_{ij} \in E, \\ 0 & \text{otherwise.} \end{cases}$ The corresponding Metropolis–Hastings transition rate matrix is constructed as $Q_{ij}^{\text{MH}} = \min\left\{\frac{\pi_j}{\pi_i} q_{ji}, q_{ij}\right\}$. The weight matrix ω is given by $\omega_{ij} = \pi_i Q_{ij}^{\text{MH}}$. We compare MCMC and aMCMC on the same weighted undirected graph $G = (V, E, \omega)$. The numerical schemes are described in Section 5. In what follows, we illustrate two representative methods: the `Chi-squared` method for small-scale graphs, and the `log-Fisher` method for various graph configurations.

We choose `Chi-squared` method over `KL` method, to numerically validate Theorem 5. We show the numerical advantage of `log-Fisher` over the MH algorithm, by selecting

```
Algorithm 3: The aMCMC sampling  in MATLAB
Input: Initial distribution $\rho_0$, transition rate matrix $Q^{\mathrm{MH}}$, total states $n$, sampling
       size $M$ of total particles, step size $\Delta t$, warm-start iterations $L$ and total
       iterations $N$.
Output: Terminal distribution $\rho^{(N)}$
Initialize $\rho^{(L)} \leftarrow L$ iterations of MH updates from Algorithm 2 ;  // warm-start
Initialize $\psi^{(L)}$ from Section 5.3, $\bar{Q}^r_\psi$ from (23), $P \leftarrow I_n + \bar{Q}^r_\psi \Delta t$;
// the strategy for $\psi^{(L)}$ depends on the choices of $\theta$.
Initialize bin $\leftarrow$ histcounts(randsample($n, M$, true, $\rho^{(L)}, n$);
// bin(i):  the number of particles in state i.
for iter $\leftarrow L+1$ to $N$ do
    for state $\leftarrow 0$ to $n$ do in parallel
        tmp(state, :) $\leftarrow$ histcounts(randsample($n$, bin(state), true, $P$(state, :)), $n$);
        // tmp(state,i):  the number of particles that jump from
           previous state to state i.
    end
    bin($i$) $\leftarrow \sum_j$ tmp($j, i$) ;                              // $i = 1, 2, \ldots, n$
    $\rho^{(\mathtt{iter})} \leftarrow \frac{1}{\text{total particles}}$bin;  // total particles may change in the restart
     block.
    $\psi^{(\mathtt{iter})}$ from (21b), $\bar{Q}^r_\psi$ from (23) ;      // $\psi$ is updated by an ODE solver.
    // Use adaptive step sizes to obtain the matrix $P$
    while $P$ not a transition probability do
        $\Delta t^{(\mathtt{iter}+1)}$ decreases by a factor of 10;
        $P \leftarrow I_n + \bar{Q}^r_\psi \Delta t^{(\mathtt{iter}+1)}$;
    end
    // restart block, see Section 5.3
    if necessary  then
        $\psi^{(\mathtt{iter})}$ from Section 5.3;
        add particles to empty bin($i$)
    end
end
```

some damping parameter by convention or suggested by `con-Fisher` in the asymptotic limit $\frac{p_i}{\pi_i} \approx 1$. In all experiments, comparisons are made based on a fixed total number of iterations, while allowing step sizes (and hence the effective time span) to vary across different methods.

6.1 Acceleration and Accuracy via `Chi-squared` Method

We start with a toy example on the circular graph C_3 (see Figure 1a). The target distribution is $\pi = [0.9913, 0.0044, 0.0043]$, so that the largest negative eigenvalue of Q^{MH} is close to zero. More specific, $\alpha_* \approx -0.5044$. By choosing the damping parameter $\gamma(t) = 2\sqrt{|\alpha_*|} = 1.4220$ as suggested by Theorem 5, we obtain the largest negative eigenvalue of matrix L (25) is $\mu_* = -0.7102 < \alpha_*$, thereby suggesting an accelerated convergence rate by the `Chi-squared` method.

We use the initialization discussed in Section 5.3. Aside from $\psi(0)$, warm-start is not employed in this experiment. The parameters for MH and `Chi-squared` methods are chosen identically:

$$\text{sampling size } M = 10^6, \qquad \text{step size } \Delta t = 0.1 \text{ or } 0.01, \qquad \text{total iterations } N = 650 \text{ or } 6500.$$

We use small step sizes and a large sampling size to accurately approximate the proposed damped Hamiltonian dynamics, as validated by Figure 2a. Notably, neither adaptive step size adjustment nor restart mechanisms were required in these experiments, yielding a consistent effective time span $[0, 65]$ across all examples shown in Figure 2.

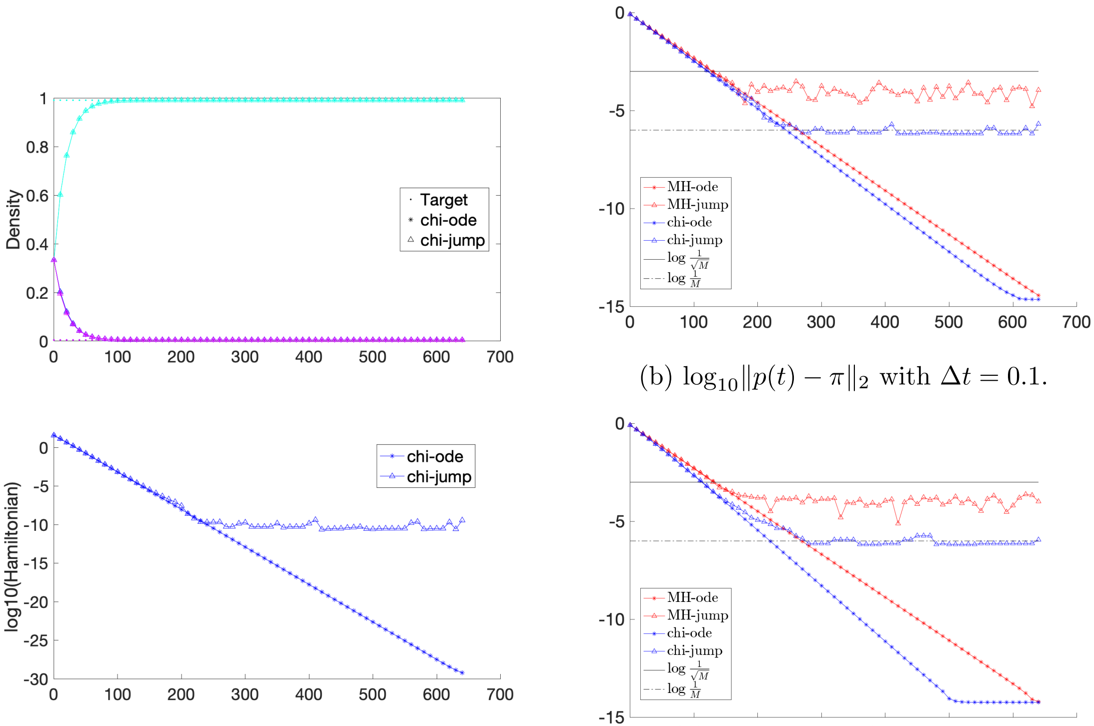
Figure 2b and Figure 2c corroborate the theoretical results in Theorem 5 under different step sizes, demonstrating that `Chi-squared` achieves a faster convergence than the MH update in both ODE and jump process. Notably, jump process sampling achieves a higher order of accuracy $\mathcal{O}(\frac{1}{M})$. When a smaller step size is used, as shown in Figure 2c, the numerical dynamics adhere more closely to the damped Hamiltonian, further accelerating convergence.

6.2 Acceleration and Accuracy via `log-Fisher` Method

Now we test `log-Fisher` method on the other graph configurations in Figure 1. Certain geometric features of a Markov Chain influence its mixing time, with bottlenecks being a notable example (See Levin et al., 2009). A two-loop graph (see Figure 1b) bridged by a thin "bottleneck" is one of those examples. We use `log-Fisher` method on this configuration first. The target distribution is given by $\pi = [\frac{4}{27}, \frac{4}{27}, \frac{4}{27}, \frac{1}{18}, \frac{1}{18}, \frac{4}{27}, \frac{4}{27}, \frac{4}{27}]$ so that the largest negative eigenvalue to the associated Q^{MH} is -3.79×10^{-2}. The initialization on $\psi_i(0) = -\log\frac{p_i(0)}{\pi_i}$ follows the discussion in Section 5.3 without further warm-start period. The parameters for MH and `log-Fisher` methods are chosen identically:

$$\text{sampling size } M = 10^4, \qquad \text{step size } \Delta t = 0.1 \qquad \text{total iterations } N = 1000.$$

We use the following damping parameter inspired by the Nesterov's accelerated gradient method. We also set a lower bound on the damping parameter as suggested by `con-Fisher`



(b) $\log_{10}\|p(t) - \pi\|_2$ with $\Delta t = 0.1$.

(a) Validation of `Chi-squared`.

(c) $\log_{10}\|p(t) - \pi\|_2$ with $\Delta t = 0.01$.

Figure 2: Sampling on C_3 graph Figure 1a via `Chi-squared` method. x-axes are in iterations and all results are over time span $[0, 65]$ when both ODE solvers achieve machine precision. (a) Validation of the `Chi-squared` method using step size $\Delta t = 0.1$, showing density convergence and the decay of $\log_{10}(\mathcal{H}(p(t), \psi(t)))$. (b) and (c) show the decay of $\log_{10}\|p(t) - \pi\|_2$ for step sizes $\Delta t = 0.1$ and $\Delta t = 0.01$ respectively. Notably, chi-jump achieves an error of order $\mathcal{O}(1/M)$.

method in Section 4.4:

$$\gamma(t) = \begin{cases} 0.5 & t < 3; \\ \max\left\{\dfrac{3}{t-2}, 0.6\right\} & t \geqslant 3. \end{cases}$$

With the above choice of parameters, restart or adaptive step sizes are not needed in this experiment, and we compare the dynamics driven by the relative Fisher information with the MH update. As shown in Figure 3, the jump process generated by the `log-Fisher` method exhibits faster convergence and achieves higher accuracy compared to the MH counterpart, when the MH sampling reaches to its accuracy limit $\mathcal{O}(\frac{1}{\sqrt{M}})$. Though the ODE integrator via `log-Fisher` attains machine precision more slowly than that via MH, the convergence rate can potentially be improved by tuning the damping parameter when higher accuracy is desired.

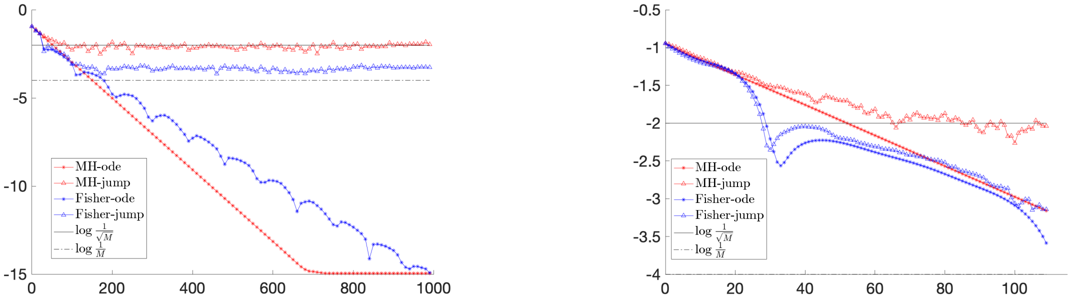


Figure 3: Sampling on a two-loop graph of node 8 (see Figure 1b) via `log-Fisher` method. x-axes are in iterations with the step size $\Delta t = 0.1$. y-axes is on error $\log_{10}\|p(t) - \pi\|_2$. The right figure zooms in the first 120 iterations of the left figure. The jump process via `log-Fisher` method exhibits faster convergence and higher accuracy.

In the second example, we compare the jump process on a hypercube graph (see Figure 1c) of 64 nodes. The target distribution is designed to place significant mass at two maximally distant vertices in the hypercube, with the remaining nodes sharing uniformly small mass. Specifically, we set $\pi = \frac{1}{Z}[16, 1, \ldots, 1, \ldots, 1, 16]$. Note that the normalizing constant Z is not required in `log-Fisher` method. The parameters are specified as follows:

$$\text{sampling size } M = 10^4, \qquad \text{step size } \Delta t = 0.01 \qquad \text{total iterations } N = 6 \times 10^3.$$

Note that largest negative eigenvalue $\alpha_* \approx -0.0468$ of Q^{MH}. We choose the damping parameter inspired by NAG and `con-Fisher` method as:

$$\gamma(t) = \max\left\{\frac{2\sqrt{-\alpha_*}}{t}, 0.17\right\} \qquad t \geqslant 1,$$

while we run a warm-start of 100 iterations (i.e., $t < 1$). With this choice of parameters, restart or adaptive step sizes are not needed in our experiments.

In the final example, we evaluate the performance of jump process via `log-Fisher` method on a 25×25 lattice graph shown in Figure 1d. The target distribution is constructed

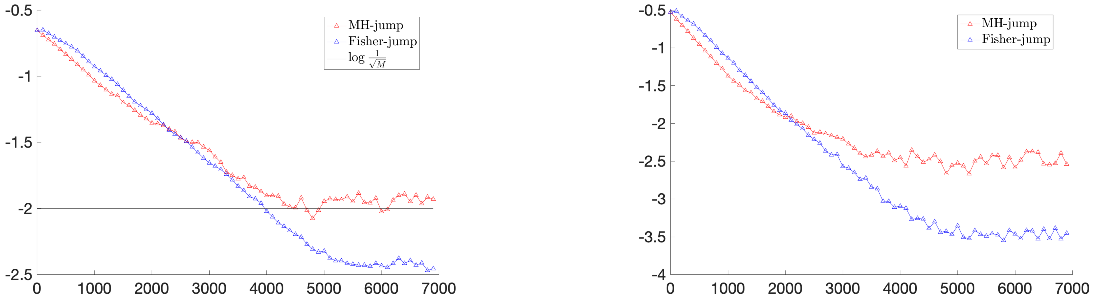


Figure 4: Sampling on a hypercube graph of 64 nodes (see Figure 1c) via `log-Fisher` method. x-axes are in iterations with step size $\Delta t = 0.01$. The left figure shows the approximation error $\log_{10}\|p(t) - \pi\|_2$ w.r.t the target distribution. The right figure shows the approximation error $\log_{10}|\sum_{i=1}^{n} p_i(t) \log \frac{p_i(t)}{Z\pi} - (-\log Z)|$ w.r.t the normalizing constant Z.

as a mixture of two Gaussian distributions on $[0, 1]^2$:

$$\pi(x) = \frac{1}{Z}\left[\exp\left(-10\|x - x_1\|_2^2\right) + \exp\left(-40\|x - x_2\|_2^2\right)\right], \tag{32}$$

where $x_1 = (0.25, 0.25)$ and $x_2 = (0.75, 0.75)$. This setting poses a challenging multimodal sampling problem. The parameters are specified as follows:

$$\text{sampling size } M = 5 \times 10^5, \quad \text{default step size } \Delta t = 0.01 \quad \text{total iterations } N = 1.5 \times 10^5.$$

In the numerical example presented in Figure 5 and Figure 6, we initialize the Hamiltonian dynamics driven by the relative Fisher information using a warm-start by MH over the first 2999 iterations (i.e., $t < 30$, accounting for 2.00% of the total iterations), as discussed in Section 5.3. Beyond this point, we employ a constant damping parameter $\gamma(t) = 2\sqrt{\lambda_*} = 0.0065$, where λ_* is obtained from the Rayleigh quotient suggested by **`con-Fisher`**. The MH maintains a fixed effective step size throughout the simulation, resulting in a total effective time span of of $[0, 1500]$. In contrast, the `log-Fisher` method employs an adaptive step size scheme, leading to a slightly reduced effective time span of approximately $[0, 1449.9]$, which corresponds to 96.66% of duration. When the restart mechanism is triggered at the k-th iteration, damping is effectively disabled. Over the full trajectory, this restart procedure is activated 1243 times in the jump process (approximately 0.85% of the total iterations). The frequency of restarts per 1000 iterations ranges from 0 to 14.

7 Discussions

In this paper, we design accelerated Markov Chain Monte Carlo (aMCMC) algorithms to sample target distributions on a discrete domain, by incorporating Nesterov's acceleration into the Metropolis-Hastings (MH) algorithm. The MCMC sampling algorithm can be formulated as an optimization problem in the space of probabilities. We study the MH

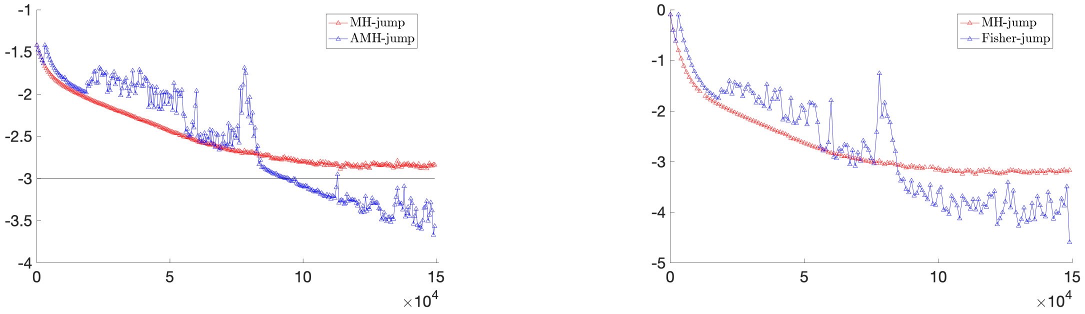


Figure 5: Sampling on a 25×25 lattice graph (see Figure 1d) via `log-Fisher` method for the target distribution as a mixture of two Gaussian distributions. x-axes are in iterations. The left figure shows the approximation error $\log_{10}\|p(t)-\pi\|_2$ w.r.t the target distribution. The right figure shows the approximation error $\log_{10}|\sum_{i=1}^{n} p_i(t)\log\frac{p_i(t)}{Z\pi} - (-\log Z)|$ w.r.t the normalizing constant Z. The jump process via `log-Fisher` achieves to a higher accuracy when that via MH is approaching to $\mathcal{O}(\frac{1}{\sqrt{M}})$.

algorithm as a gradient descent method for the KL divergence in the discrete Wasserstein-2 space. Our algorithm utilizes moment-based accelerated gradient methods in discrete Wasserstein-2 spaces and then develops an interacting Markov particle process in discrete domains. Numerical examples on various graphs illustrate the effectiveness of the aMCMC algorithm.

In addition, the proposed accelerated sampling algorithm is related to the score functions on discrete domains. In Lou et al. (2024), the discrete score function $\nabla \log p = \frac{\nabla p}{p}$ is approximated by $[\frac{p_j(t)-p_i(t)}{p_i(t)}]_{j\neq i}$. In this paper, we approximate the score function of a detailed-balanced MCMC algorithm by $\log\frac{p_j}{p_i} = \log p_j - \log p_i$ and the logarithmic mean (2). Our approach applies to the Nesterov methods of optimization problems on probability manifolds. In future work, we shall study these differences in the approximation of discrete score functions in terms of sampling complexities.

We will investigate several open problems related to the proposed accelerated MCMC algorithm. First, we shall analyze the choice of the damping parameter $\gamma(t)$ with a given objective function. The choice of optimal parameters is related to the geometric calculations in the generalized Wasserstein-2 manifold Li (2025). Moreover, one could design a generalized discretization scheme that allows large time step sizes and study its convergence properties. Another interesting direction is to investigate the efficiency of score estimation in aMCMC. This estimation can be formulated as a variational framework based on time-reversible diffusion on discrete states.

Acknowledgments and Disclosure of Funding

S. Liu, W. Li, B. Zhou and X. Zuo are supported by AFOSR YIP award No. FA9550-23-1-0087, NSF RTG: 2038080 and FRG: 2245097. All authors would like to thank the support of the AMS MRC conference: Ricci Curvatures of Graphs and Applications to Data Science

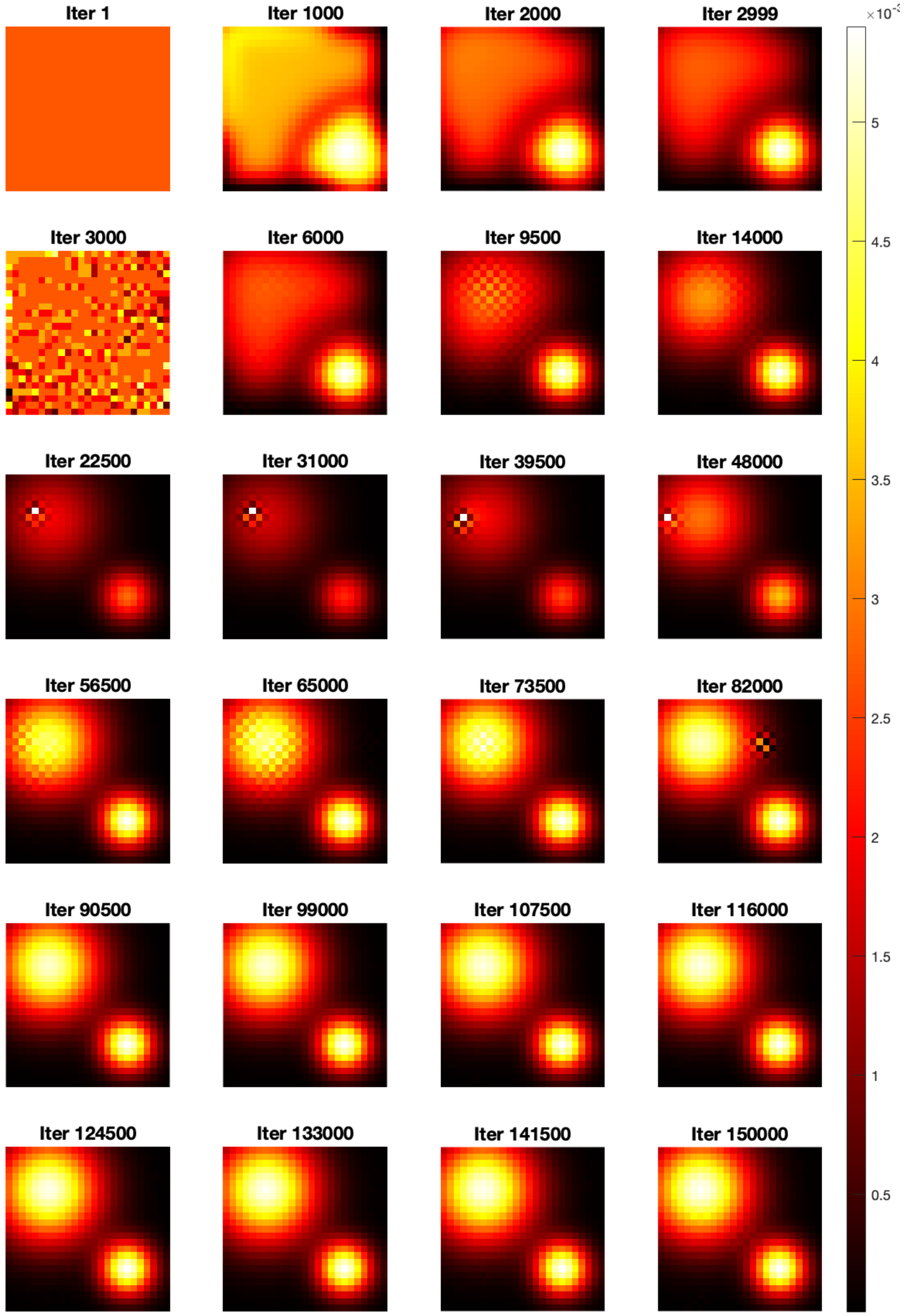


Figure 6: Sampling on a 25×25 lattice for a mixture of two Gaussian distributions supported on $[0,1]^2$, using the jump process induced by the `log-Fisher` method. The abrupt change between the 2999th and 3000th iterations suggests that a more refined choice of the damping parameter may be beneficial. Nonetheless, this simple parameter choice still demonstrates promising performance relative to the MH update.

in 2023. The discussion of this manuscript was initialized in the workshop. B. Zhou would like to extend appreciation to Andy Wan for a fruitful discussion.

Appendix A. Graph Laplacian and detailed derivation of Wasserstein gradient

Lemma 8 *Suppose the Markov chain is irreducible and reversible with respect to some strictly positive distribution π. Assume $p \in \mathbb{P}(V)$ is strictly positive. Then, the graph Laplacian $\mathbb{K}(p)$ is positive semi-definite with* $\mathrm{Ker}(\mathbb{K}(p)) = \mathrm{span}(\mathbb{1}_n)$.

Proof Recall $\mathbb{K}(p)$ is the graph Laplacian of a weighted undirected graph with a weight matrix $(\omega_{ij}\theta_{ij}(p))$, for any $x = [x_1, \ldots, x_n]$, the quadratic form is

$$x\mathbb{K}(p)x^\top = -\sum_{1\leqslant i<j\leqslant n} \mathbb{K}_{ij}(p)(x_i - x_j)^2 = -\sum_{1\leqslant i<j\leqslant n} \omega_{ij}\theta_{ij}(p)(x_i - x_j)^2 \leqslant 0.$$

The last inequality is due to $\omega_{ij} \geqslant 0$ and $\theta_{ij}(p) = \frac{\frac{p_i}{\pi_i} - \frac{p_j}{\pi_j}}{f'(\frac{p_i}{\pi_i}) - f'(\frac{p_j}{\pi_j})} > 0$. This verifies that $\mathbb{K}(p)$ is negative definite.

Furthermore, assume $y \in \mathrm{Ker}(\mathbb{K}(p))$, this leads to $y\mathbb{K}(p)y^\top = 0$, which further leads to $\mathbb{K}_{ij}(p)(y_i - y_j)^2 = 0$ for any $1 \leqslant i < j \leqslant n$.

Recall that the Markov chain is reducible; for arbitrary $1 \leqslant k < l \leqslant n$, we can find a path of nodes $k = i_1 \to i_2 \to \cdots \to i_m = l$ such that $\omega_{i_j\, i_{j+1}} > 0$ for $j = 1, \ldots, m-1$, which yields $\mathbb{K}_{i_j\, i_{j+1}}(p) > 0$. This leads to $y_k = y_{i_1} = \cdots = y_{i_m} = y_l$. This verifies that $y \in \mathrm{span}(\mathbb{1}_n)$, thus $\mathrm{Ker}(\mathbb{K}(p)) \subset \mathrm{span}(\mathbb{1}_n)$. It is clear that $\mathrm{span}(\mathbb{1}_n) \subset \mathrm{Ker}(\mathbb{K}(p))$ since $\mathbb{K}(p)$ is row-sum-zero and symmetric. Thus we have proved $\mathrm{Ker}(\mathbb{K}(p)) = \mathrm{span}(\mathbb{1}_n)$. ■

Derivation of $\mathrm{grad}\mathrm{D}_f(p\|\pi)$: Denote the tangent space of $\mathbb{P}(V)$ at p by $\mathrm{T}_p\mathbb{P}(V)$, and denote the Wasserstein gradient on $(\mathbb{P}(V), g_W)$ by $\mathrm{grad}\mathrm{D}_f(p\|\pi) \in \mathrm{T}_p\mathbb{P}(V) = \mathbb{1}_n^\perp$. We consider arbitrary curve $\{p(t)\}$ passing through p at $t = 0$, by the definition of the gradient operator on Riemannian manifold (see Lee, 2018, Chapter 2)

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{D}_f(p_t\|\pi)\Big|_{t=0} = g_W(\mathrm{grad}\mathrm{D}_f(p\|\pi),\ \dot{p}(0)).$$

Denote $\nabla_p\mathrm{D}_f(p\|\pi)$ as the flat gradient of f-divergence with respect to p, for any $\dot{p}(0) \in \mathbb{1}_n^\perp$, we have

$$\nabla_p\mathrm{D}_f(p\|\pi)(\dot{p}(0))^\top = \mathrm{grad}\mathrm{D}_f(p\|\pi)\mathbb{K}(p)^\dagger(\dot{p}(0))^\top.$$

Let $\Pi : \mathbb{R}^n \to \mathbb{R}^n$ be the orthogonal projection onto $\mathrm{T}_p\mathbb{P}(V) = \mathbb{1}_n^\perp$, the above equation yields

$$\Pi(\mathrm{grad}\mathrm{D}_f(p\|\pi)\mathbb{K}(p)^\dagger) = \Pi(\nabla_p\mathrm{D}_f(p\|\pi)).$$

Since $\mathbb{K}(p)^\dagger$ is the Moore-Penrose inverse of $\mathbb{K}(p)$, we have $\Pi = \mathbb{K}(p)^\dagger\mathbb{K}(p) = \mathbb{K}(p)\mathbb{K}(p)^\dagger$. The above equation then leads to

$$(\mathrm{grad}\mathrm{D}_f(p\|\pi)\mathbb{K}(p)^\dagger)\mathbb{K}(p)\mathbb{K}(p)^\dagger = \nabla_p\mathrm{D}_f(p\|\pi)\mathbb{K}(p)\mathbb{K}(p)^\dagger.$$

The left-hand side equals $\mathrm{grad}\mathrm{D}_f(p\|\pi)\mathbb{K}(p)^\dagger$, as $\mathrm{grad}\mathrm{D}_f(p\|\pi) \in \mathrm{T}_p\mathbb{P}(V) = \mathbb{1}_n^\perp$. Then,

$$[\mathrm{grad}\mathrm{D}_f(p\|\pi) - \nabla_p\mathrm{D}_f(p\|\pi)\mathbb{K}(p)]\,\mathbb{K}(p)^\dagger = 0.$$

Furthermore, as

$$\mathrm{Ker}(\mathbb{K}(p)^{\dagger}) = \mathrm{Ker}(\mathbb{K}(p)) = \mathrm{span}(\mathbb{1}_n),$$

we have

$$\mathrm{grad}\mathrm{D}_f(p\|\pi) - \nabla_p \mathrm{D}_f(p\|\pi)\mathbb{K}(p) \in \mathrm{span}(\mathbb{1}_n).$$

On the other hand, notice that $\mathrm{Ran}(\mathbb{K}(p)) = \mathrm{Ker}(\mathbb{K}(p)^{\top})^{\perp} = \mathrm{Ker}(\mathbb{K}(p))^{\perp} = \mathbb{1}_n^{\perp}$, we have $\nabla_p \mathrm{D}_f(p\|\pi)\mathbb{K}(p) \in \mathbb{1}_n^{\perp}$. As a result,

$$\mathrm{grad}\mathrm{D}_f(p\|\pi) - \nabla_p \mathrm{D}_f(p\|\pi)\mathbb{K}(p) \in \mathrm{span}(\mathbb{1}_n) \cap \mathbb{1}_n^{\perp} = 0.$$

Appendix B. Eigenvalues and eigenvectors of Q-matrix

While the majority of the MCMC literature emphasizes the transition probability matrix P, rather than the transition rate matrix Q, the two are closely related. For completeness, we include a self-contained discussion.

Recall $\omega = (\omega_{ij}) = \mathrm{diag}(\pi)Q$ is a symmetric and row-zero-sum matrix.

Lemma 9 (Sylvester's law of inertia) *Two symmetric square matrices A, B of the same size have the same number of positive, negative and zero eigenvalues if and only if they are congruent, that is, $B = QAQ^T$ for some non-singular matrix Q.*

Lemma 10 *Given a nondegenerate target distribution π and an irreducible transition rate matrix Q that satisfy the detailed balance condition. Now define $\omega = \mathrm{diag}(\pi)Q$. Then*

a) ω is a negative semi-definite matrix. Furthermore, $\mathrm{Ker}(\omega) = \mathrm{span}(\mathbb{1}_n)$, and ω is negative definite on $\mathbb{1}_n^{\perp}$.

b) The eigenvalues to Q consist of 0 with algebraic multiplicity 1, along with $(n-1)$ negative eigenvalues.

Proof For part a), follow the same proof technique in Theorem 8, we can show that for arbitrary $x = [x_1, \ldots, x_n]$, $x\omega x^{\top} \leqslant 0$, as well as $\mathrm{Ker}(\omega) = \mathrm{span}(\mathbb{1}_n)$.

Combining these facts, together with ω being symmetric matrix, it is straightforward to prove that ω is negative definite on $\mathbb{1}_n^{\perp}$. Since 0 is apparently an eigenvalue to ω with the eigenvector $\mathbb{1}_n$, Recall all eigenvalues to a real symmetric matrix are real, thus ω persists $(n-1)$ real negative eigenvalues.

For part b), note that

$$Q = \mathrm{diag}(\pi)^{-1}\omega = \sqrt{\mathrm{diag}(\pi)}^{-1}(\sqrt{\mathrm{diag}(\pi)}^{-1}\omega\sqrt{\mathrm{diag}(\pi)}^{-1})\sqrt{\mathrm{diag}(\pi)},$$

Since ω and $(\sqrt{\mathrm{diag}(\pi)}^{-1}\omega\sqrt{\mathrm{diag}(\pi)}^{-1})$ are congruent, by the Sylverster's law of inertia, they have the same number of negative eigenvalues and zero eigenvalue. $(\sqrt{\mathrm{diag}(\pi)}^{-1}\omega\sqrt{\mathrm{diag}(\pi)}^{-1})$ and Q are similar, thus they have the same eigenvalues with the same algebraic multiplicity. In summary, Q has the same number of negative eigenvalues and zero eigenvalue as ω. ■

Alternatively, given an irreducible transition rate matrix Q, we may define $P = I_n + Q\Delta t$ for small enough Δt such that P is an irreducible, aperiodic transition probability matrix. Once applying Perron-Frobenius Theorem onto P and combining the detailed balance, we obtain the same conclusion about Q.

Lemma 11 (Perron-Frobenius Theorem) *Any irreducible, aperiodic transition matrix P has an eigenvalue $\lambda_0 = 1$ with algebraic multiplicity 1 as the Perron-Frobenius eigenvalue, and all other eigenvalues λ_i satisfy $|\lambda_i| < 1$.*

Appendix C. Hessian and geodesic convexity

We follow the setting and notations of discussions in Mielke (2013). Consider the space $\mathbb{P}(V)$ of probability measures on a graph $G = (V, E, \omega)$, with the Onsager's response matrix $\mathbb{K}$ defined in (9), let $\{p(t)\}_{t\geq 0}$ be a geodesic on $\mathbb{P}(V)$, then it satisfies the geodesic equations:

$$\begin{cases} \dot{p}(t) = \psi\mathbb{K} \\ \dot{\psi}(t) = -\nabla_p\left(\frac{1}{2}\psi\mathbb{K}\psi^\top\right), \end{cases}$$

Any functional $\mathcal{U}(p)$ that is geodesically λ-strongly convex can be characterized by

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathcal{U}(p(t)) \geqslant \lambda\dot{p}(t)\mathbb{K}^\dagger(p(t))(\dot{p}(t))^\top.$$

In our case, $\mathbb{K}(p)$ does not depend on p. Here we denote $\mathrm{D}^2\mathcal{U}(p)$ as the Hessian of $\mathcal{U}(p)$ in the Euclidean space.

The geodesic equation is simplified as

$$\begin{cases} \dot{p}(t) = \psi K \\ \dot{\psi}(t) = 0\,. \end{cases}$$

Thus $\mathcal{U}(p)$ is geodesically λ-strongly convex if and only if $\dot{p}\mathrm{D}^2\mathcal{U}(p)\dot{p}^\top \geqslant \lambda\dot{p}K^\dagger\dot{p}^\top$, or equivalently, $\psi K\mathrm{D}^2\mathcal{U}(p)K^\top\psi^\top \geqslant \lambda\psi K\psi^\top$. Thus under the flat metric K, we can say $\mathcal{U}(p)$ is geodescially λ-strongly convex, if there exists some $\lambda > 0$, such that for any t, the Rayleigh quotient problem has a lower bound along the dynamics:

$$\lambda \leqslant \min_{\psi\mathbb{1}_n^\top=0} \frac{\psi K\mathrm{D}^2\mathcal{U}(p(t))K^\top\psi^\top}{\psi K\psi^\top}.$$

In the `con-Fisher` method, recall $\mathcal{U}(p) = \mathrm{I}(p\|\pi) = \frac{1}{4}\sum_{i=1}^n\sum_{j\neq i}\omega_{ij}\theta_{ij}(\log\frac{p_i}{\pi_i} - \frac{p_j}{\pi_j})^2$. Thus $\partial_{ii}\mathcal{U}(p) = \sum_{j\neq i}\frac{\omega_{ij}\theta_{ij}}{p_i^2}[1 - \log\left(\frac{\pi_j p_i}{\pi_i p_j}\right)]$ and $\partial_{ij}\mathcal{U}(p) = -\dfrac{\omega_{ij}\theta_{ij}}{p_i p_j}$, thus the Hessian $\mathrm{D}^2\mathcal{U}(p)$ is given by

$$\begin{bmatrix} \displaystyle\sum_{j\neq 1}\frac{\omega_{1j}\theta_{1j}}{p_1^2}[1-\log\left(\frac{\pi_j p_1}{\pi_1 p_j}\right)] & -\dfrac{\omega_{12}\theta_{12}}{p_1p_2} & \cdots & -\dfrac{\omega_{1n}\theta_{1n}}{p_1p_n} \\ -\dfrac{\omega_{21}\theta_{21}}{p_2p_1} & \displaystyle\sum_{j\neq 2}\frac{\omega_{2j}\theta_{2j}}{p_2^2}[1-\log\left(\frac{\pi_j p_2}{\pi_2 p_j}\right)] & \cdots & -\dfrac{\omega_{2n}\theta_{2n}}{p_2p_n} \\ \vdots & \cdots & \ddots & \vdots \\ -\dfrac{\omega_{n1}\theta_{n1}}{p_np_1} & -\dfrac{\omega_{n2}\theta_{n2}}{p_np_2} & \cdots & \displaystyle\sum_{j\neq n}\frac{\omega_{nj}\theta_{nj}}{p_n^2}[1-\log\left(\frac{\pi_j p_n}{\pi_n p_j}\right)] \end{bmatrix}$$

In the asymptotical limit $|\frac{p_i}{\pi_i} - 1| \propto \mathcal{O}(10^{-k})$ for some $k \gg 1$, then

$$\log\frac{\pi_j p_i}{\pi_i p_j} \approx (c_i - c_j)10^{-k}; \qquad \frac{1}{p_i^2} \approx \frac{1}{\pi_i^2} - \frac{2c_i}{p_i^2}10^{-k}; \qquad \text{and} \qquad \frac{1}{p_i p_j} \approx \frac{1}{\pi_i \pi_j} - \frac{c_i + c_j}{\pi_i\pi_j}10^{-k}.$$

Thus the Hessian $\mathrm{D}^2\mathcal{U}(p)$ is approximated by

$$\begin{bmatrix} \frac{1}{\pi_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\pi_2} & \vdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \frac{1}{\pi_n} \end{bmatrix} \begin{bmatrix} \sum_{j\neq 1}\omega_{1j}\theta_{1j} & -\omega_{12}\theta_{12} & \cdots & -\omega_{1n}\theta_{1n} \\ -\omega_{21}\theta_{21} & \sum_{j\neq 2}\omega_{2j}\theta_{2j} & \vdots & -\omega_{2n}\theta_{2n} \\ \vdots & \cdots & \ddots & \vdots \\ -\omega_{n1}\theta_{n1} & \omega_{n2}\theta_{n2} & \cdots & \sum_{j\neq n}\omega_{nj}\theta_{nj} \end{bmatrix} \begin{bmatrix} \frac{1}{\pi_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\pi_2} & \vdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \frac{1}{\pi_n} \end{bmatrix}$$
$$= \mathrm{diag}(\pi^{-1})K\mathrm{diag}(\pi^{-1}).$$

Moreover, $\mathrm{D}^2\mathcal{U}(p)|_{p=\pi} = \mathrm{diag}(\pi^{-1})K\mathrm{diag}(\pi^{-1})$.

Let's go back to the Rayleigh quotient problem. As K is symmetric, and $\mathrm{Ker}(K) = \mathrm{span}(\mathbb{1}_n)$, the singular value decomposition (SVD) of K is given by

$$K = U\Sigma U^\top = \widehat{U}\widehat{\Sigma}\widehat{U}^\top.$$

Here $U \in \mathbb{R}^{n\times n}$ is an orthogonal matrix, $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_{n-1}, 0)$ with $\sigma_1 \geq \cdots \geq \sigma_{n-1} > 0$, $\widehat{U} \in \mathbb{R}^{n\times(n-1)}$ denotes the matrix formed by the first $n-1$ columns of U, and $\widehat{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_{n-1})$. Then, we can reformulate the Rayleigh quotient as

$$\min_{\psi\mathbb{1}_n^\top = 0} \frac{\psi\widehat{U}\widehat{\Sigma}\widehat{U}^\top D^2\mathcal{U}(p)\widehat{U}\widehat{\Sigma}\widehat{U}^\top\psi^\top}{\psi\widehat{U}\widehat{\Sigma}\widehat{U}^\top\psi^\top}$$

Denote $\varphi = \psi\widehat{U}\sqrt{\widehat{\Sigma}}$. As both $\widehat{U} \in \mathbb{R}^{n\times(n-1)}$, $\widehat{\Sigma} \in \mathbb{R}^{(n-1)\times(n-1)}$ are matrices with full rank, $\varphi \in \mathbb{R}^{n-1}$ explores the entire space $\mathbb{R}^{n-1}$ as ψ goes over $\mathbb{R}^n$. We can further rewrite it as

$$\min_{\varphi\in\mathbb{R}^{n-1}} \frac{\varphi\sqrt{\widehat{\Sigma}}\widehat{U}^\top D^2\mathcal{U}(p)\widehat{U}\sqrt{\widehat{\Sigma}}\varphi^\top}{\varphi\varphi^\top} = \lambda_{\min}(\sqrt{\widehat{\Sigma}}\widehat{U}^\top D^2\mathcal{U}(p)\widehat{U}\sqrt{\widehat{\Sigma}}). \tag{33}$$

Appendix D. Other proofs

Proof [Proof to Lemma 1] Recall

$$\frac{\partial \mathrm{I}(p\|\pi)}{\partial p_i} = \frac{1}{2}\sum_{j=1,j\neq i}^{n} \frac{\omega_{ij}}{\pi_i}(\log\frac{p_i}{\pi_i} - \log\frac{p_j}{\pi_j}) + \frac{\omega_{ij}}{p_i}(\frac{p_i}{\pi_i} - \frac{p_j}{\pi_j}).$$

It is straightforward to verify that when $p = \pi$, we have $\nabla_p \mathrm{I}(p\|\pi) = 0$.

On the other hand, suppose $\nabla_p \mathrm{I}(p\|\pi) = 0$ and assume $\frac{p_{i_1}}{\pi_{i_1}} \geqslant \frac{p_j}{\pi_j}$ for any $j \in V$. We will in the end show that $\frac{p_j}{\pi_j} = \frac{p_{i_1}}{\pi_{i_1}}$, thus $p = \pi$ in $\mathbb{P}(V)$.

For any $i \in V$, denote the neighbors of vertex i as $N(i) = \{j \in V \mid \omega_{ij} > 0\}$. Note that

$$0 = \frac{\partial \mathrm{I}(p\|\pi)}{\partial p_{i_1}} = \frac{1}{2}\sum_{j\in N(i_1)} \frac{\omega_{i_1 j}}{\pi_{i_1}}(\log\frac{p_{i_1}}{\pi_{i_1}} - \log\frac{p_j}{\pi_j}) + \frac{\omega_{i_1 j}}{p_{i_1}}(\frac{p_{i_1}}{\pi_{i_1}} - \frac{p_j}{\pi_j}) \geqslant 0,$$

as $\frac{p_{i_1}}{\pi_{i_i}} \geqslant \frac{p_j}{\pi_j}$ and $\omega_{i_1 j} > 0$. This implies that for any $j \in N(i_1)$, $\frac{p_j}{\pi_j} = \frac{p_{i_1}}{\pi_{i_i}}$.

As the graph $G = (V, E, \omega)$ is connected, for any node j, there exists a path $i_1 \to i_2 \to \cdots \to i_m = j$ for some $m \leqslant n$, such that $\omega_{i_k i_{k+1}} > 0$ for $k = 1, \ldots, m-1$. From the previous step, we have shown that $i_2 \in N(i_1)$ implies that $\frac{p_{i_2}}{\pi_{i_2}} = \frac{p_{i_1}}{\pi_{i_i}}$. Thus we can repeat this process for $i_2, \ldots, i_m$, and prove that

$$\frac{p_{i_1}}{\pi_{i_1}} = \frac{p_{i_2}}{\pi_{i_2}} = \cdots = \frac{p_j}{\pi_j}.$$

In conclusion, we have shown that $\frac{p_j}{\pi_j} = \frac{p_{i_1}}{\pi_{i_1}}$ for all $j \in V$. Since p is a probability function and must sum to one, it follows that $p = \pi$.

■

Proof [Proof of Theorem 2] Fix r_0, $F_{r_0}(r)$ is invertible on $(0, r_0]$, and we denote its inverse by $F_{r_0}^{-1}(C)$, whose domain is $[0, \infty)$ and range is $(0, r_0]$. Note that for $r \in (0, r_0]$ and $C \in (0, \infty)$,

$$F_{r_0}(r) \leqslant C, \qquad \text{if and only if} \qquad F_{r_0}^{-1}(C) \leqslant r. \tag{34}$$

Pick $C \in (0, \infty)$. Let us suppose $F_{r_0}^{-1}(C) = r$ for some $r \in (0, r_0)$, that is $F_{r_0}(r) = C$. For any $r_1 > r_0$, $F_{r_1}(r) > F_{r_0}(r) = C$ implies that $F_{r_1}^{-1}(C) > r = F_{r_0}^{-1}(C)$. Now for fixed C, define $G_C(r) = F_r^{-1}(C)$, thus $G_C(r)$ is a strictly increasing function, satisfying that $G_C(r) > 0$ as long as $C > 0$ and $r > 0$. And

$$G_C(r_0) = F_{r_0}^{-1}(C) \leqslant r, \qquad \text{if and only if} \qquad F_{r_0}(r) \leqslant C. \tag{35}$$

Given that the Hamiltonian $\mathcal{H}(p(t), \psi(t))$ is nonincreasing as shown in (20), and $\mathcal{H}(p(0), \psi(0))$ is bounded, let $r_i = \frac{p_i}{\pi_i}$, for any t,

$$\mathcal{U}(p(t)) = \frac{1}{4} \sum_{i,j=1}^{n} F_{r_i}(r_j)\omega_{ij} \leqslant \mathcal{H}(p(t), \psi(t)) \leqslant C,$$

which implies that $F_{r_i}(r_j)\omega_{ij} \leqslant C$ for any pair (i, j).

Assume $r_{i_1} \geqslant r_i$ for any $i \in V$. Note that there exists an index i such that $p_i \geqslant \frac{1}{n}$. Then

$$r_{i_1} = \max r_i \geqslant r_i = \frac{p_i}{\pi_i} \geqslant \frac{1}{n\pi_i} \geqslant \frac{1}{n \max \pi_i} > 0.$$

Since the graph G is connected, for any node j, there exists a path $i_1 \to i_2 \to \cdots \to i_m = j$ for some $m \leqslant n$, such that $\omega_{i_k i_{k+1}} > 0$ for $k = 1, \ldots, m-1$. Let $\bar{\omega} = \min_k \omega_{i_k i_{k+1}} > 0$. For $k = 1, \ldots, m-1$, $F_{r_{i_k}}(r_{i_{k+1}})\omega_{i_k i_{k+1}} \leqslant C$ implies that $F_{r_{i_k}}(r_{i_{k+1}}) \leqslant \frac{C}{\omega_{i_k i_{k+1}}} \leqslant \frac{C}{\bar{\omega}}$.

From node i_1 to i_2, given that $r_{i_1} \geqslant r_{i_2}$ and $F_{r_{i_1}}(r_{i_2}) \leqslant \frac{C}{\bar{\omega}}$, by (34), (35) and $G_C(r)$ is strictly increasing, we have

$$r_{i_2} \geqslant F_{r_{i_1}}^{-1}(\frac{C}{\bar{\omega}}) = G_{C/\bar{\omega}}(r_{i_1}) \geqslant G_{C/\bar{\omega}}(\frac{1}{n \max \pi_i}) > 0.$$

From node i_2 to i_3, if $r_{i_2} \geqslant r_{i_3}$, we can repeat the above step and yield $r_{i_3} \geqslant G_{C/\bar{\omega}}(r_{i_2})$; otherwise $r_{i_3} > r_{i_2} \geqslant G_{C/\bar{\omega}}(r_{i_1}) \geqslant G_{C/\bar{\omega}}(r_{i_2})$. In either case,

$$r_{i_3} \geqslant G_{C/\bar{\omega}}(r_{i_2}) \geqslant G_{C/\bar{\omega}}(G_{C/\bar{\omega}}(r_{i_1})) \geqslant G^{(2)}_{C/\bar{\omega}}(\frac{1}{n \max \pi_i}) > 0.$$

We may repeat this argument and yield that $r_{i_k} > 0$ for any node i_k on this path,

$$r_{i_k} \geqslant G^{(k-1)}_{C/\bar{\omega}}(\frac{1}{n \max \pi_i}) > 0.$$

Thus $\varepsilon = (\min \pi_i) \cdot G^{(n-1)}_{C/\bar{\omega}}(\frac{1}{n \max \pi_i}) > 0$ is the lower bound of $p_i(t)$. ■

Proof [Proof to Lemma 7] Follow the notations used in Appendix C, Recall (33) and

$$\mathrm{D}^2\mathcal{U}(p)|_{p=\pi} = \mathrm{diag}(\pi^{-1}) K \mathrm{diag}(\pi^{-1}) = \mathrm{diag}(\pi)^{-1} \widehat{U} \widehat{\Sigma} \widehat{U}^\top \mathrm{diag}(\pi)^{-1}.$$

The minimal positive eigenvalue λ at $p = \pi$ is

$$\begin{aligned}\lambda &= \lambda_{\min}(\sqrt{\widehat{\Sigma}}\, \underbrace{\widehat{U}^\top \mathrm{diag}(\pi)^{-1} \widehat{U}}_{S}\, \widehat{\Sigma}\, \underbrace{\widehat{U}^\top \mathrm{diag}(\pi)^{-1} \widehat{U}}_{S^\top}\, \sqrt{\widehat{\Sigma}}) \\ &= \lambda_{\min}((S^\top \sqrt{\widehat{\Sigma}})^\top \widehat{\Sigma} (S^\top \sqrt{\widehat{\Sigma}})).\end{aligned}$$

Since the non-singular conjugate of positive definite matrix is still positive definite, we can show that $\lambda > 0$ as long as $S^\top \sqrt{\widehat{\Sigma}}$ is non-singular.

In fact, suppose for $x \in \mathbb{R}^{n-1}$, $S^\top \sqrt{\widehat{\Sigma}} x^\top = 0$, thus $(\sqrt{\widehat{\Sigma}} x^\top)^\top S^\top \sqrt{\widehat{\Sigma}} x^\top = 0$. From the change of variable $y^\top = \widehat{U} \sqrt{\widehat{\Sigma}} x^\top \in \mathbb{R}^n$, we obtain $y \mathrm{diag}(\pi)^{-1} y^\top = 0$. This leads to $y = 0$, i.e., $\widehat{U} \sqrt{\widehat{\Sigma}} x^\top = 0$. Now since $\widehat{U} \in \mathbb{R}^{n \times (n-1)}$ is full rank, $\sqrt{\widehat{\Sigma}}$ is non-singular, we have $x = 0$. As a result, $S^\top \sqrt{\Sigma}$ is non-singular. ■

References

F. Alimisis, A. Orvieto, G. Bécigneul, and A. Lucchi. A continuous-time perspective for modeling acceleration in Riemannian optimization. In International Conference on Artificial Intelligence and Statistics, pages 1297–1307. PMLR, 2020.

L. Ambrosio, N. Gigli, and G. Savaré. Gradient flows in metric spaces and in the space of probability measures. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008. ISBN 978-3-7643-8721-1.

J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. Numer. Math., 84(3):375–393, 2000. ISSN 0029-599X,0945-3245. doi: 10.1007/s002110050002.

S. Chen, Q. Li, O. Tse, and S. J. Wright. Accelerating optimization over the space of probability measures. Journal of Machine Learning Research, 26(31):1–40, 2025. URL http://jmlr.org/papers/v26/23-1288.html.

X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In Proceedings of the 31st Conference On Learning Theory, volume 75 of Proceedings of Machine Learning Research, pages 300–323, 2018. URL https://proceedings.mlr.press/v75/cheng18a.html.

S.-N. Chow, W. Huang, Y. Li, and H. Zhou. Fokker-Planck equations for a free energy functional or Markov process on a graph. Arch. Ration. Mech. Anal., 203(3):969–1008, 2012. ISSN 0003-9527,1432-0673. doi: 10.1007/s00205-011-0471-6.

M. Erbar and J. Maas. Ricci curvature of finite Markov chains via convexity of the entropy. Arch. Ration. Mech. Anal., 206(3):997–1038, 2012. ISSN 0003-9527,1432-0673. doi: 10.1007/s00205-012-0554-z.

Y. Gao, J.-G. Liu, and W. Li. Master equations for finite state mean field games with nonlinear activations. Discrete Contin. Dyn. Syst. Ser. B, 29(7):2837–2879, 2024. ISSN 1531-3492,1553-524X. doi: 10.3934/dcdsb.2023204.

C. J. Geyer. Introduction to Markov chain Monte Carlo. In Handbook of Markov chain Monte Carlo, Chapman & Hall/CRC Handb. Mod. Stat. Methods, pages 3–48. CRC Press, Boca Raton, FL, 2011. ISBN 978-1-4200-7941-8.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57(1):97–109, 1970. ISSN 0006-3444,1464-3510. doi: 10.1093/biomet/57.1.97.

G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. Neural Comput., 18(7):1527–1554, 2006. ISSN 0899-7667,1530-888X. doi: 10.1162/neco.2006.18.7.1527.

P. Jäckel. Monte Carlo methods in finance. John Wiley & Sons, 2002.

E. Jacquier, N. G. Polson, and P. E. Rossi. Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. J. Econometrics, 122(1):185–212, 2004. ISSN 0304-4076,1872-6895. doi: 10.1016/j.jeconom.2003.09.001.

M. Jerrum and A. Sinclair. The Markov chain Monte Carlo method: an approach to approximate counting and integration. In Approximation Algorithms for NP-Hard Problems, page 482–520. PWS Publishing Co., 1996. ISBN 0534949681.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In Learning in Graphical Models, page 105–161. MIT Press, Cambridge, MA, USA, 1999. ISBN 0262600323.

I. Karatzas, J. Maas, and W. Schachermayer. Trajectorial dissipation and gradient flow for the relative entropy in Markov chains. Commun. Inf. Syst., 21(4):481–536, 2021. ISSN 1526-7555,2163-4548. doi: 10.4310/CIS.2021.v21.n4.a1.

S. G. Kou and H. Wang. Option Pricing under a Double Exponential Jump Diffusion Model. Management Science, 50(9):1178–1192, 2004. ISSN 00251909, 15265501. URL http://www.jstor.org/stable/30046226.

D. P. Landau and K. Binder. A guide to Monte Carlo simulations in statistical physics. Cambridge University Press, Cambridge, fourth edition, 2015. ISBN 978-1-107-07402-6. doi: 10.1017/CBO9781139696463.

J. M. Lee. Introduction to Riemannian manifolds, volume 176 of Graduate Texts in Mathematics. Springer, Cham, second edition, 2018. ISBN 978-3-319-91754-2; 978-3-319-91755-9.

D. A. Levin, Y. Peres, and E. L. Wilmer. Markov chains and mixing times. American Mathematical Society, Providence, RI, 2009. ISBN 978-0-8218-4739-8. doi: 10.1090/mbk/058.

R. Li, H. Zha, and M. Tao. Hessian-Free High-Resolution Nesterov Acceleration For Sampling. In Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pages 13125–13162, 2022. URL https://proceedings.mlr.press/v162/li22z.html.

W. Li. Geometric calculations on probability manifolds from reciprocal relations in Master equations, 2025. URL https://arxiv.org/abs/2504.19368.

A. Lou, C. Meng, and S. Ermon. Discrete Diffusion Modeling by Estimating the Ratios of the Data Distribution, 2024. URL https://arxiv.org/abs/2310.16834.

Y.-A. Ma, N. S. Chatterji, X. Cheng, N. Flammarion, P. L. Bartlett, and M. I. Jordan. Is there an analog of Nesterov acceleration for gradient-based MCMC? Bernoulli, 27(3): 1942 – 1992, 2021. doi: 10.3150/20-BEJ1297.

J. Maas. Gradient flows of the entropy for finite Markov chains. J. Funct. Anal., 261(8): 2250–2292, 2011. ISSN 0022-1236,1096-0783. doi: 10.1016/j.jfa.2011.06.009.

C. J. Maddison, D. Paulin, Y. W. Teh, B. O'Donoghue, and A. Doucet. Hamiltonian Descent Methods, 2018. URL https://arxiv.org/abs/1809.05042.

K. L. Mengersen and R. L. Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. Ann. Statist., 24(1):101–121, 1996. ISSN 0090-5364,2168-8966. doi: 10.1214/aos/1033066201.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. The Journal of Chemical Physics, 21(6):1087–1092, 06 1953. ISSN 0021-9606. doi: 10.1063/1.1699114.

A. Mielke. Geodesic convexity of the relative entropy in reversible Markov chains. Calc. Var. Partial Differential Equations, 48(1-2):1–31, 2013. ISSN 0944-2669,1432-0835. doi: 10.1007/s00526-012-0538-8.

R. M. Neal. MCMC using Hamiltonian dynamics. In Handbook of Markov chain Monte Carlo, Chapman & Hall/CRC Handb. Mod. Stat. Methods, pages 113–162. CRC Press, Boca Raton, FL, 2011. ISBN 978-1-4200-7941-8.

Y. Nesterov. Introductory lectures on convex optimization: A basic course, volume 87 of Applied Optimization. Kluwer Academic Publishers, Boston, MA, 2004. ISBN 1-4020-7553-7. doi: 10.1007/978-1-4419-8853-9.

Y. E. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. Dokl. Akad. Nauk SSSR, 269(no. 3,):543–547, 1983. ISSN 0002-3264.

F. Otto. The geometry of dissipative evolution equations: the porous medium equation. Comm. Partial Differential Equations, 26(1-2):101–174, 2001. ISSN 0360-5302,1532-4133. doi: 10.1081/PDE-100002243.

M. Ottobre. Markov chain Monte Carlo and irreversibility. Rep. Math. Phys., 77(3):267–292, 2016. ISSN 0034-4877,1879-0674. doi: 10.1016/S0034-4877(16)30031-3.

C. P. Robert and G. Casella. Monte Carlo statistical methods. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2004. ISBN 0-387-21239-6. doi: 10.1007/978-1-4757-4145-2.

G. O. Roberts and R. L. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. Biometrika, 83(1):95–110, 1996a. ISSN 0006-3444. doi: 10.1093/biomet/83.1.95.

G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. Bernoulli, 2(4):341–363, 1996b. ISSN 1350-7265. doi: 10.2307/3318418.

R. Salakhutdinov and G. Hinton. Deep Boltzmann Machines. In Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, volume 5 of Proceedings of Machine Learning Research, pages 448–455, 2009. URL `https://proceedings.mlr.press/v5/salakhutdinov09a.html`.

J. Schnakenberg. Network theory of microscopic and macroscopic behavior of master equation systems. Rev. Modern Phys., 48(4):571–585, 1976. ISSN 0034-6861,1539-0756. doi: 10.1103/RevModPhys.48.571.

W. Su, S. Boyd, and E. J. Candès. A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights. Journal of Machine Learning Research, 17(153):1–43, 2016. URL `http://jmlr.org/papers/v17/15-084.html`.

A. Taghvaei and P. Mehta. Accelerated Flow for Probability Distributions. In Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 6076–6085, 2019. URL `https://proceedings.mlr.press/v97/taghvaei19a.html`.

Y. W. Teh, M. Welling, S. Osindero, and G. E. Hinton. Energy-based models for sparse overcomplete representations. Journal of Machine Learning Research, 4(7-8):1235–1260, 2004. ISSN 1532-4435,1533-7928. doi: 10.1162/jmlr.2003.4.7-8.1235.

C. Villani. Optimal transport: Old and new, volume 338. Springer-Verlag, Berlin, 2009. ISBN 978-3-540-71049-3. doi: 10.1007/978-3-540-71050-9.

Y. Wang and W. Li. Accelerated information gradient flow. J. Sci. Comput., 90(1):Paper No. 11, 47, 2022. ISSN 0885-7474,1573-7691. doi: 10.1007/s10915-021-01709-3.

A. Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In Proceedings of the 31st Conference On Learning Theory, volume 75 of Proceedings of Machine Learning Research, pages 2093–3027, 2018. URL https://proceedings.mlr.press/v75/wibisono18a.html.

H. Zhang and S. Sra. Towards Riemannian Accelerated Gradient Methods, 2018. URL https://arxiv.org/abs/1806.02812.

X. Zuo, S. Osher, and W. Li. Gradient-adjusted underdamped Langevin dynamics for sampling, 2024. URL https://arxiv.org/abs/2410.08987.