Discrete Langevin Samplers via Wasserstein Gradient Flow

Haoran Sun[1]
hsun349@gatech.edu
Georgia Tech

Hanjun Dai
hadai@google.com
Google Research

Bo Dai
bodai@google.com
Google Research & Georgia Tech

Haomin Zhou
hmzhou@math.gatech.edu
Georgia Tech

Dale Schuurmans
schuurmans@google.com
Google Research & University of Alberta

Abstract

It is known that gradient based MCMC samplers for continuous spaces, such as Langevin Monte Carlo (LMC), can be derived as particle versions of a gradient flow that minimizes KL divergence on a Wasserstein manifold. The superior efficiency of such samplers has motivated several recent attempts to generalize LMC to discrete spaces. However, a fully principled extension of Langevin dynamics to discrete spaces has yet to be achieved, due to the lack of well-defined gradients in the sample space. In this work, we show how the Wasserstein gradient flow can be generalized naturally to discrete spaces. Given the proposed formulation, we demonstrate how a discrete analogue of Langevin dynamics can subsequently be developed. With this new understanding, we reveal how recent gradient based samplers in discrete spaces can be obtained as special cases by choosing particular discretizations. More importantly, the framework also allows for the derivation of novel algorithms, one of which, *Discrete Langevin Monte Carlo* (DLMC), is obtained by a factorized estimate of the transition matrix. The DLMC method admits a convenient parallel implementation and time-uniform sampling that achieves larger jump distances. We demonstrate the advantages of DLMC on various binary and categorical distributions.

Proceedings of the 26[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

1 INTRODUCTION

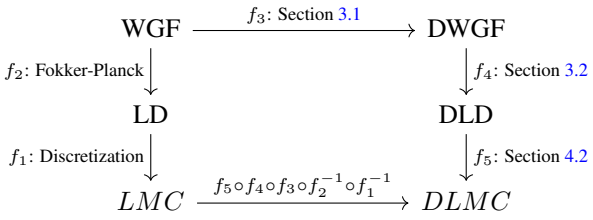
The Markov Chain Monte Carlo (MCMC) algorithm is one of the most widely used methods for sampling from intractable distributions (Robert & Casella, 2013). However, it is known to mix slowly in complex, high-dimensional models. In response, several gradient based MCMC methods have been developed over the past decades that leverage gradient information to guide proposals toward high probability regions (Neal et al., 2011). By simulating Langevin dynamics (LD), the Langevin Monte Carlo method (LMC) (Rossky et al., 1978) and its variants (Welling & Teh, 2011; Girolami & Calderhead, 2011) have substantially improved sampling efficiency in both theory and practice. In seminal work, Jordan et al. (1998) and Otto (2001) have shown that Langevin dynamics simulate $\frac{d}{dt}\rho^t = -\nabla_\rho D_{\mathrm{KL}}(\rho^t||\pi)$, which is a Wasserstein gradient flow (WGF) that minimizes the KL-divergence to a target distribution π. This connection not only provides a tool for algorithm design (Ma et al., 2015; Liu et al., 2019) but also help theoretical analysis (Cheng & Bartlett, 2018).

Despite these advances, progress in gradient based methods has generally focused on continuous spaces. Recently, a family of locally balanced (LB) samplers (Zanella, 2020; Grathwohl et al., 2021; Sun et al., 2021a; Zhang et al., 2022; Sun et al., 2022; Rhodes & Gutmann, 2022) have leveraged gradient information for proposals in discrete spaces via LB functions, achieving significant success. However even though Zanella (2020) and Sun et al. (2021a) have proved that LB functions are asymptotically optimal for leveraging gradient information in a proposal distribution, a principled extension of LMC from continuous to discrete spaces remains lacking in finite dimensional problems. Consequently, existing LB samplers suffer from inefficiencies arising from a suboptimal imitation of LMC. For example, Sun et al. (2021a, 2022) flip multiple sites in

[1]Work done during an internship at Google.

order, which prevents parallel implementation; Zhang et al. (2022) and Rhodes & Gutmann (2022) restrict a Gaussian proposal to discrete states, which ignores the difference between continuous diffusion and discrete jump processes.

To migrate LMC from continuous to discrete spaces, we consider an alternative perspective that lifts the commonly used particle level to a more principled distribution level view. In particular, we start from the fact that every stochastic process X^t has a corresponding probability density $\rho^t(x)$ in the Wasserstein manifold (Villani, 2009). Instead of designing jump processes as discrete analogues of the diffusion process in LMC, we instead consider the *Discrete Wasserstein Gradient Flow* (DWGF) ρ^t that minimizes the KL-divergence $D_{\mathrm{KL}}(\rho^t||\pi)$ to the target distribution π. We then derive a *Discrete Langevin Dynamics* (DLD) X^t as a particle realization of this gradient flow ρ^t. Unsurprisingly, previous LB samplers can be interpreted as different discretizations of X^t, which explains their success in finite dimensional problems. However, the key benefit of this alternative perspective is the development of new algorithms that more faithfully follow the derivation of LMC from LD in continuous spaces. Using a more efficient discretization of DLD, we develop a novel sampler, *Discrete Langevin Monte Carlo* (DLMC), that factorizes X^t into sub-processes X_n^t, where for each sub-process, the transition P_n^h after a simulation time h is estimated and a new state proposed according to $P_n^h(X_n^t, \cdot)$. In this way, DLMC is (i) computationally more efficient than previous methods, as it decouples dimensions and permits convenient parallel implementation, and (ii) statistically more efficient, since it uses a time-uniform discretization of DLD. The main ideas behind the overall framework can be summarized in the following diagram that relates the derivation of LMC from LD and WGF to their discrete counterparts.



An experimental evaluation demonstrates that DLMC enjoys better proposal quality and greater efficiency than traditional samplers as well as other LB samplers. These advantages are demonstrated both in sampling and learning tasks involving binary and categorical distributions, including the Bernoulli distribution, Ising model, Potts model, factorial hidden Markov model, restricted Boltzmann machine, and deep energy based models (EBMs). Code for reproducing all the experiments can be found at https://github.com/google-research/google-research/tree/master/dwgf.

2 PRELIMINARIES

We first revisit the framework of WGF $\rightarrow$ LD $\rightarrow$ LMC in the above diagram, which derives LMC as a discretization of LD that minimizes the KL-divergence to a target distribution π.

Wasserstein Gradient Flow The Wasserstein manifold

$$\mathcal{P}_2(\mathbb{R}^d) = \{\mu : \mathbb{R}^d \rightarrow \mathbb{R}_{\geq 0} : \int_{\mathbb{R}^d} \mu(x)dx = 1\} \tag{1}$$

is the set of probability measures on $\mathbb{R}^d$ where we define the distance between two measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ by the Wasserstein-2 distance (Villani, 2009):

$$W_2(\mu,\nu) = \left(\inf_{\Pi} \int d^2(x,y)d\Pi(x,y)\right)^{\frac{1}{2}}, \tag{2}$$

$$\text{s.t.} \int \Pi(x,y)dy = \mu(x), \int \Pi(x,y)dx = \nu(y), \tag{3}$$

where $d(x,y)$ is the distance on the underlying Euclidean space $\mathbb{R}^d$, and Π is a joint distribution satisfying the marginal constraints (3). For a target distribution π and a current distribution ρ on $\mathcal{P}_2(\mathbb{R}^d)$, the gradient of the KL-divergence with respect to ρ is:

$$\nabla_\rho D_{\mathrm{KL}}(\rho||\pi) = \nabla \cdot [\log \pi(x)\rho(x)] - \Delta\rho(x). \tag{4}$$

A flow ρ_t on $\mathcal{P}_2(V)$ that satisfies

$$\frac{\partial}{\partial t}\rho^t(x) = -\nabla \cdot [\log \pi(x)\rho^t(x)] + \Delta\rho^t(x), \tag{5}$$

is the *Wasserstein gradient flow* (WGF) that minimizes the KL-divergence to the target distribution π.

Langevin Dynamics Jordan et al. (1998) and Otto (2001) have established the elegant connection between the WGF (4) and its particle level realization via the Fokker-Planck Equation. In particular, assume X^t is a time dependent random variable that has movement described by the following stochastic differential equation:

$$dX^t = \nabla \log \pi(X^t)dt + \sqrt{2}dW^t, \tag{6}$$

where $W^t \in \mathbb{R}^d$ is a Wiener process. Then the Fokker-Planck equation asserts that ρ^t, the distribution of X^t, evolves over time in a way that satisfies (5). The stochastic process in (6) is called *Langevin Dynamics* (LD).

Langevin Monte Carlo Since LD is a particle realization of the gradient flow (5), one can use it to efficiently generate samples from the target distribution π via discrete time simulation:

$$X^{t+\epsilon} = X^t + \epsilon\nabla \log \pi(X^t) + \sqrt{2\epsilon}\xi, \tag{7}$$

where $\xi \sim \mathcal{N}(0, I_d)$ is a standard normally distributed random variable. In practice, the discrete time simulation in (7) has an approximation error, so a Metropolis-Hastings acceptance test (MH) is commonly used to correct any bias (Metropolis et al., 1953; Hastings, 1970). Specifically, given the current state x, proposal distribution $Q(x, \cdot)$ and new state y, the MH test accepts y with probability

$$\min\{1, \pi(y)Q(y, x)/\pi(x)Q(x, y)\} \tag{8}$$

to guarantee the Markov chain is π-reversible. Rewriting the simulation (7) as a Gaussian proposal distribution

$$x_{t+\epsilon} \sim Q(x, \cdot) = \mathcal{N}(\cdot; x + \epsilon \nabla \log \pi(x), 2\epsilon I), \tag{9}$$

one obtains the *Langevin Monte Carlo* (LMC) sampling algorithm for continuous spaces.

3 DISCRETE LANGEVIN FRAMEWORK

Given a finite set $V = \{1, 2, ..., M\}$, a distribution on V is an M-dimensional vector. These vectors form a manifold as the set of $M - 1$ dimensional simplex:

$$\mathcal{P}(V) = \big\{\rho \in \mathbb{R}^M : \sum_{i=1}^{M} \rho_i = 1, \rho_i \geq 0\big\}. \tag{10}$$

For a point $\rho \in \mathcal{P}(V)$, the associated tangent space at ρ (Do Carmo & Flaherty Francis, 1992) is

$$T_\rho \mathcal{P}(V) = \big\{\sigma \in \mathbb{R}^M : \sum_{i=1}^{M} \sigma_i = 0\big\}. \tag{11}$$

We assume the target distribution $\pi \in \mathcal{P}(V)$ is determined by an energy function f, such that

$$\pi_i = \exp(-f_i)/\sum_{k \in V} \exp(-f_k). \tag{12}$$

Each state $x \in V$ corresponds to a distribution $\rho^0 \in \mathcal{P}(V)$ as a one-hot vector with the k-th site equal to 1. To find efficient MCMC algorithms for drawing samples x_t from a target π, we first consider the gradient flow ρ_t that minimizes the KL-divergence $D_{\text{KL}}(\rho^t||\pi)$.

3.1 Discrete Wasserstein Gradient Flow

The gradient flow depends on both the loss function $D_{\text{KL}}(\rho_t||\pi)$ and the metric in the space. To established the Wasserstein distance in $\mathcal{P}(V)$, we follow Chow et al. (2012, 2017) to rewrite the Wasserstein distance in the language of fluid dynamics via Benamou-Brenier formula (Benamou & Brenier, 2000):

$$W_2^2(\rho^0, \rho^1) := \inf_v \left\{ \int_0^1 \langle v^t, v^t \rangle_{\rho^t} dt : \frac{d\rho^t}{dt} = -\nabla \cdot (\rho^t v^t) \right\}, \tag{13}$$

where $v \in \mathbb{R}^d \to \mathbb{R}^d$ is a vector field on $\mathbb{R}^d$, and $\langle v, v \rangle_\rho = \frac{1}{2}\int \langle v(x), v(x) \rangle \rho(x) dx$ is the total kinetic energy.

On $\mathcal{P}(V)$, we have a natural generalization of the vector field $v : V \to \mathbb{R}^M$, where $v_i = (v_{ij})_{j=1}^M$ characterize the amount of the transportation from node i to node j. For the divergence and inner product, instead of using the canonical form, we first introduce the conductance

$$c_{ij}(\rho) = c_{ji}(\rho) \geq 0 \tag{14}$$

between two nodes $i, j \in V$, depending on the current distribution ρ, to characterize the conductivity. Then we define the divergence of a vector field v as:

$$\text{div}_\rho(v) := -\Big(\sum_{i \neq j} c_{ij}(\rho) v_{ji}\Big)_{j=1}^M \in T_\rho \mathcal{P}(V), \tag{15}$$

and inner product between two vector fields u, v as:

$$\langle u, v \rangle_\rho := \frac{1}{2} \sum_{i,j} c_{ij}(\rho) u_{ij} v_{ij}. \tag{16}$$

Such a divergence $\text{div}_\rho(\cdot)$ and inner product $\langle \cdot, \cdot \rangle_\rho$ induce a 2-Wasserstein distance $W_2(\rho, \nu)$ via the Benamou-Brenier formula in (13) and make the manifold a Riemannian manifold $\mathcal{P}_2(V)$ (Chow et al., 2012).

$$W_2^2(\rho^0, \rho^1) := \inf_v \left\{ \int_0^1 \langle v^t, v^t \rangle_{\rho^t} dt : \frac{d\rho^t}{dt} = \text{div}_{\rho^t}(v^t) \right\}. \tag{17}$$

Now, we can characterize the Wasserstein gradient flow in Theorem 3.1; see Appendix A for a complete proof.

Theorem 3.1. *On $\mathcal{P}_2(V)$, the gradient flow that minimizes the KL-divergence $D_{KL}(\rho^t||\pi)$ is:*

$$\frac{d\rho^t}{dt} = \Big(\sum_{i \neq j} c_{ij}(\rho^t)(f_i + \log \rho_i^t - f_j - \log \rho_j^t)\Big)_{j=1}^M. \tag{18}$$

3.2 Discrete Langevin Dynamics

Equation (18) gives a discrete analogue of the Wasserstein gradient flow (5). Naturally, the discrete analogue of the Langevin dynamics should be a Markov jump process, as a particle level realization of (18). Denote such a Markov jump process as X^t, with rate matrix Q^t. Then the Kolmogorov forward equation that characterizes the distribution ρ^t for X^t is given by:

$$\frac{d\rho^t}{dt} = \rho^t Q^t = \Big(\sum_i \rho_i^t Q_{ij}^t\Big)_{j=1}^M. \tag{19}$$

When $c_{ij}(\rho) = c_{ij}$ is a constant, computing the rate Q^t via (18) and (19) requires knowledge of the current distribution ρ^t, which is typically intractable. However, we find that, inspired by physics, a proper choice of $c_{ij}(\rho)$ as the

conductance in nonequilibrium chemical reactions (Qian & Beard, 2005) can avoid computing ρ^t. In particular, set

$$c_{ij}(\rho) = w_{ij}\frac{g(\pi_j/\pi_i)\rho_i - g(\pi_i/\pi_j)\rho_j}{f_i + \log\rho_i - f_j - \log\rho_j}, \tag{20}$$

where w_{ij} satisfying $w_{ij} = w_{ji} \in \mathbb{R}$ is an inherent variability between i and j, independent of π and ρ; and $g(\cdot)$ is the locally balanced (LB) function satisfying $g(a) = ag(\frac{1}{a})$ broadly used in recent LB samplers (Zanella, 2020); a more detailed derivation of $c_{ij}(\rho)$ is given in Appendix B.4. Such a $c_{ij}(\rho)$ can significantly simplify (18) to

$$\frac{d\rho^t}{dt} = \left(\sum_{i\neq j} w_{ij}\left[\rho_i^t g\left(\frac{\pi_j}{\pi_i}\right) - \rho_j^t g\left(\frac{\pi_i}{\pi_j}\right)\right]\right)_{j=1}^{M}. \tag{21}$$

In this case, we can set the rate matrix Q^t in (19) as a tractable, time homogeneous matrix Q such that:

$$Q_{ij} = \begin{cases} w_{ij}g\left(\frac{\pi_j}{\pi_i}\right), & i \neq j \\ -\sum_{k\neq i} w_{ik}g\left(\frac{\pi_k}{\pi_i}\right), & i = j \end{cases}. \tag{22}$$

Accordingly, the Markov jump process X^t associated with the Wasserstein gradient flow (18) can be characterized as a differential equation with respect to the transition probabilities:

$$\frac{d}{dh}\mathbb{P}(X^{t+h} = j | X^t = i) = w_{ij}g\left(\frac{\pi_j}{\pi_i}\right). \tag{23}$$

Since the Markov jump process determined by (23) gives a particle realization of the DWGF in (21), we refer to it as *discrete Langevin dynamics* (DLD).

4 SAMPLING ALGORITHM

Next, we study how to efficiently simulate the DLD to sample from a discrete space. We consider the state space $V = \mathcal{C}^N = \{1, ..., C\}^N$, where N is the dimension and $\mathcal{C}$ is a code book with elements represented by one-hot vectors. We follow the commonly used assumption (Grathwohl et al., 2021; Sun et al., 2021a; Zhang et al., 2022) that the energy function $f(\cdot)$ in (12) is differentiable. Since the approximation error of $\langle\nabla f(x), y - x\rangle$ for $f(y) - f(x)$ is repaired by the MH test (8), we ignore such differences in this section.

Generally, there exist many different choices for simulating the DLD. For example, the Gillespie algorithm (Gillespie, 1977) can simulate a continuous-time trajectory exactly, but is computationally expensive. In this work, since we are only interested in the target distribution π, we can focus on more efficient MCMC algorithms.

4.1 Casting Previous Samplers as DLD

We first show that previous locally balanced (LB) samplers are essentially simulating the *discrete Langevin dynamics* (DLD) by choosing particular discretizations.

Single Jump Consider a special case $V = \{0, 1\}^N$. Denote $x, y \in V$ as the current and the next state, LB-1 (Zanella, 2020) and GWG (Grathwohl et al., 2021) set

$$w_{xy} = 1_{\{\sum_{n=1}^N \|x_n - y_n\| = 1\}}; \tag{24}$$

that is to say, $w_{xy} = 1$ if and only if there exists an index n, such that $y_n \neq x_n$ and for any other $i = 1, ..., n-1, n+1, ..., N$, $x_i = y_i$. Such a weight w restricts the new state y to lie within the 1-Hamming ball of x. Then, LB-1 and GWG propose the new state y with probability

$$q(x, y) \propto w_{xy}g\left(\frac{\pi(y)}{\pi(x)}\right). \tag{25}$$

Such a categorical distribution is exactly the first transition probability for X^t satisfying DLD (23). Specifically, we denote the jumping time of X^t as $J_1, ..., J_m, ...$, and assume X^t jumps to y from x at time J_m, then we have

$$\mathbb{P}(X^{J_m} = y | X^{J_{m-1}} = x) \propto w_{xy}g\left(\frac{\pi(y)}{\pi(x)}\right). \tag{26}$$

A more detailed derivation for (26) is given in Appendix B.1. Hence, one can claim that LB-1 and GWG propose the new state by simulating the first jump of DLD.

Multiple Jumps Denote the current state as $\sigma^0 \in V$ and the path length as L. PAS (Sun et al., 2021a, 2022) propose a new state $\sigma^L \in V$ along the auxiliary path σ. Specifically, they sequentially propose $\sigma^{l-1} = x, \sigma^l = y$ from

$$q(x, y) \propto w_{xy}g\left(\frac{\pi(y)}{\pi(x)}\right). \tag{27}$$

Similar to the analysis of single jump samplers, for process X^t with the first jump after J_{l-1} occurs at time J_l, we have:

$$\mathbb{P}(X^{J_l} = y | X^{J_{l-1}} = x) \propto w_{xy}g\left(\frac{\pi(y)}{\pi(x)}\right). \tag{28}$$

Hence, one can claim that PAS propose a new state by simulating the first L jumps of DLD.

Parallel Jumps Denote the current state as $x \in V$. Recent work (Zhang et al., 2022; Rhodes & Gutmann, 2022) has generalized the Gaussian proposal in LMC (9) to discrete spaces by restricting the proposal distribution to discrete points:

$$q(x, y) \propto \prod_{n=1}^N \exp(-r_n(x, y_n)) \tag{29}$$

$$r_n(x, y_n) = \langle y_n - x_n, \frac{\partial}{\partial x_n}f(x)\rangle + \frac{\|x_n - y_n\|}{2\alpha} \tag{30}$$

where each dimension for $y = (y_1, ..., y_N)$ are sampled independently, which allows convenient parallel implementation. The proposal (29) can be seen as setting

$$w_{xy} = 1, \quad h = \exp\Big(-\frac{1}{2\alpha}\Big) \tag{31}$$

and using the forward Euler's method to approximate (23) with simulation time h:

$$\mathbb{P}(X_n^{t+h} = y_n | X^t = x) \propto h w_{xy} g\Big(\frac{\pi(x_{\backslash n}, y_n)}{\pi(x_{\backslash n}, x_n)}\Big). \tag{32}$$

This value is correct when $y_n \neq x_n$. However, such a Gaussian proposal, copied from a continuous diffusion process, does not use the correct diagonal rate for a discrete jump process. In particular, (29) has

$$\mathbb{P}(X_n^{t+h} = x_n | X^t = x) \propto 1, \tag{33}$$

which corresponds to a rate $r_n(x, x_n) = 0$ on the diagonal in the rate matrix. However, in a jump process, the correct rate on the diagonal should be $r_n(x, x_n) = -\sum_{y_n \neq x_n} r(x, y_n)$ the negative summation of the off-diagonal entries. Such a mismatch reduces the quality of the proposal distribution; see Appendix B.2 for a more detailed discussion. Also, one can effectively improve the sampling efficiency via correcting the rate in diagonal; see more results in Appendix C.

4.2 Discrete Langevin Monte Carlo

The framework induced by the *discrete Wasserstein gradient flow* (DWGF) in (21) provides a more principled way to design gradient based MCMC algorithms by estimating the transition probability matrix. In particular, since the gradient flow is a Markov jump process, we have the closed form for the transition:

$$\rho^{t+h} = \rho^t P^h, \quad P^h = \exp(Qh), \tag{34}$$

for $Q \in \mathbb{R}^{C^N \times C^N}$ in (22). Of course, it is impractical to directly calculate the matrix exponential $\exp(Qs)$ for a large rate matrix Q. Instead, by factorizing the jump process $X^t = (X_1^t, X_2^t, ..., X_N^t) = (x_1^t, x_2^t, ..., x_N^t)$, one can simulate each sub-processes X_n^t with initial value x_n^t, independently. In this case, the distribution ρ_n^t for X_n^t has the following closed form expression for the transition:

$$\rho_n^{t+h} = \rho_n^t P_n^h(x^t), \quad P_n^h(x^t) = \exp(Q_n(x^t)h), \tag{35}$$

where the rate matrix $Q_n(x^t) \in \mathbb{R}^{C \times C}$ depends on the current state x^t. In particular, for index $i \neq j \in \mathcal{C}$, the rate matrix $Q_n(x^t)$ satisfies

$$Q_n(x^t)(i, j) = w_{ij} g\left(\pi(x_{\backslash n}^t, j)/\pi(x_{\backslash n}^t, i)\right). \tag{36}$$

For simplicity, we will drop the x^t and only use $P_n^h(i, j)$, $Q_n(i, j)$ when this does not cause ambiguity. For the binary case $C = 2$, denoting $\alpha = Q_n(1, 2)$ and $\beta = Q_n(2, 1)$, then the transition matrix P_n^h has a closed form expression (37):

Algorithm 1: DLMC MH step

Input: current state x^t, step time h, target π
Output: new state x^{t+h}
for *n=1, ..., N* **do** `// Run in parallel`
  Calculate x_n^t-row of $\tilde{P}_n^h(x^t)$ in (39)
  Sample $y_n \propto \tilde{P}_n^h(x^t)(x_n^t, y_n)$
end
Compute $A = \min\{1, \frac{\pi(y)\prod_{n=1}^N \tilde{P}_n^h(y)(y_n, x_n^t)}{\pi(x)\prod_{n=1}^N \tilde{P}_n^h(x^t)(x_n^t, y_n)}\}$
if *rand*$(0, 1) < A$ **then** $x^{t+h} = y$ **else** $x^{t+h} = x^t$;

$$P_n^h = \begin{pmatrix} \frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta}e^{-(\alpha+\beta)h} & \frac{\alpha}{\alpha+\beta} - \frac{\alpha}{\alpha+\beta}e^{-(\alpha+\beta)h} \\ \frac{\beta}{\alpha+\beta} - \frac{\beta}{\alpha+\beta}e^{-(\alpha+\beta)h} & \frac{\alpha}{\alpha+\beta} + \frac{\beta}{\alpha+\beta}e^{-(\alpha+\beta)h} \end{pmatrix}. \tag{37}$$

One can sample y_n from the x_n^t-th row of P_n^g, a categorical distribution, for $n = 1, ..., N$ in parallel. Hence, the new state $y = (y_1, ..., y_N)$ can be efficiently obtained.

For the categorical case $C > 2$, we do not have a simple closed form expression of P_n^h for all C. Instead, we generalize the expression in (37). Denote

$$\nu_n(x^t)(j) = \pi(x_{\backslash n}^t, j)/\textstyle\sum_{i=1}^C \pi(x_{\backslash n}^t, i) \tag{38}$$

as the stationary distribution induced by $Q_n(x^t)$. For simplicity, we drop x^t and only use $\nu_n(j)$ when this does not cause ambiguity. We approximate the transition as:

$$\tilde{P}_n^h(i, j) = \begin{cases} \nu_n(i) + \sum_{k \neq i} \nu_n(k) e^{-h\frac{Q_n(i,k)}{\nu_n(k)}}, & i = j \\ \nu_n(j) - \nu_n(j) e^{-h\frac{Q_n(i,j)}{\nu_n(j)}}, & i \neq j \end{cases}. \tag{39}$$

Such an approximation is consistent with the special case $C = 2$ in (37), and satisfies the boundary conditions $\tilde{P}_n^0 = P_n^0$, $\tilde{P}_n^\infty = P_n^\infty$, $\frac{d}{dh}\tilde{P}_n^h|_{h=0} = \frac{d}{dh}P_n^h|_{h=0}$ for arbitrary C. Hence, Equation (39) provides a better approximation than the forward Euler's method. In practice, we find (39) does not lose much proposal quality compared to calculating the matrix exponential for P_n^h in (35). On the other hand, we only need to compute the x_n^t-row in $\tilde{P}_n^h$, with computational cost is $O(C)$. This is much more efficient than a generic numerical approximation of the matrix exponential with cost $O(C^3)$ (Al-Mohy & Higham, 2010).

In a concurrent work (Sun et al., 2023), the "globally balanced" phenomenon is observed, where using $g(a) = a^\alpha$ with $\alpha \in (0.5, 1]$ can have better performance than that with $\alpha = 0.5$ in some distributions. Such a phenomenon occurs when the target distribution is nearly factorized, for example Bernoulli distribution in section 6.2. In DLMC, the selection of α is implicitly done when tuning the simulation time h. Specifically, when h is small, the transition probability (39) is equivalent with using forward Euler's

method with $g(a) = a^{\frac{1}{2}}$. When h is large, the transition probability degenerates to the stationary distribution ν in (38), whose value can be obtained by using $g(a) = a$. This explains the "globally balanced" phenomenon is originated from a large simulation time of DLD.

Combining the DLD and the discretization via (39), we obtain the *Discrete Langevin Monte Carlo* (DLMC) algorithm. The discritization in (39) not only provides a factorized proposal distribution for parallel computing, but also gives a time-uniform slicing of DLD. By contrast, PAS (Sun et al., 2021a, 2022) only flips a fixed number of sites in each MH step, thereby has a simulation time that depends on the current state. Specifically, PAS has a shorter simulation time at states with a larger jump rate, and longer simulation time at states with a smaller jump rate. Consequently, PAS realizes a non-uniform time slicing of DLD, which leads to more proposal rejections in comparison to DLMC; see Appendix B.3 for more details. Pseudo code for an MH step of DLMC is given in Algorithm 1.

5 RELATED WORK

Gradient based MCMC algorithms that simulate Langevin dynamics (Rossky et al., 1978; Girolami & Calderhead, 2011; Welling & Teh, 2011) or Hamiltonian dynamics (Duane et al., 1987; Neal et al., 2011; Hoffman et al., 2014), can substantially improve sampling efficiency in both theory and practice. The seminal work of Jordan et al. (1998) and Otto (2001) shows that the Langevin dynamics simulates the gradient flow on the 2-Wasserstein space $\mathcal{P}_2(\mathbb{R}^D)$ (Villani, 2009). Subsequent work has directly studied the Wasserstein gradient flow (Mokrov et al., 2021) or extended the result to Hamiltonian dynamics (Ambrosio & Gangbo, 2008; Liu et al., 2019; Chow et al., 2020) and particle variational inference (Chen et al., 2018; Liu et al., 2019). By contrast, the corresponding theory for sampling algorithms in discrete spaces is less well understood. Mielke (2011); Maas (2011); Chow et al. (2012) introduce 2-Wasserstein distances on finite graphs via the Benamou-Brenier formula (Benamou & Brenier, 2000). However, these works do not investigate Langevin dynamics or sampling algorithms in discrete space.

A number of samplers for discrete spaces construct invertible mappings between discrete and continuous spaces via auxiliary variables, uniform dequantization, or VAE flow (Zhang et al., 2012; Pakman & Paninski, 2013; Nishimura et al., 2017; Han et al., 2020; Jaini et al., 2021). Such methods work in some scenarios, but a key challenge is that embedding the discrete space in a continuous space can destroy the inherent discrete structure, resulting in irregular target distributions in the continuous space such that compromises performance in high dimensional discrete spaces (Grathwohl et al., 2021).

Another group of methods work directly on discrete spaces. Dai et al. (2020); Titsias & Yau (2017) augment the discrete space with an auxiliary variable, but still rely on slow Gibbs sampling for improvement. Zanella (2020) introduces an informed proposal for discrete spaces, and proves that a family of locally balanced (LB) functions is asymptotically optimal. Following this work, Grathwohl et al. (2021); Sun et al. (2021a); Zhang et al. (2012); Rhodes & Gutmann (2022); Sun et al. (2022, 2023) propose various LB samplers. Despite these LB samplers substantially improving sampling efficiency in discrete spaces by mimicking LMC, their lack of a principled connection to the *discrete Langevin dynamics* (DLD) results in sub-optimal proposal distribution designs. We note that special cases of DLD (23) have been mentioned in previous work (Sohl-Dickstein et al., 2009; Power & Goldman, 2019) but without realizing the connection to gradient flow.

6 SAMPLING FROM CLASSICAL EBMS

6.1 Settings

Models We demonstrate the advantage of DLMC in sampling tasks on four classical models: the Bernoulli model (Bernoulli), Ising model (Ising) (Ising, 1924), factorial hidden Markov model (FHMM) (Ghahramani & Jordan, 1995), and the restricted Boltzmann machine (RBM) (McClelland et al., 1987). For each model, we use a binary version $C = 2$ with high or low temperature, a 4-category version $C = 4$, and an 8-category version $C = 8$. Compared to the low temperature model, the high temperature model is smoother and has larger entropy. Here we only report the results on high temperature version and 8-category version. More description of the models and additional results are given in Appendix C.

Baselines We consider the LB samplers GWG (Grathwohl et al., 2021), PAS (Sun et al., 2022), and DMALA (Zhang et al., 2022). Note that, for PAS, we follow the implementation in Sun et al. (2022), which is computationally more efficient compared to the PAS in the original paper (Sun et al., 2021a). Also, NCG (Rhodes & Gutmann, 2022) is equivalent to DMALA, so we do not report results for NCG. For these LB samplers, we consider two commonly used weight functions $g(a) = \sqrt{a}$ and $g(a) = \frac{a}{a+1}$. Also, we select the optimal hyperparameters by tuning the average acceptance rate to 0.574 following the result (Sun et al., 2022). In particular, we tune U, how many sites to flip per MH step for PAS, α, the step size for DMALA, and h, the simulation time for DLMC. Although the optimality for 0.574 is only proved for PAS, we find it robustly produce good results for DMALA and DLMC, so we still use this technique.

We also compare with classical discrete samplers: random walk Metropolis (RWM), the Hamming Ball sampler (HB) (Titsias & Yau, 2017), and block Gibbs (BG). Following

Grathwohl et al. (2021), we use a block size of 10 and hamming distance 1 for HB, and a block size of 2 for BG. Silimar to PAS, we also select an optimal U, how many sites to flip per MH step, for RWM by setting the average acceptance rate to 0.234 to achieve the otpimal efficiency (Sun et al., 2022).

Metrics We use effective sample size (ESS) to evaluate each sampler (Lenth, 2001). To reduce the effects of implementation, we report ESS normalized by the number of energy evaluations, and the running time. For methods requires gradients, we count each gradient backpropagation as one call of the energy function as they have the similar computational cost. The former measure ignores the computational cost for sampling a new state from the proposal distribution and focuses on proposal quality. The latter measure reflects proposal efficiency. For each setting and sampler, we run 100 chains for 100,000 steps, with 50,000 burn-in steps to ensure the chain mixes.

6.2 Results

Bernoulli and Categorical The Bernoulli distribution is the simplest distribution in a discrete space, consisting of independent binary random variables. The categorical distribution is a simple generalization to categorical random variables. For $x \in \mathcal{C}^N$, the energy function is:

$$f(x) = \sum_{n=1}^{N} \langle x_n, \theta_n \rangle \tag{40}$$

We report the results in the first row of Figure 1. We can see that the DLMC significantly outperforms all the other samplers. On Bernoulli and categorical distributions, DLMC has ESS with respect to energy evaluation larger than 10^4. Since each step of DLMC requires 4 evaluations of energy function, it basically means the 50k samples collected by DLMC are independent. The reason is that DLMC does not lose accuracy in factorizing. With a simulation time large enough, the proposal distribution in (39) is exactly the target distribution.

Another interesting observation is that, compared to PAS, DMALA has a smaller ESS with respect to the number of energy evaluations, but a larger ESS with respect to the running time. The reason is that DMALA does not correctly simulate the WGF, which reduces its proposal quality. PAS generates the new state by constructing an auxiliary path sequentially, where the lack of parallelism reduces efficiency.

Ising and Potts The Ising model (Ising, 1924) is a mathematical model of ferromagnetism in statistical mechanics. It consists of binary random variables arranged in a lattice graph $G = (V, E)$ and allows each node to interact with its neighbors. The Potts model (Potts, 1952) is a generalization of the Ising model where the random variables are categorical. The energy function is:

$$f(x) = -\sum_{n=1}^{N} \langle x_n, \theta_n \rangle - \lambda \sum_{(i,j)\in E} \delta(x_i, x_j) \tag{41}$$

We report the results in the second row of Figure 1. We can see that the advantage of DLMC narrows compared to Bernoulli model, as Ising and Potts models are not factorized, but the gap is still significant. Also, one can notice an interesting phenomenon that all LB samplers, except for DLMC, have large ESS with $g(a) = \frac{a}{a+1}$ in binary models and with $g(a) = \sqrt{a}$ in categorical models.

FHMM The Factorial Hidden Markov Model (Ghahramani & Jordan, 1995) uses latent variables to characterize time series data. In particular, it assumes the continuous data $y \in \mathbb{R}^L$ is generated by hidden state $x \in \mathcal{C}^{L\times K}$. When $C = 2$, we call it a binary FHMM (binFHMM) and when $C > 2$, we call it a categorical FHMM (catFHMM). The probability function is:

$$p(x) = p(x_1) \prod_{l=2}^{L} p(x_t | x_{t-1}) \tag{42}$$

$$p(y|x) = \prod_{l=1}^{L} \mathcal{N}(y_t; \sum_{k=1}^{K} \langle W_k, x_{l,k} \rangle + b; \sigma^2) \tag{43}$$

We report the results in the third row of Figure 1. Similar to the Ising model, we can see that all LB samplers demonstrate good efficiency and DLMC still leads the performance. In FHMM, we can see that using $g(a) = \sqrt{a}$ is more efficient than using $g(a) = \frac{a}{a+1}$ across all LB samplers. One possible reason is that using $g(a) = \sqrt{a}$ is more likely to jump to high probability states that makes the sampling more efficient on smooth target distributions, but less robust on nonsmooth target distributions (Livingstone & Zanella, 2019).

RBM The Restricted Boltzmann Machine (Smolensky, 1986) is an unnormalized latent variable model, with a visible random variable $v \in \mathcal{C}^N$ and a hidden random variable $h \in \{0, 1\}^M$. When $C = 2$, we call it a binary RBM (binRBM) and when $C > 2$, we call it a categorical RBM (catRBM). The energy function is:

$$f(v) = \sum_{h} \Big[-\sum_{n=1}^{N} \langle v_n, \theta_n \rangle - \sum_{m=1}^{M} \beta_m h_m - \sum_{n,m} \langle h_m \theta_{m,n}, v_n \rangle \Big] \tag{44}$$

Unlike the previous three models, where the parameters are hand designed, we train binRBM on MNIST (LeCun et al., 1998) and catRBM on Fashion-MNIST (Xiao et al., 2017) using contrastive divergence (Hinton, 2002). The learned

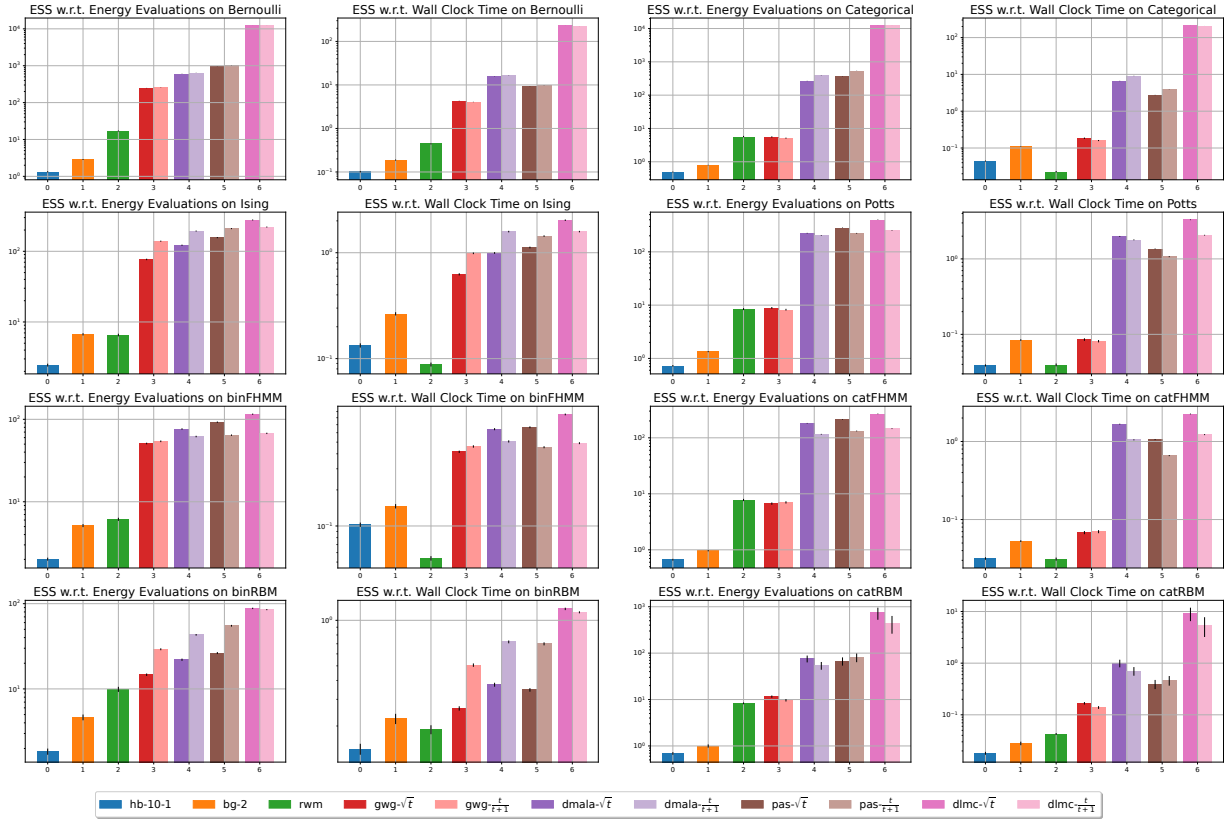


Figure 1: Effective Sample Size (↑) on Various Distributions in log scale

RBMs have stronger multi-modality compared to previous models and are harder to sample from. We report the results in the fourth row of Figure 1. We can see that DLMC is significantly more efficient than all the other samplers with respect to both number of energy evaluations and the running time. Also, we can see that in binRBM, although $g(a) = \frac{a}{a+1}$ is significantly more efficient in other LB samplers, DLMC still has larger ESS using $g(a) = \sqrt{a}$.

7 LEARNING DEEP EBMS

Deep energy-based models (EBMs) have gained increasing popularity. Recent advances including tempered Langevin samplers (Nijkamp et al., 2020), large persistent chains (Du & Mordatch, 2019), and amortized sampling (Dai et al., 2019, 2020), which have enabled deep EBMs to become a competitive approach for generative modeling (Song et al., 2020; Sun et al., 2021b; Xie et al., 2021; Bakhtin et al., 2021). However, learning an EBM is challenging. Given data sampled from a true distribution π, we maximize the likelihood of the target distribution $\pi_\theta(x) \propto e^{-f_\theta(x)}$ parameterized by θ. The gradient of the likelihood is:

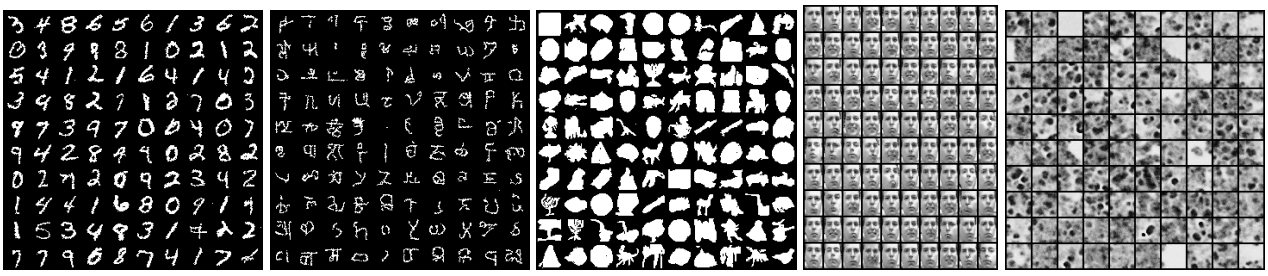
$$\nabla_\theta \log \pi_\theta(x) = \mathbb{E}_\pi[\nabla_\theta f_\theta(x)] - \mathbb{E}_{\pi_\theta}[\nabla_\theta f_\theta(x)] \quad (45)$$

The first expectation can be estimated using the data from the true distribution. The second expectation requires samples from the current model, which is are typically obtained via MCMC. The speed of EBM training is determined by how fast the MCMC algorithm can obtain a good estimate of the second expectation. Following Grathwohl et al. (2021) and Sun et al. (2021a), we train deep EBMs parameterized by residual networks (He et al., 2016) on binary and grayscale image datasets using PCD (Tieleman, 2008) with a replay buffer (Du & Mordatch, 2019). The grayscale images were treated as 1-of-256 categorical data.

We present the test-set likelihoods in Table 1 and Table 2. Likelihoods are estimated using annealed importance sampling (Neal, 2001). Since the quality of the learned EBMs will be similar as long as the sampler is good enough with certain steps per model update, we measure the efficiency of samplers by the minimum number of MCMC steps needed to chain a decent EBM. We compare the performance of DLMC to Variational Autoencoder (Kingma & Welling, 2013), an RBM, a deep belief network (DBN) (Hinton, 2009) and EBMs trained by Gibbs, GWG (Grathwohl et al., 2021), PAS (Sun et al., 2021a, 2022), and DMALA (Zhang et al., 2022). We use weight function $g(t) = \sqrt{t}$ for all LB samplers. On all datasets, the DLMC samplers enable deep EBMs to become competitive on high dimensional discrete data. We also present long-run samples from the EBMs trained by DLMC in Figure 2.

Table 1: Evaluation of effectiveness on learning binary EBMs.

Data Type	Dataset	VAE (MLP)	VAE (Conv)	RBM	DBN	EBM (GWG)	EBM (Gibbs)	EBM (PAS)	EBM (DMALA)	EBM (DLMC)
Binary	Static MNIST	-86.05	-82.41	-86.39	-85.67	-80.01	-117.17	-79.58	-79.46	**-79.13**
	Dynamic MNIST	-82.42	-80.40	–	–	-80.51	-121.19	-79.59	-79.54	**-78.84**
log-likelihood ↑	Omniglot	-103.52	97.65	-100.47	-100.78	-94.72	-142.06	**-90.75**	-91.11	-90.84
	Caltech Silhouettes	-112.08	-106.35	–	–	-96.20	163.50	-84.56	-87.82	**-77.04**



Binary MNIST Omniglot Caltech Frey Faces Histopathology

Figure 2: Samples from learned deep EBMs using proposed LBJ sampler.

Table 2: Evaluation on learning categorical EBMs.

Data Type	Dataset	VAE (MLP)	VAE (Conv)	EBM (GWG)	EBM (PAS)	EBM (DLMC)
Categorical	Frey Faces	4.61	4.49	4.65	4.74	**4.33**
(bits/dim ↓)	Histopathology	5.82	5.59	5.08	5.1	**4.91**

8 DISCUSSION

We have described the *discrete Langevin dynamics* (DLD) and showed that it simulates the *discrete Wasserstein gradient flow* (DWGF) to minimize the KL-divergence to a target distribution. Such a view provides an unified framework to design gradient based samplers for discrete spaces. Based on this perspective, we proposed a new algorithm, *discrete Langevin Monte Carlo* (DLMC), that improves the efficiency of existing locally balanced samplers in both sampling and learning tasks across various discrete distributions. Despite the success of DLMC presented in this work, there remain several interesting problems to investigate.

The DWGF and DLD are determined by topological structure via w_{ij} and $g(\cdot)$ (100). For w_{ij}, in complicated discrete spaces like Hamiltonian cycles, one can resort to powerful heuristics to construct shortcuts between a current state i and a new state j. For the weight function, we empirically evaluated the most commonly used alternatives $g(t) = \sqrt{t}$ and $g(t) = \frac{t}{t+1}$ in Sec. 6.2 and found that each has its own advantages on different models. Although Sansone (2021) has some initial attempts to learn $g(\cdot)$ as a linear combination of 4 commonly used weight functions, a principled understanding of the weight function is still missing. In future work, we will investigate whether DWGF can be used as a tool to analyze the choice of $g(\cdot)$.

After obtaining DLD, one can use any kind of discretization to obtain a sampling algorithm. Besides the DLMC and previous LB samplers, there are many other choices. For example, we can use forward Euler's method to simulate DLD, which we call DLMCf. DLMCf can be seen as DMALA with a corrected diagonal rate. In Appendix C, we show that DLMCf has comparable performance with DLMC on several models, and is substantially more efficient than DMALA. Considering fruitful numerical methods like Heun's method and Runge-Kutta method can outperforms forward Euler's method in many scenarios, we believe there is ample room to improve the discretization.

Overall, DWGF and DLD provide a new framework for sampling algorithms in discrete spaces. We believe this is a milestone for sampling in discrete spaces and expect further development upon this framework.

Acknowledgements

We thank four anonymous reviewers for their helpful comments to improve the manuscript. Dale Schuurmans gratefully acknowledges support from a CIFAR Canada AI Chair, NSERC and Amii.

References

Al-Mohy, A. H. and Higham, N. J. A new scaling and squaring algorithm for the matrix exponential. *SIAM Journal on Matrix Analysis and Applications*, 31(3):970–989, 2010.

Ambrosio, L. and Gangbo, W. Hamiltonian odes in the wasserstein space of probability measures. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 61(1): 18–53, 2008.

Bakhtin, A., Deng, Y., Gross, S., Ott, M., Ranzato, M., and Szlam, A. Residual energy-based models for text. *J. Mach. Learn. Res.*, 22:40–1, 2021.

Beard, D. A. and Qian, H. Relationship between thermodynamic driving force and one-way fluxes in reversible processes. *PloS one*, 2(1):e144, 2007.

Benamou, J.-D. and Brenier, Y. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

Chen, C., Zhang, R., Wang, W., Li, B., and Chen, L. A unified particle-optimization framework for scalable bayesian sampling. *arXiv preprint arXiv:1805.11659*, 2018.

Cheng, X. and Bartlett, P. Convergence of langevin mcmc in kl-divergence. In *Algorithmic Learning Theory*, pp. 186–211. PMLR, 2018.

Chow, S.-N., Huang, W., Li, Y., and Zhou, H. 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.

Chow, S.-N., Li, W., and Zhou, H. Entropy dissipation of fokker-planck equations on graphs. *arXiv preprint arXiv:1701.04841*, 2017.

Chow, S.-N., Li, W., and Zhou, H. Wasserstein hamiltonian flows. *Journal of Differential Equations*, 268(3): 1205–1219, 2020.

Christensen, O. F., Roberts, G. O., and Rosenthal, J. S. Scaling limits for the transient phase of local metropolis–hastings algorithms. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):253–268, 2005.

Dai, B., Liu, Z., Dai, H., He, N., Gretton, A., Song, L., and Schuurmans, D. Exponential family estimation via adversarial dynamics embedding. *Advances in Neural Information Processing Systems*, 32, 2019.

Dai, H., Singh, R., Dai, B., Sutton, C., and Schuurmans, D. Learning discrete energy-based models via auxiliary-variable local exploration. *Advances in Neural Information Processing Systems*, 33:10443–10455, 2020.

Do Carmo, M. P. and Flaherty Francis, J. *Riemannian geometry*, volume 6. Springer, 1992.

Du, Y. and Mordatch, I. Implicit generation and generalization in energy-based models. *arXiv preprint arXiv:1903.08689*, 2019.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

Ghahramani, Z. and Jordan, M. Factorial hidden markov models. *Advances in Neural Information Processing Systems*, 8, 1995.

Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81 (25):2340–2361, 1977.

Girolami, M. and Calderhead, B. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

Grathwohl, W., Swersky, K., Hashemi, M., Duvenaud, D., and Maddison, C. J. Oops i took a gradient: Scalable sampling for discrete distributions. *arXiv preprint arXiv:2102.04509*, 2021.

Han, J., Ding, F., Liu, X., Torresani, L., Peng, J., and Liu, Q. Stein variational inference for discrete distributions. In *International Conference on Artificial Intelligence and Statistics*, pp. 4563–4572. PMLR, 2020.

Hastings, W. K. Monte carlo sampling methods using markov chains and their applications. 1970.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Hinton, G. E. Deep belief networks. *Scholarpedia*, 4(5): 5947, 2009.

Hoffman, M. D., Gelman, A., et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

Ising, E. *Beitrag zur theorie des ferro-und paramagnetismus*. PhD thesis, Grefe & Tiedemann, 1924.

Jaini, P., Nielsen, D., and Welling, M. Sampling in combinatorial spaces with survae flow augmented mcmc. In *International Conference on Artificial Intelligence and Statistics*, pp. 3349–3357. PMLR, 2021.

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

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lenth, R. V. Some practical guidelines for effective sample size determination. *The American Statistician*, 55(3): 187–193, 2001.

Liu, C., Zhuo, J., and Zhu, J. Understanding mcmc dynamics as flows on the wasserstein space. In *International Conference on Machine Learning*, pp. 4093–4103. PMLR, 2019.

Livingstone, S. and Zanella, G. The barker proposal: combining robustness and efficiency in gradient-based mcmc. *arXiv preprint arXiv:1908.11812*, 2019.

Ma, Y.-A., Chen, T., and Fox, E. A complete recipe for stochastic gradient mcmc. *Advances in neural information processing systems*, 28, 2015.

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

McClelland, J. L., Rumelhart, D. E., Group, P. R., et al. *Parallel Distributed Processing, Volume 2: Explorations in the Microstructure of Cognition: Psychological and Biological Models*, volume 2. MIT press, 1987.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

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

Mokrov, P., Korotin, A., Li, L., Genevay, A., Solomon, J. M., and Burnaev, E. Large-scale wasserstein gradient flows. *Advances in Neural Information Processing Systems*, 34:15243–15256, 2021.

Neal, R. M. Annealed importance sampling. *Statistics and computing*, 11(2):125–139, 2001.

Neal, R. M. et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Nijkamp, E., Hill, M., Han, T., Zhu, S.-C., and Wu, Y. N. On the anatomy of mcmc-based maximum likelihood learning of energy-based models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5272–5280, 2020.

Nishimura, A., Dunson, D., and Lu, J. Discontinuous hamiltonian monte carlo for sampling discrete parameters. *arXiv preprint arXiv:1705.08510*, 853, 2017.

Otto, F. The geometry of dissipative evolution equations: the porous medium equation. 2001.

Pakman, A. and Paninski, L. Auxiliary-variable exact hamiltonian monte carlo samplers for binary distributions. *arXiv preprint arXiv:1311.2166*, 2013.

Potts, R. B. Some generalized order-disorder transformations. In *Mathematical proceedings of the cambridge philosophical society*, volume 48, pp. 106–109. Cambridge University Press, 1952.

Power, S. and Goldman, J. V. Accelerated sampling on discrete spaces with non-reversible markov processes. *arXiv preprint arXiv:1912.04681*, 2019.

Qian, H. and Beard, D. A. Thermodynamics of stoichiometric biochemical networks in living systems far from equilibrium. *Biophysical chemistry*, 114(2-3):213–220, 2005.

Rhodes, B. and Gutmann, M. Enhanced gradient-based mcmc in discrete spaces. *arXiv preprint arXiv:2208.00040*, 2022.

Robert, C. and Casella, G. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

Rossky, P. J., Doll, J., and Friedman, H. Brownian dynamics as smart monte carlo simulation. *The Journal of Chemical Physics*, 69(10):4628–4633, 1978.

Sansone, E. Lsb: Local self-balancing mcmc in discrete spaces. *arXiv preprint arXiv:2109.03867*, 2021.

Smolensky, P. Information processing in dynamical systems: Foundations of harmony theory. Technical report, Colorado Univ at Boulder Dept of Computer Science, 1986.

Sohl-Dickstein, J., Battaglino, P., and DeWeese, M. R. Minimum probability flow learning. *arXiv preprint arXiv:0906.4779*, 2009.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Sun, H., Dai, H., Xia, W., and Ramamurthy, A. Path auxiliary proposal for mcmc in discrete space. In *International Conference on Learning Representations*, 2021a.

Sun, H., Dai, H., and Schuurmans, D. Optimal scaling for locally balanced proposals in discrete spaces. *arXiv preprint arXiv:2209.08183*, 2022.

Sun, H., Dai, B., Sutton, C., Schuurmans, D., and Dai, H. Any-scale balanced samplers for discrete space. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=lEkl0jdSb7B.

Sun, R., Dai, H., Li, L., Kearnes, S., and Dai, B. Towards understanding retrosynthesis by energy-based models. *Advances in Neural Information Processing Systems*, 34, 2021b.

Tieleman, T. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, pp. 1064–1071, 2008.

Titsias, M. K. and Yau, C. The hamming ball sampler. *Journal of the American Statistical Association*, 112 (520):1598–1611, 2017.

Tran, T., Phung, D., and Venkatesh, S. Mixed-variate restricted boltzmann machines. In *Asian conference on machine learning*, pp. 213–229. PMLR, 2011.

Villani, C. *Optimal transport: old and new*, volume 338. Springer, 2009.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688. Citeseer, 2011.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Xie, Y., Shi, C., Zhou, H., Yang, Y., Zhang, W., Yu, Y., and Li, L. Mars: Markov molecular sampling for multi-objective drug discovery. *arXiv preprint arXiv:2103.10432*, 2021.

Zanella, G. Informed proposals for local mcmc in discrete spaces. *Journal of the American Statistical Association*, 115(530):852–865, 2020.

Zhang, R., Liu, X., and Liu, Q. A langevin-like sampler for discrete distributions. In *International Conference on Machine Learning*, pp. 26375–26396. PMLR, 2022.

Zhang, Y., Ghahramani, Z., Storkey, A. J., and Sutton, C. Continuous relaxations for discrete hamiltonian monte carlo. *Advances in Neural Information Processing Systems*, 25:3194–3202, 2012.

Appendices

A COMPLETE PROOFS

In order to formally prove theorem 3.1, we need the following two lemmas.

Lemma A.1. *For any $\rho \in \mathcal{P}(V)$ and any vector field u, the minimizer $v^* = \arg\min_v \langle v, v\rangle$, subject to $div_\rho(v) = div_\rho(u)$, is a potential field $\nabla\Phi$. That is, there exists a function $\Phi : V \to \mathbb{R}$, such that $v^*_{ij} = \nabla\Phi_{ij} = (\Phi_i - \Phi_j)1_{\{(i,j)\in E\}}$.*

We note that a potential field is invariant up to a constant shift, meaning that if Φ is a potential function and $\Phi' = \Phi + c = (\Phi_i + c)_{i=1}^M$, then $\nabla\Phi' = \nabla\Phi$. Hence, we consider an equivalence class $[\Phi] = \{\Phi' \in \mathbb{R}^M : \exists c \in \mathbb{R}, \Phi' = \Phi + c\}$ and denote $P^M = \{[\Phi] : \Phi \in \mathbb{R}^M\}$

Lemma A.2. *For any $\rho \in \mathcal{P}(V)$, the mapping $\zeta([\Phi]) = div_\rho(\nabla\Phi)$ is a linear isomorphism between the set of equivalence classes P^M and the tangent space $T_\rho\mathcal{P}(V)$.*

In this case, the isomorphism ζ induces a metric on the tangent space $T_\rho\mathcal{P}(V)$:

Definition A.3. For any $\rho \in \mathcal{P}(V)$, we define the inner-product $\langle\cdot,\cdot\rangle_\rho$ on $T_\rho\mathcal{P}(V)$ as follows. Denote $\Phi^\sigma \in \zeta^{-1}(\sigma)$. Then for arbitrary $\sigma^1, \sigma^2 \in T_\rho\mathcal{P}(V)$, define

$$\langle\sigma^1, \sigma^2\rangle_\rho = \sum_{i=1}^M \sigma_i^1 \Phi_i^{\sigma^2} \tag{46}$$

In the following section, we first prove lemma A.1 and lemma A.2, then we justify definition A.3 is well-defined, and finally we prove theorem 3.1.

A.1 Proof for Lemma A.1

Consider $V = \{1, ..., M\}$. Denote $F(G)$ as the set of all vector fields on graph G. The divergence operator div_ρ maps a vector field $v \in F(G)$ to a vector σ in the tangent space $T_\rho\mathcal{P}(V)$. It is not hard to see that div_ρ is a surjection, but not an injection. For a $\sigma \in T_\rho\mathcal{P}(V)$, there are infinite choices of vector field v such that $\text{div}_\rho(v) = \sigma$. Lemma A.1 tells us that for Wasserstein distance defined in (17), we only need to consider vector field v as a potential field.

Proof. We show that given arbitrary vector field u, there exists a potential field $\nabla\Phi$ has the same divergence and minimizes the norm. In particular, let us consider the following optimization problem:

$$\min_v \langle v, v\rangle, \qquad \text{subject to: } \text{div}_\rho(v) = \text{div}_\rho(u) = \sigma \tag{47}$$

We introduce the dual variable $(\lambda_i)_{i=1}^M$ and we have the Lagrangian:

$$L(v, \lambda) = \frac{1}{2}\sum_{(i,j)\in E} c_{ij}v_{ij}^2 + \sum_{i=1}^M \lambda_i(\sigma_i - \sum_{j\in N(i)} c_{ji}v_{ji} - c_{ij}v_{ij}) \tag{48}$$

$$= \sum_{(i,j)\in E} [(\lambda_i - \lambda_j) + \frac{1}{2}v_{ij}]c_{ij}v_{ij} + \sum_{i=1}^M \lambda_i\sigma_i \tag{49}$$

Since u is a solution, the optimization problem (47) is feasible. Since the inner product $\langle v, v\rangle \geq 0$, the optimization problem (47) is bounded. By Slater's condition, the strong duality holds and the Lagrangian is minimized at (v^*, λ^*) with a finite value. Hence, we have $v^*_{ij} = \lambda^*_j - \lambda^*_i$. When we let $\Phi_i = \lambda^*_i$, we have $v^* = \nabla\Phi$ is a potential field. □

A.2 Proof for Lemma A.2

Lemma A.1 tells us that the minimum vector field to realize a divergence is in the form of a potential field. We can notice that a potential field is invariant up to a constant shifting. That's to say, if Φ is a potential function and $\Phi' = \Phi + c = (\Phi_i + c)_{i=1}^M$, then $\nabla\Phi' = \nabla\Phi$. Hence, we consider a equivalence class $[\Phi] = \{\Phi' \in \mathbb{R}^M : \exists c \in \mathbb{R}, \Phi' = \Phi + c\}$ and we denote $P^M = \{[\Phi] : \Phi \in \mathbb{R}^M\}$. Then, lemma A.2 gives an isomorphism between P^M and $T_\rho\mathcal{P}(V)$.

Proof. We first show $\zeta([\Phi]) = \text{div}_\rho(\nabla\Phi)$ is well-defined. For arbitrary $\Phi^1, \Phi^2 \in [\Phi]$, we have $\nabla\Phi^1 = \nabla\Phi^2$, thereby $\text{div}_\rho(\nabla\Phi^1) = \text{div}_\rho(\nabla\Phi^2)$. It indicates ζ is well-defined.

Second, we show ζ is linear. We have

$$\begin{aligned}
\zeta(\alpha[\Phi^1] + \beta[\Phi^2]) &= \zeta([\alpha\Phi^1 + \beta\Phi^2]) && (50)\\
&= \text{div}_\rho(\nabla(\alpha\Phi^1 + \beta\Phi^2)) && (51)\\
&= \alpha\text{div}_\rho(\nabla\Phi^1) + \beta\text{div}_\rho(\nabla\Phi^2) && (52)\\
&= \alpha\zeta([\Phi^1]) + \beta\zeta([\Phi^2]) && (53)
\end{aligned}$$

We have (50) holds as

$$\begin{aligned}
\psi \in \alpha[\Phi^1] + \beta[\Phi^2] &\Longleftrightarrow \exists c^1, c^2, \psi = \alpha(\Phi^1 + c^1) + \beta(\Phi^2 + c^2) && (54)\\
&\Longleftrightarrow \exists c, \psi = \alpha\Phi^1 + \beta\Phi^2 + c && (55)\\
&\Longleftrightarrow \psi \in [\alpha\Phi^1 + \beta\Phi^2] && (56)
\end{aligned}$$

Third, we show that ζ is an injection. By the property shown above, we have

$$\zeta([\Phi^1]) = \zeta([\Phi^2]) \Longleftrightarrow \zeta([\Phi^1 - \Phi^2]) = 0 \tag{57}$$

That means for any $(i, j) \in E$

$$(\Phi^1 - \Phi^2)_j - (\Phi^1 - \Phi^2)_i = 0 \tag{58}$$

Since we assume G is connected, it indicates $\Phi^1 = \Phi^2 + c$, hence $[\Phi^1] = [\Phi^2]$. As both P^M and $T_\rho\mathcal{P}(V)$ are linear space with dimension $M - 1$, we prove ζ is a linear isomorphism. □

A.3 Justification for Definition A.3

Lemma A.2 gives an immersion (Do Carmo & Flaherty Francis, 1992) from the tangent space $T_\rho\mathcal{P}(V)$ to the the set of vector fields $F(G)$. Since we define the inner-product on $F(G)$ in (16), ζ naturally induce the metric on $T_\rho\mathcal{P}(V)$. In this section, we will first justify $\langle\sigma^1, \sigma^2\rangle_\rho$ is valid. Assume $\Phi^{\sigma^2}, \Psi^{\sigma^2} \in \zeta^{-1}(\sigma^2)$, then there exists c, such that $\Phi^{\sigma^2} = \Psi^{\sigma^2} + c$. Hence we have:

$$\sum_{i=1}^{M} \sigma_i^1(\Phi_i^{\sigma^2} - \Psi_i^{\sigma^2}) = c\sum_{i=1}^{M} \sigma_i^1 = 0 \tag{59}$$

It shows that the value of $\langle\sigma^1, \sigma^2\rangle_\rho$ does not depends on the choice of the representative Φ^{σ^2}, hence it is well-defined.

To show $\langle\sigma^1, \sigma^2\rangle_\rho$ is a valid inner-product, we need to check conjugate symmetry, linearity in the first argument, and

positive-definiteness. For conjugate symmetry, we have:

$$\langle \sigma^2, \sigma^1 \rangle_\rho = \sum_{i=1}^M \sigma_i^2 \Phi_i^{\sigma^1} \tag{60}$$

$$= \sum_{i=1}^M \mathrm{div}_\rho(\nabla \Phi^{\sigma^1})_i \Phi_i^{\sigma^1} \tag{61}$$

$$= \sum_{i=1}^M \sum_{j \in N(i)} c_{ij}(\rho)(\Phi_i^{\sigma^1} - \Phi_j^{\sigma^1})\Phi_i^{\sigma^2} \tag{62}$$

$$= \frac{1}{2}\sum_{i=1}^M \sum_{j \in N(i)} c_{ij}(\rho)(\Phi_i^{\sigma^1} - \Phi_j^{\sigma^1})\Phi_i^{\sigma^2} + \frac{1}{2}\sum_{j=1}^M \sum_{i \in N(j)} c_{ij}(\rho)(\Phi_i^{\sigma^1} - \Phi_j^{\sigma^1})\Phi_i^{\sigma^2} \tag{63}$$

$$= \frac{1}{2}\sum_{i=1}^M \Phi_i^{\sigma^2} \sum_{j \in N(i)} c_{ij}(\rho)(\Phi_i^{\sigma^1} - \Phi_j^{\sigma^1}) + \frac{1}{2}\sum_{j=1}^M \Phi_j^{\sigma^2} \sum_{i \in N(j)} c_{ij}(\rho)(\Phi_i^{\sigma^1} - \Phi_j^{\sigma^1}) \tag{64}$$

$$= \frac{1}{2}\sum_{i=1}^M \Phi_i^{\sigma^2} \sum_{j \in N(i)} c_{ij}(\rho)(\Phi_i^{\sigma^1} - \Phi_j^{\sigma^1}) + \frac{1}{2}\sum_{i=1}^M \Phi_i^{\sigma^2} \sum_{j \in N(i)} c_{ij}(\rho)(\Phi_j^{\sigma^1} - \Phi_i^{\sigma^1}) \tag{65}$$

$$= \frac{1}{2}\sum_{(i,j) \in E} c_{ij}(\rho)(\Phi_i^{\sigma^1} - \Phi_j^{\sigma^1})(\Phi_i^{\sigma^2} - \Phi_j^{\sigma^2}) \tag{66}$$

We can see that (66) does not depend on the order of σ^1 and σ^2, hence we have:

$$\langle \sigma^1, \sigma^2 \rangle = \langle \sigma^2, \sigma^1 \rangle \tag{67}$$

The linearity for the first argument is trivial to see. For positive-definiteness, if we have $\langle \sigma, \sigma \rangle_\rho = 0$, then by (66), we have:

$$\sum_{(i,j) \in E} c_{ij}(\rho)(\Phi_i^\sigma - \Phi_j^\sigma)^2 = 0 \tag{68}$$

Since by our assumption, G is connected and $c_{ij}(\rho) > 0$, it indicates $\sigma = \nabla \Phi^\sigma = 0$. Finally, from (66), we can see that:

$$\langle \sigma^1, \sigma^2 \rangle_\rho = \langle \nabla \Phi^{\sigma^1}, \nabla \Phi^{\sigma^2} \rangle_\rho \tag{69}$$

This means the inner-product we defined in (46) is compatible with the immersion ζ.

A.4 Proof for Theorem 3.1

We prove the theorem in a more general form in terms of free energy $F : \mathcal{P}_2(V) \to \mathbb{R}$. Once we find the gradient flow for F, theorem 3.1 can be seen as a special case. In particular, when we define

$$F(\rho) = \sum_{i=1}^M \rho_i f_i - \sum_{i=1}^M \rho_i \log \rho_i \tag{70}$$

we have $F(\rho) = D_{\mathrm{KL}}(\rho || \pi)$.

Proof. The gradient flow in terms of the free energy is:

$$\frac{d\rho}{dt} = -\nabla_\rho F(\rho) \tag{71}$$

By lemma A.2, for any $\sigma \in T_\rho \mathcal{P}_2(V)$, we have $\Phi^\sigma \in \zeta^{-1}(\sigma)$ such that $\sigma = \mathrm{div}_\rho(\Phi^\sigma)$. On the left hand side, we have:

$$\langle \frac{d\rho}{dt}, \sigma \rangle = \sum_{i=1}^M \frac{d\rho_i}{dt} \Phi_i^\sigma \tag{72}$$

On the right hand side, we have:

$$\langle \nabla_\rho F(\rho), \sigma \rangle = \sum_{i=1}^{M} \frac{\partial F(\rho)}{\partial \rho_i} \sum_{j \in N(i)} c_{ij}(\rho)(\Phi_i^\sigma - \Phi_j^\sigma) \tag{73}$$

$$= \sum_{i=1}^{M} \frac{\partial F(\rho)}{\partial \rho_i} \sum_{j \in N(i)} c_{ij}(\rho)\Phi_i^\sigma - \sum_{i=1}^{M} \frac{\partial F(\rho)}{\partial \rho_i} \sum_{j \in N(i)} c_{ij}(\rho)\Phi_j^\sigma \tag{74}$$

$$= \sum_{i=1}^{M} \Phi_i^\sigma \sum_{j \in N(i)} c_{ij}(\rho)\frac{\partial F(\rho)}{\partial \rho_i} - \sum_{j=1}^{M} \Phi_j^\sigma \sum_{i \in N(j)} c_{ij}(\rho)\frac{\partial F(\rho)}{\partial \rho_i} \tag{75}$$

$$= \sum_{i=1}^{M} \Phi_i^\sigma \sum_{j \in N(i)} c_{ij}(\rho)\frac{\partial F(\rho)}{\partial \rho_i} - \sum_{i=1}^{M} \Phi_i^\sigma \sum_{j \in N(i)} c_{ij}(\rho)\frac{\partial F(\rho)}{\partial \rho_j} \tag{76}$$

$$= \sum_{i=1}^{m} \left(\sum_{j \in N(i)} c_{ij}(\rho)\big(\frac{\partial F(\rho)}{\partial \rho_i} - \frac{\partial F(\rho)}{\partial \rho_j}\big) \right) \Phi_i^\sigma \tag{77}$$

Hence we have:

$$\sum_{i=1}^{M} \frac{d\rho_i}{dt}\Phi_i^\sigma = -\sum_{i=1}^{m} \left(\sum_{j \in N(i)} c_{ij}(\rho)\big(\frac{\partial F(\rho)}{\partial \rho_i} - \frac{\partial F(\rho)}{\partial \rho_j}\big) \right) \Phi_i^\sigma \tag{78}$$

holds for arbitrary Φ^σ. Then we prove:

$$\frac{d\rho_i}{dt} = \sum_{j \in N(i)} c_{ij}(\rho) \left(\frac{\partial F(\rho)}{\partial \rho_j} - \frac{\partial F(\rho)}{\partial \rho_i} \right) \tag{79}$$

Plug in the value of free energy in (70), we have:

$$\frac{d\rho_i}{dt} = \sum_{j \in N(i)} c_{ij}(\rho)[f_j + \log \rho_j - f_i - \log \rho_i] \tag{80}$$

Thus we prove the theorem. □

B SAMPLER DETAILS

We discuss different discretizations for DLD. Since we decompose the Markov jump process into independent sub-processes X_n^t, the distribution satisfies the following equation:

$$\frac{d}{dt}\rho_n^t = \rho_n^t Q_n(x^t), \tag{81}$$

where x^t is the current state. The rate matrix Q_n satisfies:

$$Q_n(x^t)(i,j) = \begin{cases} w_{ij} g\left(\frac{\pi(x^t_{\backslash n}, j)}{\pi(x^t_{\backslash n}, i)}\right), & i \neq j \\ -\sum_{k \neq i} w_{ik} g\left(\frac{\pi(x^t_{\backslash n}, k)}{\pi(x^t_{\backslash n}, i)}\right), & i = j \end{cases} \tag{82}$$

Accordingly, the transition matrix satisfies:

$$\frac{d}{dh}P_n^h(x^t) = P_n^h Q_n(x^t), \quad P_n^0(x^t) = I_C \tag{83}$$

which has the following closed form solution:

$$P_n^h(x^t) = \exp(\int_0^h Q_n(x^t)ds) = \exp(Q_n(x^t)h) \tag{84}$$

To simplify the notation, we drop x^t, n, and only use $Q(i,j), P^h(i,j)$ if it does not cause ambiguity.

B.1 Single Jump

We show that the first jump satisfies a categorical distribution. We define the holding time

$$S_i := \inf_t \{t > 0 : X(0) = i, X(t) \neq i\} \tag{85}$$

as the time the process $X(t)$ stays at state x. We first show that S_x is memoryless. That's to say:

$$\mathbb{P}(S_i > r + t | S_i > t, X(0) = i) = \mathbb{P}(S_i > r + t | X(t) = i) = \mathbb{P}(S_i > r | X(0) = i) \tag{86}$$

Since the only continuous memoryless distribution is exponential distribution, we know S_x satisfies an exponential distribution $\lambda e^{-\lambda t}$. To estimate λ, we have:

$$\lambda = -\frac{d}{dt}|_{t=0} e^{-\lambda t} = -\frac{d}{dt}|_{t=0} \mathbb{P}(S_i > t | X(0) = i) = \lim_{h \to 0} \frac{1 - \mathbb{P}(S_i > h | X(0) = i)}{h} \tag{87}$$

$$= \lim_{h \to 0} \frac{1 - \mathbb{P}(X(h) = i | X(0) = i) + o(h)}{h} = \lim_{h \to 0} \frac{1 - (1 + q_{ii}h + o(h))}{h} = -q_{ii} \tag{88}$$

To derive the transition probability, we condition on the holding time belongs to a small interval $(t, t + h]$, and let $h \to 0$, we have:

$$\tilde{P}_{ij} = \lim_{h \to 0} \mathbb{P}(X(t + h) = j | X(0) = i, t < S_i \leq t + h) \tag{89}$$

$$= \lim_{h \to 0} \mathbb{P}(X(t + h) = j | X(t) = i, t < S_i \leq t + h) \tag{90}$$

$$= \lim_{h \to 0} \frac{\mathbb{P}(X(t + h) = j | X(t) = i)}{\mathbb{P}(t < S_i \leq t + h | X(t) = i)} \tag{91}$$

$$= \lim_{h \to 0} \frac{q_{ij}h + o(h)}{-q_{ii}h + o(h)} \tag{92}$$

$$= -\frac{q_{ij}}{q_{ii}} \tag{93}$$

Hence, the first jump that leaves state i to $j \in N(i)$ satisfies the multinomial distribution $q(i, j) \propto w_{ij} g(\pi_j / \pi_i)$.

B.2 DLMCf

One of the most straightforward methods to estimate P^h is forward Euler's method. Specifically, we can let:

$$\tilde{P}_f^h = P^0 + h \frac{d}{dh} P^0 = P^0 + hQ \tag{94}$$

where we use f to indicate the method is based on forward Euler. The transition matrix can be written as:

$$\tilde{P}_f^h = \begin{bmatrix} 1 - h\sum_{j \neq 1} Q(1, j) & hQ(1, 2) & \cdots & hQ(1, C) \\ hQ(2, 1) & 1 - h\sum_{j \neq 2} Q(2, j) & \cdots & hQ(2, n) \\ \vdots & \vdots & \ddots & \vdots \\ hQ(n, 1) & hQ(n, 2) & \cdots & 1 - h\sum_{j \neq C} Q(C, j) \end{bmatrix} \tag{95}$$

One constraint we should take care of is that we need to restrict the simulation time h such that the diagonal of $\tilde{P}^h$ is always non-negative.

Comparison with DMALA The DLMCf can be seen as a correction of DMALA (Zhang et al., 2022). Specifically, choosing $h = \exp(-\frac{1}{2})$, the transition matrix $\tilde{P}^h_{\text{DMALA}}$ of DMALA has the same off-diagonal value as $\tilde{P}_f^h$. However, the diagonal of the $\tilde{P}^h_{\text{DMALA}}$ is always 1. This systematic mismatch reduces the efficiency DMALA. After correction, DLMCf has substantial improvements in efficiency. See more results in Appendix C.

B.3 DLMC

We denote the stationary distribution for X_n^t as

$$\nu_n(x^t)(j) = \frac{\pi(x^t_{\setminus n}, j)}{\sum_{i\in\mathcal{C}} \pi(x^t_{\setminus n}, i)} \tag{96}$$

Again, we drop x^t and only use $\nu(j)$ if it does not cause ambiguity. The transition matrix for (39) can be written as:

$$\tilde{P}_n^h = \begin{bmatrix} \nu(1) + \sum_{j\neq 1}\nu(j)e^{-h\frac{Q(1,j)}{\nu(j)}} & \nu(2) - \nu(2)e^{-h\frac{Q(1,2)}{\nu(2)}} & \cdots & \nu(C) - \nu(C)e^{-h\frac{Q(1,C)}{\nu(C)}} \\ \nu(1) - \nu(1)e^{-h\frac{Q(2,1)}{\nu(1)}} & \nu(2) + \sum_{j\neq 2}\nu(j)e^{-h\frac{Q(2,j)}{\nu(j)}} & \cdots & \nu(C) - \nu(C)e^{-h\frac{Q(2,C)}{\nu(C)}} \\ \vdots & \vdots & \ddots & \vdots \\ \nu(1) - \nu(1)e^{-h\frac{Q(n,1)}{\nu(1)}} & \nu(2) - \nu(2)e^{-h\frac{Q(n,2)}{\nu(2)}} & \cdots & \nu(C) + \sum_{j\neq C}\nu(j)e^{-h\frac{Q(C,j)}{\nu(C)}} \end{bmatrix} \tag{97}$$

We can notice that when $C = 2$, the estimation

$$\tilde{P}^h = \begin{bmatrix} \nu(1) + \nu(2)e^{-h\frac{Q(1,2)}{\nu(2)}} & \nu(2) - \nu(2)e^{-h\frac{Q(1,2)}{\nu(2)}} \\ \nu(1) - \nu(1)e^{-h\frac{Q(2,1)}{\nu(1)}} & \nu(2) + \nu(1)e^{-h\frac{Q(2,1)}{\nu(1)}} \end{bmatrix} = P_n^h \tag{98}$$

is exact. For $C > 2$, We have:

$$\tilde{P}_n^0 = I_C = P_n^0, \quad \tilde{P}_n^\infty = \nu^T 1 = P_n^\infty, \quad \frac{d}{dh}\tilde{P}_n^h|_{h=0} = Q_n = \frac{d}{dh}\tilde{P}_n^h|_{h=0} \tag{99}$$

Comparison with PAS The PAS (Sun et al., 2021a, 2022) flips a given number of sites R per M-H step. As a results, it is equivalent with simulating the DLD via non-uniform time slice. Specifically, when the current state x^t has small jump rate Q, to flip R sites, PAS needs to simulate a longer time $h_+ > h$ in this M-H step. On the contrary, when the current state x^t has large jump rate Q, PAS needs to simulate a shorter time $h_- < h$ in this M-H step. Consequently, the Markov chain obtained by PAS is more self correlated than DLMC. Also, since the PAS chain is likely to sample more frequently at the states with larger jump rates, M-H test need to reject more proposals to guarantee the chain is π-reversible. As a result, DLMC will be more efficient than PAS; see results in Appendix C. One disadvantage of DLMC is that, the simulation time h needed for transient and stationary phases are very different (Christensen et al., 2005), which makes tuning the scaling via average acceptance rate less robust comparing to PAS.

B.4 Choice of Conductance

Inspired by physics, we can define the conductance as

$$c_{ij}(\rho) = \frac{m_{ij}(\rho) - m_{ji}(\rho)}{\log(m_{ij}(\rho)) - \log(m_{ji}(\rho))}, \ \forall j \neq i. \tag{100}$$

The logarithmic mean in (100) is known as *conductance* in the stoichiometric network theory of chemical reactions (Qian & Beard, 2005), where m_{ij} represents the amount of the transition from i to j, such that the numerator is the flux and the denominator is the driving force in nonequilibrium systems (Beard & Qian, 2007). We assume that the amount of the transition

$$m_{ij}(\rho) = w_{ij}g(\frac{\pi_j}{\pi_i})\rho_i, \ \forall j \neq i, \tag{101}$$

is only determined by the transition speed $w_{ij}g(\frac{\pi_j}{\pi_i})$ multiplying the current amount ρ_i. Here, w_{ij} satisfies $w_{ij} = w_{ji}$ as an inherent scalar that measure the variability between i and j, independent with both the target distribution π and current distribution ρ. The second term $g(\frac{\pi_j}{\pi_i})$ is an external force caused by the target probability ratio and also relies on the choice of the weight function $g(\cdot)$. Furthermore, a reasonable assumption is that the transition should reach the equilibrium at the target distribution:

$$m_{ij}(\pi) = m_{ji}(\pi). \tag{102}$$

Plug (101) into (102), one can solve

$$g(\frac{\pi_j}{\pi_i}) = \frac{\pi_j}{\pi_i}g(\frac{\pi_i}{\pi_j}) \ \Rightarrow \ g(t) = tg(\frac{1}{t}), t > 0, \tag{103}$$

which is exactly the *locally balanced* (LB) function used in recent locally balanced samplers (Zanella, 2020). Plug (101) and (103) in (100), one can rewrite the conductance as:

$$c_{ij}(\rho) = w_{ij} \frac{g(\pi_j/\pi_i)\rho_i - g(\pi_i/\pi_j)\rho_j}{f_i + \log \rho_i - f_j - \log \rho_j}. \tag{104}$$

C EXPERIMENTAL DETAILS

We focus on discrete spaces of the form $V = \mathcal{X}^D$ where $\mathcal{X} = \{e_1, ..., e_n\}$ is a finite set of one-hot vectors. We evaluate our methods on Bernoulli model, Ising model, factorial hidden Markov model and restricted Boltzmann machine. For each model, we consider both binary and categorical versions. For binary model, we use one high temperature setting and one low temperature setting. For categorical model, we use $n = 4$ and $n = 8$. We report the detailed descriptions of the models and corresponding results in the following. The running time for all methods across all models are summarized in Table 3.

Table 3: Running time (second) for all samplers on all target distributions with 100k steps

Mehtod	hb-10-1	bg-2	rwm	gwg	dmala	pas	dlmc	dlmcf
Bernoulli (low)	144	61	76	231	153	406	213	149
Bernoulli (high)	143	61	73	197	156	245	217	150
Bernoulli ($n = 4$)	351	112	500	115	150	514	184	170
Bernoulli ($n = 8$)	794	458	501	121	161	526	231	194
Ising (low)	203	101	149	491	490	557	548	476
Ising (high)	205	106	158	519	514	589	584	507
Potts ($n = 4$)	484	198	412	409	452	824	494	476
Potts ($n = 8$)	1335	1022	428	416	451	832	486	470
binFHMM (low)	216	141	228	490	476	555	542	469
binFHMM (high)	216	141	228	495	475	559	549	474
catFHMM ($n = 4$)	450	204	492	398	433	800	470	450
catFHMM ($n = 8$)	1490	1168	499	393	436	804	475	456
binRBM	144	83	105	225	235	304	296	229
binRBM	142	82	102	229	234	305	298	229
catRBM ($n = 4$)	783	357	369	195	236	590	269	255
catRBM ($n = 8$)	2721	2283	389	276	304	684	322	313

C.1 Bernoulli Model

The Bernoulli distribution is the simplest distribution in a discrete space, where each site is independent with others. For $x \in \mathcal{C}^N$, the energy function is:

$$f(x) = \sum_{n=1}^{N} \langle x_n, \theta^d \rangle \tag{105}$$

where $\theta^d \in \mathbb{R}^C$. Across all settings, we use the entries in θ_n independently sampled from centered normal distribution $\mathcal{N}(0, \sigma^2)$. For binary model we consider $D = 10000$. We use $\sigma^2 = 0.125$ in the high temperature setting and $\sigma^2 = 12.5$ in the low temperature setting. For categorical model, we consider $D = 2000$. We use $\sigma^2 = 1.125$ for both $n = 4$ and $n = 8$. The results are reported in Figure 3 and Figure 4. We can see that DLMC and DLMCf have substantial better efficiencies compared to other samplers. The weight function $g(t) = \frac{t}{t+1}$ has better performance compared to $g(t) = \sqrt{t}$ as proved in Zanella (2020). Moreover, the advantage of $g(t) = \frac{t}{t+1}$ is more significant when the target distributions are sharper, which is consistent with the observation in continuous space (Livingstone & Zanella, 2019).

C.2 Ising Model

The Ising model (Ising, 1924) is a mathematical model of ferromagnetism in statistical mechanics. It consists of binary random variables arranged in a lattice graph $G = (V, E)$ and allows node to interact with its neighbors. The Potts model

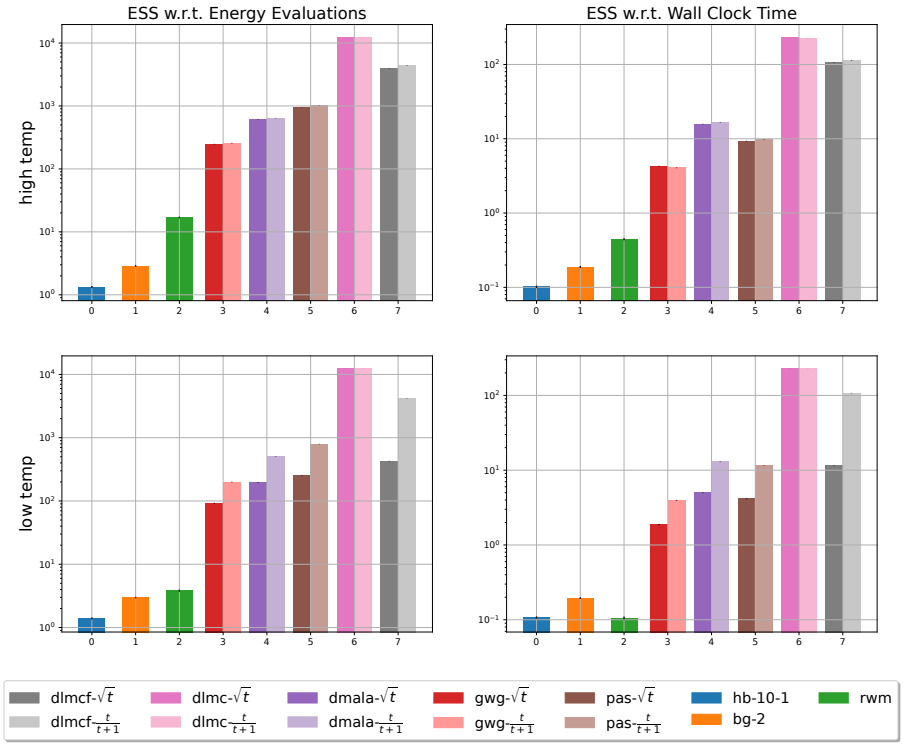


Figure 3: Evaluation on Bernoulli Models

(Potts, 1952) is a generalization of the Ising model where the random variables are categorical. The energy function for Ising model and Potts model can be described as:

$$f(x) = -\sum_{n=1}^{N} \langle x_n, \theta_n \rangle - \lambda \sum_{(i,j)\in E} \delta(x_i, x_j) \tag{106}$$

where $\theta^d \in \mathbb{R}^n$, $\delta(x, y) = 1_{\{x=y\}}$. For Ising model, we consider $N = 2500$ where G is a 50×50 square lattice, and we follow the settings in Zanella (2020). In high temperature setting, we use $\theta^d \sim$ uniform$(-2, 1)$ for the outer part of the lattice graph, and $\theta^d \sim$ uniform$(-1, 2)$ for the inner part of the lattice graph. The connection strength is chosen as $\lambda = 0.5$. In low temperature setting, we use $\theta^d \sim$ uniform$(-4, 2)$ for the outer part of the lattice graph, and $\theta^d \sim$ uniform$(-2, 4)$ for the inner part of the lattice graph. The connection strength is chosen as $\lambda = 1.0$. For potts model, we consider $N = 900$ where G is a 30×30 square lattice. For both $C = 4, 8$, we use entries in external field $\theta_i^d \sim$ uniform$(-1.5, 1.5) - 0.5\frac{i}{C}$ for the outer part of the lattice graph, and $\theta_i^d \sim$ uniform$(-1.5, 1.5) + 0.5\frac{i}{C}$ for the inner part of the lattice graph, where $i = 1, ..., C$. The connection strength is chosen as $\lambda = 1.0$. The results are reported in Figure 5 and Figure 6. We can see that all LB samplers exhibit good performance. Among them, DLMC and DLMCf are the most efficient. The weight functions $g(t) = \sqrt{t}$ and $g(t) = \frac{t}{1+t}$ each demonstrate advantages for different samplers.

C.3 Factorial Hidden Markov Model

FHMM (Ghahramani & Jordan, 1995) uses latent variables to characterize time series data. In particular, it assumes the continuous data $y \in \mathbb{R}^L$ is generated by hidden state $x \in \mathcal{C}^{L\times K}$. The probability function is:

$$p(x) = p(x_1)\prod_{l=2}^{L} p(x^t|x^{t-1}), \quad p(y|x) = \prod_{l=1}^{L} \mathcal{N}(y_t; \sum_{k=1}^{K} \langle W_k, x_{l,k} \rangle + b; \sigma^2) \tag{107}$$

In particular, for binary model, we consider $\mathbb{P}(x_1 = 0) = 0.9, \mathbb{P}(x^t = x^{t-1}|x^{t-1}) = 0.8$, $\sigma = 2.0$. We use $L = 200, K = 10$ for high temperature setting and $L = 100, K = 20$ in low temperature setting. For categorical model, we use $p(x_1|x_1 \neq 0)$ and $p(x^t|x^{t-1}, x^t \neq x^{t-1})$ as uniform distribution and we use $L = 200$, $K = 10$. We report the results

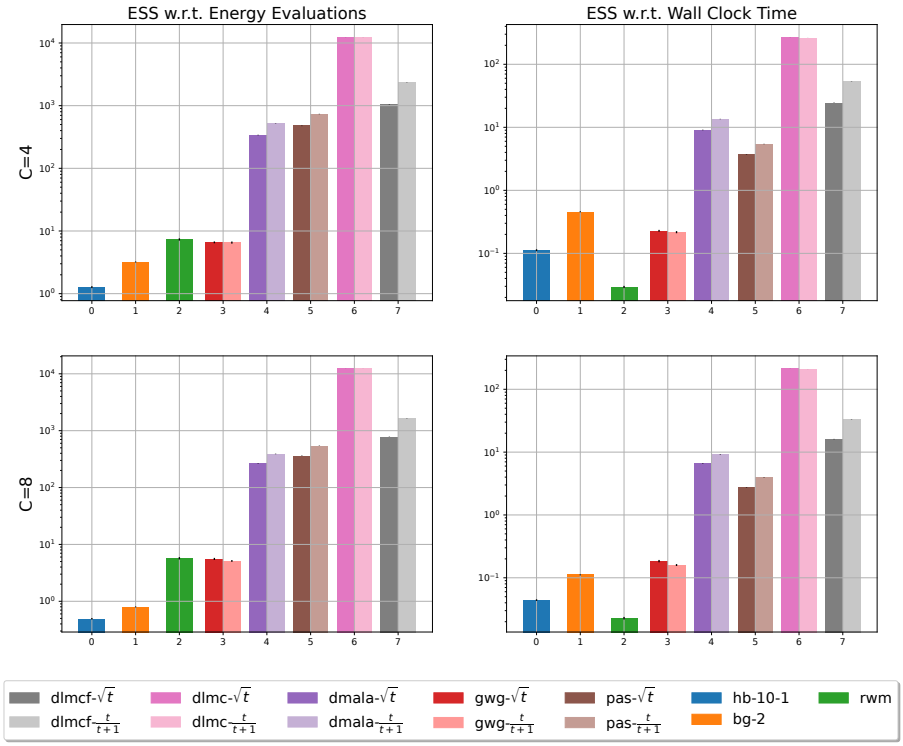


Figure 4: Evaluation on Categorical Models

in Figure 7 and Figure 8. Similar to the Ising model, we can see that all locally balanced samplers demonstrate good performance. In FHMM, LBP has an efficiency very close to the DLMC samplers. We believe this is because the energy change rate is stable in FHMM and the magnitude of the gradient changes steadily. Hence the Hamming distance works as a good metric. We also note that the weight function $g(t) = \sqrt{t}$ is systematically better than $g(t) = \frac{t}{t+1}$ on FHMM. This is consistent with the observation in Livingstone & Zanella (2019) that $g(t) = \sqrt{t}$ performs better on smooth target distributions and $g(t) = \frac{t}{t+1}$ performs better on nonsmooth target distributions, although Livingstone & Zanella (2019) focus on the sampling in continuous spaces.

C.4 Restricted Boltzmann Machine

The RBM is an unnormalized latent variable model, with a visible random variable $v \in \mathcal{C}^N$ and a hidden random variable $h \in \{0, 1\}^M$. When v is binary, we call it a binary RBM (binRBM) and when v is categorical, we call it a categorical RBM (catRBM). The energy function of both binRBM and catRBM (Tran et al., 2011) can be written as:

$$f(v) = \sum_h \left[-\sum_{n=1}^{N} \langle v_n, \theta_n \rangle - \sum_{m=1}^{M} \beta_m h_m - \sum_{d,m} \langle h_m \theta_{m,d}, v_n \rangle \right] \tag{108}$$

Unlike the previous three models, where the parameters are hand designed, we train binary RBM on MNIST (LeCun et al., 1998) and categorical RBM on Fashion-MNIST (Xiao et al., 2017) using contrastive divergence Hinton (2002). Across all settings, we have $D = 784$. For binary models, we use $M = 25$ for high temperature setting and $M = 200$ for low temperature setting. For categorical models, we use $M = 100$. We report the results in Figure 9 and Figure 10. The learned RBMs have stronger multi-modality compared to previous models. We can see that, as before, DLMC and DLMCf lead in proposal quality, while DLMC is the most efficient overall.

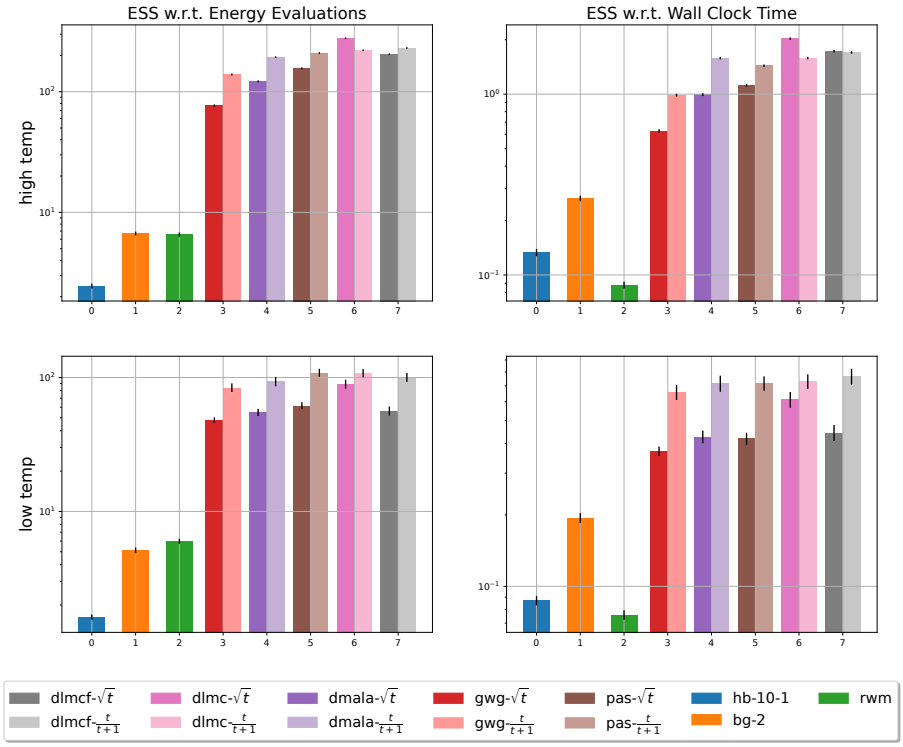


Figure 5: Evaluation on Ising Models

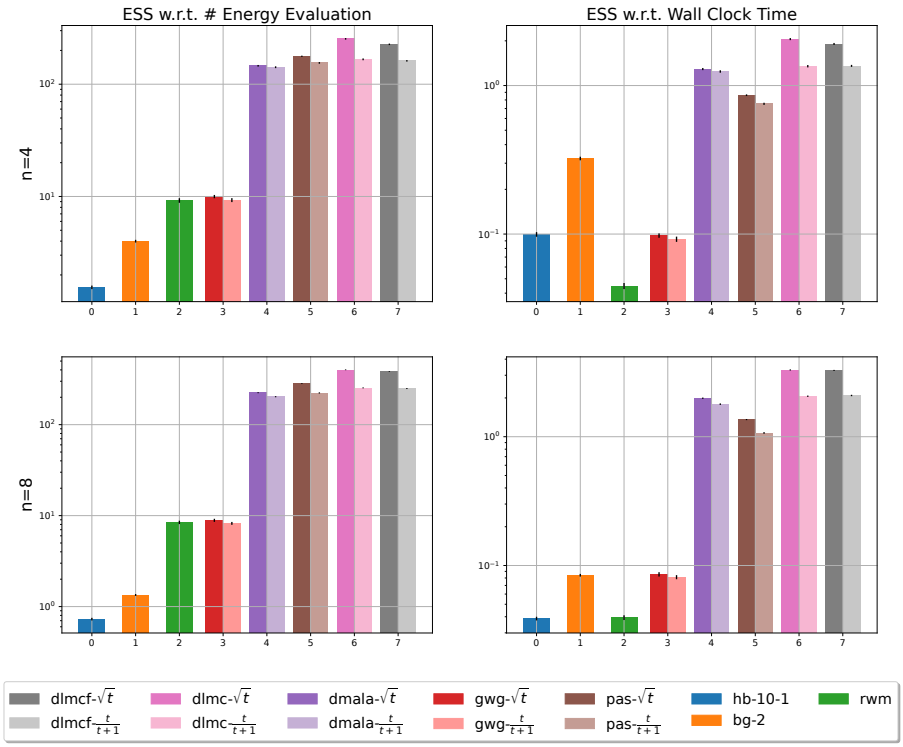


Figure 6: Evaluation on Potts Models

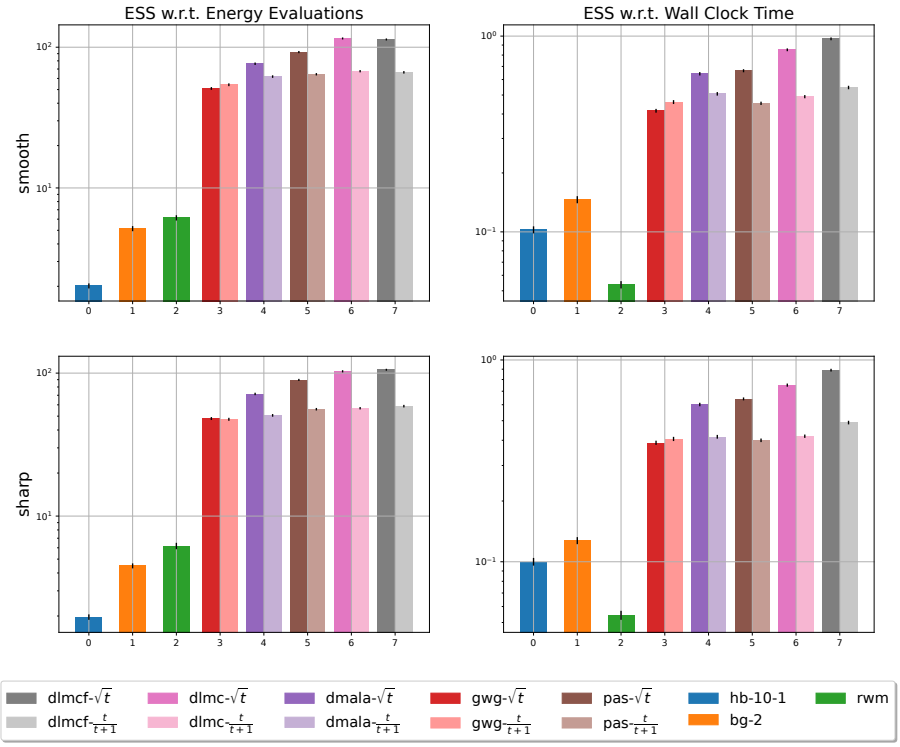


Figure 7: Evaluation on binFHMM

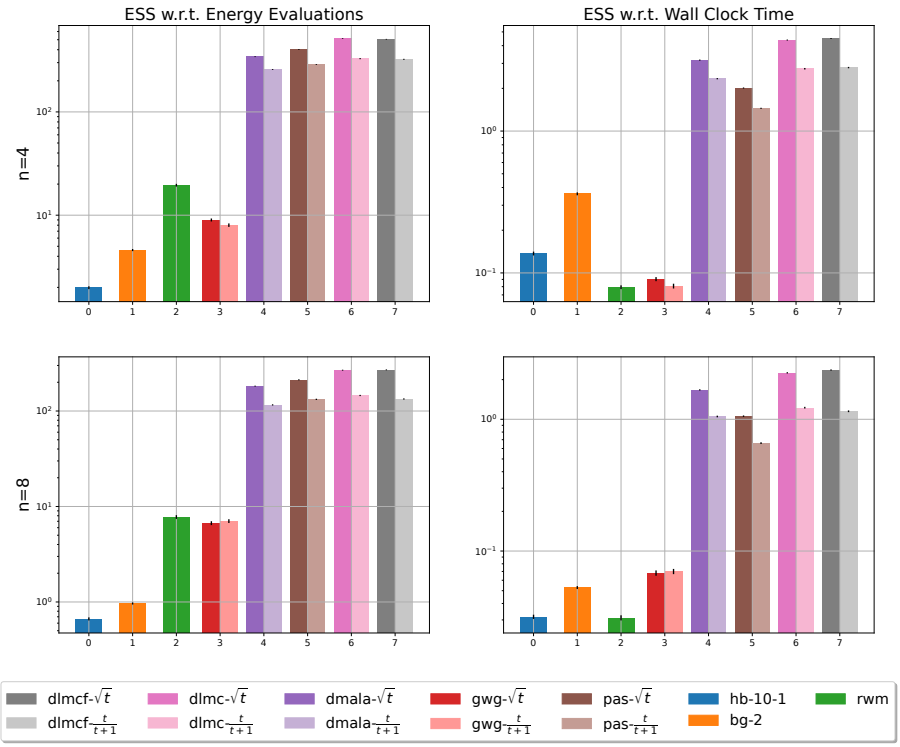


Figure 8: Evaluation on catFHMM

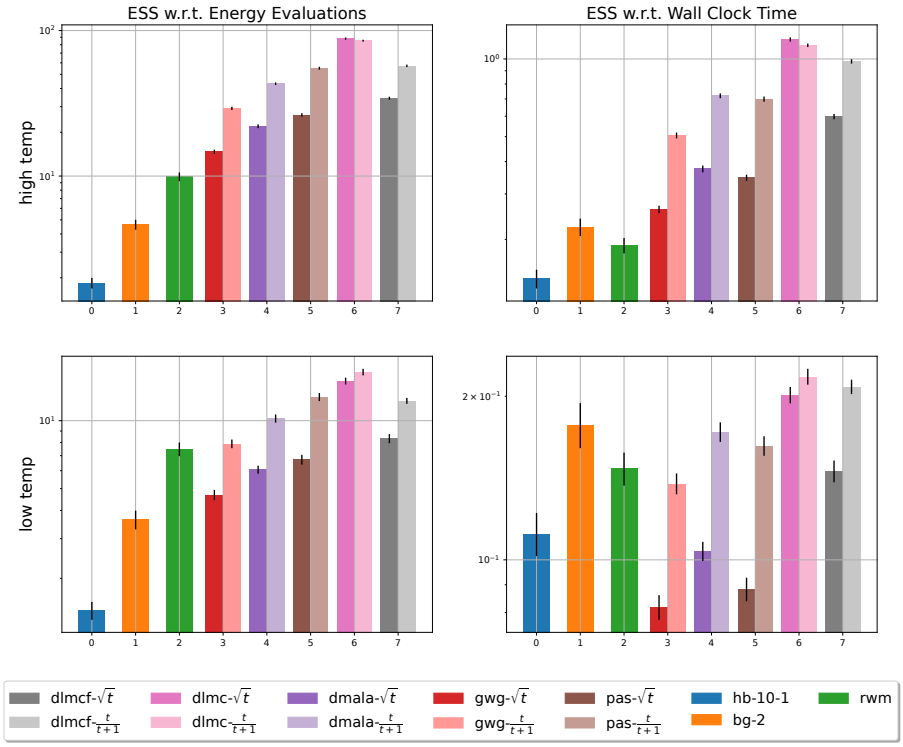


Figure 9: Evaluation on binRBM

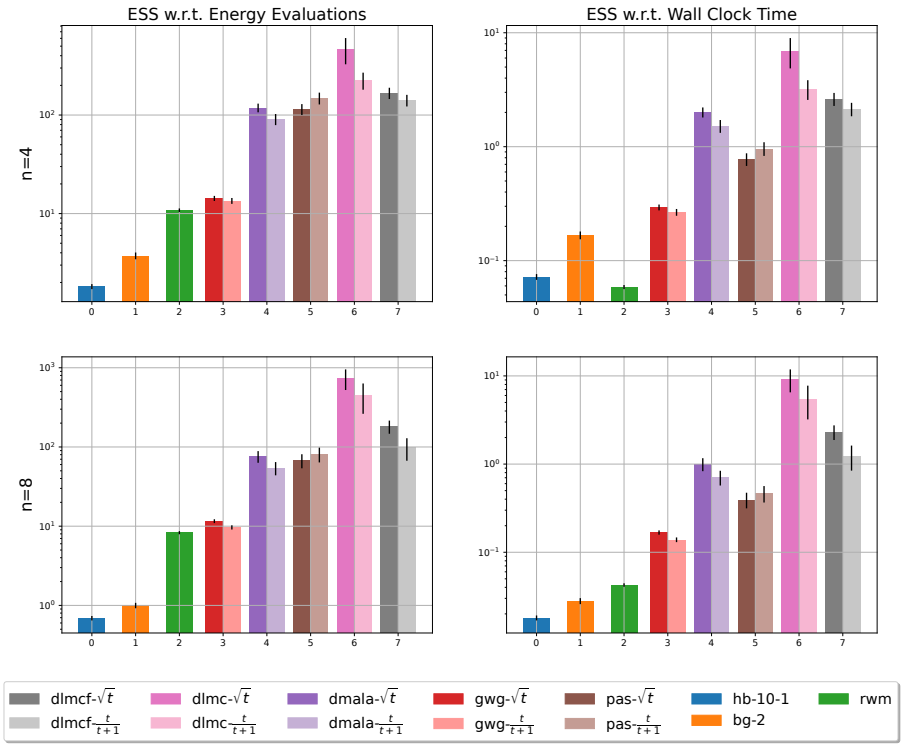


Figure 10: Evaluation on catRBM