Accelerated Markov Chain Monte Carlo Algorithms on Discrete States

Bohan Zhou[1], Shu Liu[2], Xinzhe Zuo[2], and Wuchen Li[3]

[1]UCSB Mathematics. bhzhou@ucsb.edu; [2]UCLA Mathematics. {shuliu,zxz}@math.ucla.edu;
[3]U of SC Mathematics. wuchen@mailbox.sc.edu.

Overview

- We propose a class of accelerated Markov chain Monte Carlo (aMCMC) algorithms for sampling from discrete-state spaces. This framework is inspired by the Metropolis-Hastings algorithm, the graphical Wasserstein metric, and Nesterov's accelerated gradient method.
- While MH can be viewed as a gradient descent of the KL divergence, our approach introduces a momentum-based acceleration via a damped Hamiltonian system, with user-defined potentials and mobilities.
- The accelerated gradient flow of the relative Fisher information demonstrates (acceleration and accuracy) of the algorithm, without requiring the normalizing constant while preserving positivity of probabilities.

Route Map

Discrete-time		Continuous-time
$\mathbb{P}(X^{(k+1)} = j \mid X^{(k)} = i) = P_{ij}$		$\mathbb{P}(X(t+h) = j \mid X(t) = i) \approx \delta_{ij} + Q_{ij}h$
$\downarrow$		
$p^{(k+1)} = p^{(k)}P$	$\longleftrightarrow$	$\dot{p}(t) = pQ$
		$\downarrow$
$p^{(k+1)} = p^{(k)}(I_n + Q\Delta t)$		$\dot{p}(t) = -\nabla_p D_\varphi(p\Vert\pi)\mathbb{K}(p)$
		$\downarrow$
$p^{(k+1)} = p^{(k)}(I_n + \bar{Q}^r_\psi \Delta t)$	$\longleftarrow$	$\begin{bmatrix}\dot{p}(t)\\ \dot{\psi}(t)\end{bmatrix} = \begin{bmatrix}0\\ -\gamma(t)\psi(t)\end{bmatrix} + \begin{bmatrix}0 & I\\ -I & 0\end{bmatrix}\begin{bmatrix}\partial_p\mathcal{H}\\ \partial_\psi\mathcal{H}\end{bmatrix}$

Table 1: We start with the Markov chain (first row), which can be rewrite as the forward master equation (second row). Using the graphical Wasserstein metric, this equation takes the form of a gradient flow (the third row). Finally, we introduce the damped Hamiltonian dynamics (the last row) originated from Nesterov's accelerated gradient method, along with its corresponding jump process.

Sampling on Discrete-State Spaces

We aim to sample from a target distribution π supported on a given graph by designing a dynamical (or jump) process such that the state variable $p(t)$ (or $p^{(k)}$) converges to π over time.

Input: Initial distribution $\rho^{(0)}$, total particles M, unnormalized target distribution π.

User-specified choices:

i. Baseline transition rate matrix Q that satisfies the detailed balance, e.g., the one constructed by the Metropolis-Hastings (MH) algorithm.

ii. Activation function $\theta_{ij}(p)$ that induces a graphical metric tensor $\mathbb{K}^\dagger$ on the graph; e.g., the logarithmic mean $\theta_{ij}(p) = \frac{\frac{p_i}{\pi_i} - \frac{p_j}{\pi_j}}{\log\frac{p_i}{\pi_i} - \log\frac{p_j}{\pi_j}} = \frac{p_i}{\pi_i}\cdot\frac{1-\frac{\pi_i}{\pi_j}\frac{p_j}{p_i}}{\log\left(\frac{\pi_j}{\pi_i}\frac{p_i}{p_j}\right)}$.

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

iv. Damping parameter $\gamma(t) > 0$, which may be either time-dependent or constant.

Accelerated MCMC

We select Q^{MH} (specify i.), define $\omega_{ij} = \pi_i Q^{\mathrm{MH}}_{ij}$ and expand the matrix form as

$$\begin{cases} \frac{\mathrm{d}p_i}{\mathrm{d}t} + \sum_{j\neq i}\omega_{ij}\theta_{ij}(p)(\psi_j - \psi_i) = 0, & (0.1\mathrm{a})\\ \frac{\mathrm{d}\psi_i}{\mathrm{d}t} + \gamma(t)\psi_i + \frac{1}{2}\sum_{j\neq i}\omega_{ij}\frac{\partial\theta_{ij}(p)}{\partial p_i}(\psi_i - \psi_j)^2 + \frac{\partial\mathcal{U}(p)}{\partial p_i} = 0, & (0.1\mathrm{b})\end{cases}$$

The jump process for (0.1a) can be constructed as

$$\frac{\mathrm{d}p_i}{\mathrm{d}t} = -\left[\sum_{j\neq i}\frac{\omega_{ij}\theta_{ij}(p)(\psi_i - \psi_j)_-}{p_i}\right]p_i + \sum_{j\neq i}\frac{\omega_{ji}\theta_{ji}(p)(\psi_i - \psi_j)_+}{p_j}p_j,$$

which leads to the form of forward master equation $\frac{\mathrm{d}}{dt}p = p\bar{Q}^r_\psi$ and requires positivity.

$\theta_{ij}(p)$ (specify ii.)	potential $\mathcal{U}(p)$ (specify iii.)	w/o Z	strict positivity
1	$\frac{1}{2}\sum_{i=1}^n \frac{(p_i-\pi_i)^2}{\pi_i}$	No	No
log-mean	$\sum_{i=1}^n p_i \log\frac{p_i}{\pi_i}$.	Yes	No
log-mean	$\frac{1}{4}\sum_{i,j=1}^n \omega_{ij}(\log\frac{\pi_j}{\pi_i}\frac{p_i}{p_j})(\frac{p_i}{\pi_i} - \frac{p_j}{\pi_j})$	Yes	Yes
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. First row is **Chi-squared** method; second row is **KL** method; third row is **log-Fisher** method; fourth row is **con-Fisher** method.

Analysis of aMCMC

- Convergence: If π is the unique critical point to $\mathcal{U}(p)$, then $p(t)$ converges to π.
- Normalizing constant Z of target distribution π: **KL** and **log-Fisher** do not depend on Z, provided that damping parameter $\gamma(t)$ does not depend on Z.
- Positivity of state variables: A large class of potential functions (include **log-Fisher** and **con-Fisher**) ensure there is a positive lower bound ε such that $p_i(t) > \varepsilon$ for any i and any t.
- Damping parameter and acceleration
  - **Chi-squared**: Let α_* be the largest negative eigenvalue of Q. If $|\alpha_*| < 1$, then there exists damping parameter $\gamma(t) = d \in [2\sqrt{|\alpha_*|}, |\alpha_*| + 1)$ (specify iv.), such that the largest negative eigenvalue μ_* of L satisfies $\mu_* < \alpha_*$.
  - **Log-Fisher**: the damping parameter $\gamma(t)$ in the asymptotical limit can be suggested by **con-Fisher**, via computing a Rayleigh quotient problem.

Computational Remark

- The staggered scheme with splitting method is employed.
- MH steps are triggered as a restart mechanism to restore strict positivity when it is compromised by accumulated sampling errors.
- Acceleration and Accuracy via **Chi-squared** and **log-Fisher** method are observed in numerical examples, comparing with MH method.

Sampling on hypercube and lattices

We seek to sampling $\pi = \frac{1}{Z}[16, 1, \ldots, 1, \ldots, 1, 16]$ on hypercube of 64 nodes,

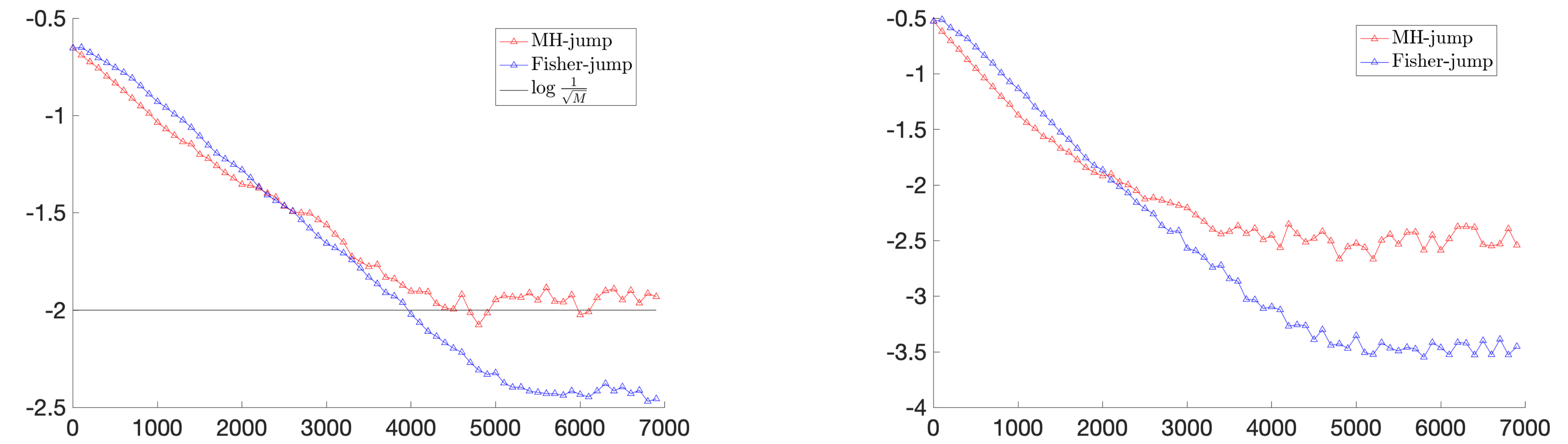

Figure 1: Sampling on a hypercube graph of 64 nodes via **log-Fisher** method. x-axes are in iterations with step size $\Delta t = 0.01$. The left figure shows the approximation error $\log_{10}\|p(t) - \pi\|_2$ w.r.t π. The right figure shows the approximation error $\log_{10}|\sum_{i=1}^n p_i(t)\log\frac{p_i(t)}{Z\pi}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

and sampling the mixture of two Gaussians on a lattice of 625 nodes:

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

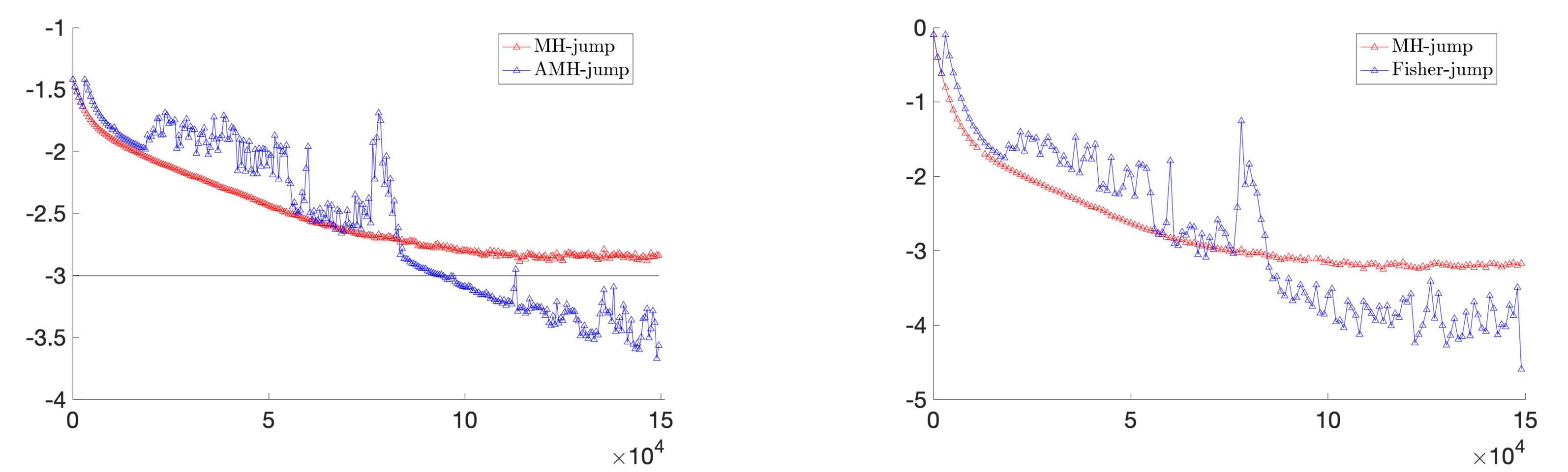

Figure 2: Sampling on a 25×25 lattice graph via **log-Fisher**. x-axes are in iterations. The left figure shows the approximation error $\log_{10}\|p(t) - \pi\|_2$ w.r.t π. The right figure shows the approximation error $\log_{10}|\sum_{i=1}^n p_i(t)\log\frac{p_i(t)}{Z\pi} - (-\log Z)|$ w.r.t Z. The jump process via `log-Fisher` achieves to a higher accuracy when that via MH is approaching to $\mathcal{O}(\frac{1}{\sqrt{M}})$.

References

[1] 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.

[2] S.-N. Chow, W. Huang, Y. Li, and H. Zhou. Fokker–Planck Equations for a Free Energy Functional or Markov Process on a Graph. *Archive for Rational Mechanics and Analysis*, 203(3):969–1008, 2012.

[3] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

[4] J. Maas. Gradient flows of the entropy for finite Markov chains. *Journal of Functional Analysis*, 261(8):2250–2292, 2011.

[5] A. Mielke. A gradient structure for reaction–diffusion systems and for energy-drift-diffusion systems. *Nonlinearity*, 24(4):1329–1346, 2011.

[6] Y. E. Nesterov. A method of solving a convex programming problem with convergence rate $\mathcal{O}(\frac{1}{k^2})$. In *Doklady Akademii Nauk*, volume 269, pages 543–547. Russian Academy of Sciences, 1983.

[7] Y. Wang and W. Li. Accelerated information gradient flow. *Journal of Scientific Computing*, 90:1–47, 2022.

This project has received funding from 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 in 2023.