Published at ProbML 2026

Accelerating Discrete Langevin Samplers via Continuous Intermediates

Guangyu Li[1] **Ruqi Zhang**[1,†]

[1]Department of Computer Science, Purdue University
[†]Correspondence to ruqiz@purdue.edu

Abstract

Sampling from discrete distributions remains a challenge in machine learning, with traditional Markov Chain Monte Carlo (MCMC) methods such as Gibbs sampling suffering from inefficiency due to single-coordinate updates. Recent gradient-based discrete samplers have improved performance but remain constrained by the original discrete structures, which potentially hinder the convergence. To address this issue, we propose a hybrid approach that enables more informed proposals by introducing a continuous exploratory intermediate before the discrete update. This method, called Discrete Langevin Samplers via Continuous intermediates (cDLS), bridges the gap between discrete and continuous sampling and significantly accelerates convergence while maintaining theoretical guarantees. We develop variants of cDLS to ensure broad applicability, including unadjusted and Metropolis-adjusted versions. Experiments on Ising models, restricted Boltzmann machines, deep energy-based models, and Bayesian binary neural networks validate the superior performance of cDLS compared to existing methods. Our results highlight the potential of hybrid continuous-discrete exploration for advancing general discrete sampling.

1. Introduction

Discrete random variables are pervasive, from text data to complex genetic sequences (Eddy, 2004; Wainwright et al., 2008; Vaswani et al., 2017). With the rapid progress of machine learning, the demand for effective sampling algorithms tailored to discrete distributions has become both urgent and essential. Markov Chain Monte Carlo (MCMC) remains a cornerstone for sampling from complex distributions, where its effectiveness highly depends on the proposal distribution. Among MCMC methods, Gibbs sampling (Geman and Geman, 1984) is particularly popular for discrete distributions, yet its efficiency suffers in high-dimensional settings due to its coordinate-wise update scheme. To address this limitation, recent works have incorporated gradient information into discrete sampling, enabling more effective exploration and faster convergence (Grathwohl et al., 2021; Zhang et al., 2022; Sun et al., 2022b, 2023a,b; Xiang et al., 2023; Pynadath et al., 2024). These methods have demonstrated strong empirical performance and important theoretical guarantees, advancing the state of the art in high-dimensional discrete sampling.

Although existing gradient-based discrete samplers benefit from differentiable extensions of the energy function, their updates remain confined to the discrete domain, preventing full exploitation of the underlying smooth geometry. To fully harness smooth structure, we propose a hybrid method, *Discrete Langevin Samplers via Continuous Intermediates* (cDLS), which introduces continuous intermediates to enhance gradient-based discrete samplers, including discrete Langevin methods (Zhang et al., 2022).

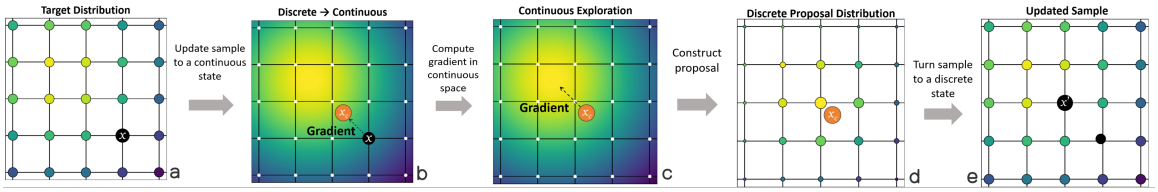


Figure 1. Overview of the proposed cDLS. Starting from a discrete state, the sampler performs a continuous update using gradient, then constructs a discrete proposal using discrete Langevin, and finally transitions to a new discrete state.

By bridging discrete and continuous domains, cDLS leverages geometric insights from continuous Langevin dynamics while preserving exact discrete updates. This hybrid design enables geometry-aware large moves and faster mixing, substantially improving convergence across various experiments. A visualization of this process is shown in Figure 1. Notably, our framework is general and can be integrated with a wide range of gradient-based discrete samplers. For clarity and brevity, we primarily focus our discussion on the Discrete Langevin Sampler[1], which serves as a representative example in the domain of discrete gradient-based sampling.

We summarize our contributions as follows:

- We propose discrete Langevin samplers via continuous intermediates (cDLS), a novel hybrid sampling framework that leverages continuous exploration to accelerate discrete Langevin sampling.
- We establish theoretical guarantees for cDLS, including asymptotic correctness for log-quadratic distributions without Metropolis–Hastings correction, controlled bias under near-quadratic deviations, and non-asymptotic convergence and inference guarantees for the MH-adjusted variant.
- We demonstrate the effectiveness of our proposed algorithm on several experiments, including Ising models, restricted Boltzmann machines, deep energy-based models, and binary Bayesian neural networks. Notably, cDLS can achieve strong performance and low bias even without the use of Metropolis-Hastings correction.

2. Preliminaries

We consider sampling from a discrete target distribution of the form:

$$\pi(\theta) = \frac{1}{Z}\exp(U(\theta)), \quad \theta \in \Theta, \tag{1}$$

where θ is a d-dimensional variable, Θ is a finite discrete variable domain, U is the energy function, and Z is the normalizing constant for π to be a distribution. Throughout this paper, we adopt standard assumptions commonly used in gradient-based discrete sampling methods (Grathwohl et al., 2021; Sun et al., 2022b; Zhang et al., 2022): (1) the energy function U admits a differentiable extension to $\mathbb{R}^d$; (2) the discrete domain factorizes coordinate-wise as $\Theta = \prod_{i=1}^{d} \Theta_i$.

1. Our framework is general and can be readily applied to other gradient-based discrete samplers, such as Gibbs with Gradients (GWG) (Grathwohl et al., 2021). Due to space constraints, we present these extensions in Appendix G.

Langevin Dynamics in Continuous Space Langevin dynamics provide a canonical gradient-based sampling mechanism in continuous domains. For a target density proportional to $\exp(U(\theta))$ with $\theta \in \mathbb{R}^d$, the Langevin stochastic differential equation is given by

$$\mathrm{d}\theta_t = \nabla U(\theta_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \tag{2}$$

where W_t denotes standard Brownian motion. A first-order Euler discretization with stepsize α yields the unadjusted Langevin algorithm (ULA),

$$\theta_{t+1} = \theta_t + \frac{\alpha}{2}\nabla U(\theta_t) + \sqrt{\alpha}\,\xi_t, \tag{3}$$

where $\xi_t \sim \mathcal{N}(0, I_{d\times d})$. This update can be interpreted as a Gaussian proposal centered at $\theta_t + \frac{\alpha}{2}\nabla U(\theta_t)$ with covariance αI, capturing both local gradient information and stochastic exploration.

Discrete Langevin Proposal To transfer the geometric intuition of Langevin dynamics to discrete spaces, the Discrete Langevin Proposal (DLP) (Zhang et al., 2022) defines a proposal distribution directly over Θ. Given a current state $\theta \in \Theta$, the proposal for $\theta' \in \Theta$ is

$$q(\theta' \mid \theta) = \frac{\exp\left(-\frac{1}{2\alpha}\left\|\theta' - \theta - \frac{\alpha}{2}\nabla U(\theta)\right\|_2^2\right)}{Z_\Theta(\theta)}, \tag{4}$$

where $Z_\Theta(\theta)$ is the normalizing constant over the discrete domain. Notably, if the proposal were defined on $\mathbb{R}^d$ rather than Θ, it would exactly recover the Gaussian proposal induced by the discretized Langevin dynamics, motivating the interpretation of DLP as a discrete analogue of Langevin sampling.

Under the coordinate-wise factorization assumption on Θ, the proposal further decomposes as

$$q(\theta_i' \mid \theta) \propto \exp\left(\frac{1}{2}\nabla U(\theta)_i(\theta_i' - \theta_i) - \frac{(\theta_i' - \theta_i)^2}{2\alpha}\right), \quad \theta_i' \in \Theta_i, \tag{5}$$

which allows parallel updates across dimensions. While DLP incorporates gradient information into discrete sampling, it operates purely within the discrete space and therefore cannot fully exploit the smooth structure of the continuous energy landscape implied by the differentiable extension of U. This observation motivates methods that bridge continuous and discrete dynamics, which we develop in the following sections.

3. Discrete Langevin Sampler via Continuous Intermediates

In this section, we introduce *Discrete Langevin Samplers via Continuous Intermediates* (cDLS), a general framework for improving discrete Langevin sampling by inserting a short, gradient-guided excursion into continuous space between discrete updates. The central idea is to temporarily lift the dynamics from the discrete domain into a continuous auxiliary space, and then induce a discrete transition informed by this continuous intermediate.

3.1. The Limitation of Purely Discrete Gradient-Based Samplers

Existing discrete gradient-based samplers (Grathwohl et al., 2021; Sun et al., 2023b; Zhang et al., 2022) incorporate gradient information by defining proposal distributions directly over the discrete domain. While this preserves exact discreteness, it fundamentally restricts how smooth geometric information of the energy function can be utilized (see Figure 1(a,d,e)).

In particular, discrete proposals are necessarily centered at a discrete state, and gradient information can only affect the relative probabilities assigned to nearby discrete configurations. Consequently,

even when the underlying energy function exhibits a strong, smooth structure in the ambient continuous space, the induced Markov chain is unable to follow gradient-aligned directions beyond the immediate discrete neighborhood, as illustrated in Figure 1(d).

This limitation becomes more pronounced in high-dimensional settings or in regions where the energy landscape is weakly curved, as informative descent or ascent directions may not align with discrete coordinate directions. As a result, purely discrete gradient-based samplers may suffer from slow exploration and reduced efficiency, despite access to gradient information (Figure 1(e)).

These observations suggest that the main bottleneck lies not in the absence of gradient information, but in the restriction that gradient-guided updates must be realized entirely within the discrete domain. This motivates the introduction of an auxiliary continuous intermediate that allows gradient information to shape the dynamics more directly, while ultimately preserving discrete transitions.

3.2. Continuous Intermediates via Gradient Flow

Given the current discrete state $\theta \in \Theta$, we construct a **continuous intermediate** by extending θ into $\mathbb{R}^d$ along the gradient of the energy function. Concretely, we perform a single gradient ascent step:

$$\theta^c_{\text{raw}} = \theta + \frac{\alpha_0}{2}\nabla U(\theta), \tag{6}$$

where $\alpha_0 > 0$ controls the strength of the continuous exploration. Since θ^c_{raw} may lie outside the relaxed feasible region, we project it onto a convex set $K \subset \mathbb{R}^d$ that contains Θ:

$$\theta^c = \text{Proj}_K(\theta^c_{\text{raw}}) = \arg\min_{y \in K} \|y - \theta^c_{\text{raw}}\|. \tag{7}$$

The choice of K is typically natural and problem-dependent. For example, when $\Theta = \{0,1\}^d$, we may take $K = [0,1]^d$. This projection prevents excessive drift away from the discrete domain and ensures that the continuous intermediate remains informative for subsequent discrete updates.

3.3. Discrete Updates Induced by Continuous Intermediates

Given the continuous intermediate θ^c, we construct a discrete proposal $\theta' \in \Theta$ by sampling from a distribution that is centered around θ^c and incorporates gradient information evaluated at this intermediate state. Intuitively, the continuous intermediate serves as a geometry-aware point from which discrete transitions are generated.

Specifically, we define the proposal distribution

$$q(\theta' \mid \theta^c) = \frac{\exp\left(-\frac{1}{2\alpha}\left\|\theta' - \theta^c - \frac{\alpha}{2}\nabla U(\theta^c)\right\|^2\right)}{Z_\Theta(\theta^c)}, \tag{8}$$

where $\alpha > 0$ controls the stochasticity of the discrete update and $Z_\Theta(\theta^c)$ denotes the normalizing constant over the discrete domain Θ.

When the discrete domain Θ factorizes across coordinates, the proposal further admits a product decomposition

$$q(\theta' \mid \theta^c) = \prod_{i=1}^{d} q_i(\theta'_i \mid \theta^c), \tag{9}$$

with coordinate-wise factors

$$q_i(\theta'_i \mid \theta^c) \propto \exp\left(\frac{1}{2}\nabla U(\theta^c)_i(\theta'_i - \theta^c_i) - \frac{(\theta'_i - \theta^c_i)^2}{2\alpha}\right), \quad \theta'_i \in \Theta_i. \tag{10}$$

This factorization enables efficient and parallelizable sampling across dimensions, while allowing discrete transitions to be guided by a smoother local geometry than would be accessible when centering proposals directly at the discrete state θ.

Importantly, the use of a continuous intermediate to mediate discrete proposals is not restricted to the specific Langevin-like form in Equation 8. The same principle can be applied to other gradient-based discrete samplers, such as Gibbs-with-Gradients (GWG) (Grathwohl et al., 2021), by replacing the underlying discrete proposal kernel with its continuous-intermediate counterpart. We discuss such extensions and Metropolis-Hastings (MH) correction (Metropolis et al., 1953; Hastings, 1970) in Appendix G and Appendix C.

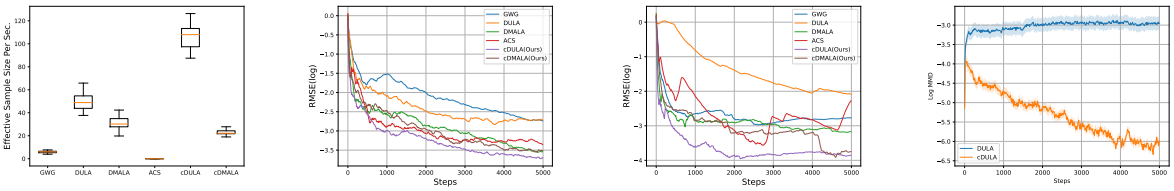
When combined with MH correction, the resulting algorithm corresponds to the *Discrete Metropolis-Adjusted Langevin Algorithm via Continuous Intermediates* (cDMALA); without MH correction, it yields the *Discrete Unadjusted Langevin Algorithm via Continuous Intermediates* (cDULA). The complete procedures are summarized in Algorithm 1. We also provide a theoretical analysis of the proposed cDLS framework, detailed in Appendix B.

4. Experiments

For both sampling and learning tasks, we compare our method against a range of representative discrete samplers, including Gibbs sampling, Gibbs-with-Gradients (GWG) (Grathwohl et al., 2021), the Discrete Unadjusted Langevin Algorithm (DULA), the Discrete Metropolis-Adjusted Langevin Algorithm (DMALA) (Zhang et al., 2022), and Automatic Cyclical Scheduling (Pynadath et al., 2024). These methods cover both classical and recent gradient-based approaches for sampling from discrete distributions. All methods are implemented in `Pytorch`, and we use the official release of code from previous papers when possible. To ensure a fair comparison, hyperparameters for all methods are tuned according to the guidelines suggested in the original works. Additional implementation details and hyperparameter settings are provided in Appendix K.

4.1. Sampling From Ising Models

We consider a 5×5 lattice Ising model, where the random variable $\theta \in \{-1, 1\}^d$ with $d = 25$. The energy function is defined as $U(\theta) = a\theta^\top W\theta + b\theta$, where W is the binary adjacency matrix of the lattice, $a = 0.1$ controls the connectivity strength, and $b = 0.2$ represents the external bias. More discussion is detailed in the Appendix K.4.



(a) ESS (b) $a = 0.1, b = 0.2$ (c) $a = 0.2, b = 0.1$ (d) Set $\alpha = 0.2$

Figure 2. Ising model sampling results. (a) cDULA yields the largest effective sample size (ESS) per second and (b)&(c) achieves the lowest or competitive RMSE error among all the methods compared. (d) Stability comparison under a larger stepsize $\alpha = 0.2$, where DULA becomes unstable while cDULA remains well-behaved.

Results Figure 2 reports the sampling performance on the 5×5 Ising model. As shown in Figure 2(a), cDULA achieves the largest effective sample size (ESS) (Lenth, 2001) per second among

all compared methods, indicating substantially reduced sample autocorrelation and more effective exploration of the discrete state space.

We further assess finite-sample accuracy by measuring the root-mean-square error (RMSE) between the empirical mean and the true mean. Figure 2(b) shows that cDULA converges faster and attains the lowest or comparable RMSE across methods. This suggests that incorporating continuous intermediates enables more informed discrete updates, allowing the chain to reach high-probability regions more efficiently under a fixed computational budget. Figures 2(c) demonstrate that this advantage persists across different interaction strengths and external fields, indicating robustness to changes in model geometry.

To further investigate stability with respect to stepsize selection, we compare DULA and cDULA under different values of α in Figure 2 and 6. When $\alpha = 0.2$, DULA exhibits clear instability and divergence, whereas cDULA remains stable and continues to explore the target distribution effectively. Even when both methods use a smaller stepsize $\alpha = 0.1$, cDULA consistently outperforms DULA, highlighting the benefit of gradient-guided continuous exploration beyond mere stepsize tuning. When parameters are jointly fine-tuned, the performance gap becomes even more pronounced.

Notably, cDULA achieves strong empirical performance even without MH correction. We attribute this to the proposal distribution induced by continuous intermediates, which better exploits local geometric structure than purely discrete proposals. Nevertheless, in more challenging learning scenarios that may exhibit divergence (e.g., Section K.6.3), incorporating MH correction provides a principled guarantee of convergence, as we have demonstrated in Section C.

Detailed Experimental Setup and Additional Experiments in the Appendix. Further experiment details and a series of additional experiments, including restricted Boltzmann machines, deep energy-based models, and binary Bayesian neural networks, are provided in Appendix K. Those widely experiments validate the superior performance of cDLS compared to existing methods.

5. Conclusions

In our paper, we propose a hybrid discrete-continuous sampler for discrete spaces that extends gradient-based updates from purely discrete settings into an augmented continuous domain. By allowing samples to explore along gradient directions in this expanded space, our method improves the efficiency of sampling from high-dimensional discrete distributions. For different usage scenarios, we develop several corresponding variants, which include the unadjusted and Metropolis-Hastings adjusted versions. We prove the asymptotic convergence of our sampler under log-quadratic and general distributions and non-asymptotic bounds for the MH adjusted version. Numerous experiments on many different tasks show that our method outperforms the baseline methods. As a simple and easy-to-use method, we hope that our method will improve the sampling efficiency of discrete distributions and provide a new perspective on the utilization of gradient information.

Acknowledgements

This research is supported in part by NSF IIS-2508145.

References

Leon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Krisztian Buza. Feedback prediction for blogs. In *Data analysis, machine learning and knowledge discovery*, pages 145–152. Springer, 2014.

Grace E Cho and Carl D Meyer. Comparison of perturbation bounds for the stationary distribution of a markov chain. *Linear Algebra and its Applications*, 335(1-3):137–150, 2001.

Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830*, 2016.

Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society Series B:Statistical Methodology*, 79(3):651–676, 2017.

Persi Diaconis, Susan Holmes, and Radford M Neal. Analysis of a nonreversible markov chain sampler. *Annals of Applied Probability*, pages 726–752, 2000.

Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models. *Advances in Neural Information Processing Systems*, 2019.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

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

Andrew B Duncan, Tony Lelievre, and Grigorios A Pavliotis. Variance reduction using nonreversible langevin samplers. *Journal of statistical physics*, 163(3):457–491, 2016.

Sean R Eddy. What is a hidden markov model? *Nature Biotechnology*, 22(10):1315–1316, 2004.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6): 721–741, 1984.

Will Grathwohl, Kevin Swersky, Milad Hashemi, David Duvenaud, and Chris J Maddison. Oops i took a gradient: Scalable sampling for discrete distributions. *International Conference on Machine Learning*, 2021.

Jun Han, Fan Ding, Xianglong Liu, Lorenzo Torresani, Jian Peng, and Qiang Liu. Stein variational inference for discrete distributions. In *International Conference on Artificial Intelligence and Statistics*, pages 4563–4572. PMLR, 2020.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

José Miguel Hernández-Lobato and Ryan Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International conference on machine learning*, pages 1861–1869. PMLR, 2015.

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

S. Mattu J. Angwin, J. Larson and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016.

Galin L. Jones. On the Markov chain central limit theorem. *arXiv Mathematics e-prints*, page math/0409112, 2004.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference for Learning Representations*, 2015.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

RV Lenth. Some practical guidelines for effective sample size determination. *AMERICAN STATISTICIAN*, 55(3):187–193, 2001.

Xingchao Liu, Xin Tong, and Qiang Liu. Sampling with trusthworthy constraints: A variational gradient framework. *Advances in Neural Information Processing Systems*, 34, 2021a.

Xingchao Liu, Mao Ye, Dengyong Zhou, and Qiang Liu. Post-training quantization with multiple points: Mixed precision without mixed precision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8697–8705, 2021b.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient mcmc. *Advances in neural information processing systems*, 28, 2015.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

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

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

Akihiko Nishimura, David B Dunson, and Jianfeng Lu. Discontinuous hamiltonian monte carlo for discrete parameters and discontinuous likelihoods. *Biometrika*, 107(2):365–380, 2020.

Patrick Pynadath, Riddhiman Bhattacharya, Arun Hariharan, and Ruqi Zhang. Gradient-based Discrete Sampling with Automatic Cyclical Scheduling. In *Neural Information Processing Systems*, 2024.

Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European conference on computer vision*, pages 525–542. Springer, 2016.

Gareth O Roberts and Osnat Stramer. Langevin diffusions and metropolis-hastings algorithms. *Methodology and computing in applied probability*, 4(4):337–357, 2002.

Paul J Schweitzer. Perturbation theory and finite markov chains. *Journal of Applied Probability*, 5 (2):401–413, 1968.

Yang Song and Diederik P Kingma. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*, 2021.

Haoran Sun, Hanjun Dai, and Dale Schuurmans. Optimal scaling for locally balanced proposals in discrete spaces. In *Neural Information Processing Systems*, volume 35, pages 23867–23880. Curran Associates, Inc., 2022a.

Haoran Sun, Hanjun Dai, Wei Xia, and Arun Ramamurthy. Path auxiliary proposal for mcmc in discrete space. In *International Conference on Learning Representations*, 2022b.

Haoran Sun, Bo Dai, Charles Sutton, Dale Schuurmans, and Hanjun Dai. Any-scale balanced samplers for discrete space. In *International Conference on Learning Representations*, 2023a.

Haoran Sun, Hanjun Dai, Bo Dai, Haomin Zhou, and Dale Schuurmans. Discrete langevin samplers via wasserstein gradient flow. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 6290–6313. PMLR, 25–27 Apr 2023b.

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

Tijmen Tieleman and Geoffrey Hinton. Using fast weights to improve persistent contrastive divergence. In *Proceedings of the 26th annual international conference on machine learning*, pages 1033–1040, 2009.

Ashish Vaswani et al. Attention is all you need. In *NeurIPS*, 2017.

Dootika Vats, James M Flegal, and Galin L Jones. Multivariate output analysis for Markov chain Monte Carlo. *Biometrika*, 106(2):321–337, April 2019. ISSN 0006-3444. doi: 10.1093/biomet/asz002. URL https://doi.org/10.1093/biomet/asz002.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Yue Xiang, Dongyao Zhu, Bowen Lei, Dongkuan Xu, and Ruqi Zhang. Efficient informed proposals for discrete distributions via newton's series approximation. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 7288–7310. PMLR, 25–27 Apr 2023.

Giacomo Zanella. Informed proposals for local MCMC in discrete spaces. *Journal of the American Statistical Association*, 115(530):852–865, 2020. Publisher: Taylor & Francis.

Ruqi Zhang, Chunyuan Li, Jianyi Zhang, Changyou Chen, and Andrew Gordon Wilson. Cyclical stochastic gradient mcmc for bayesian deep learning. In *International Conference on Learning Representations*, 2020.

Ruqi Zhang, Xingchao Liu, and Qiang Liu. A Langevin-like sampler for discrete distributions. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26375–26396. PMLR, 17–23 Jul 2022.

A. Related Work

Gradient-based Discrete Sampling Efficient sampling from high-dimensional discrete distributions remains challenging due to the lack of informative proposals and poor scalability of traditional methods. Gradient-based approaches, such as Gibbs with Gradients (GWG)(Grathwohl et al., 2021) approximate locally balanced proposals (Zanella, 2020) using first-order Taylor expansion to guide coordinate selection, significantly reducing computational cost. Then, Sun et al. (2022b) introduced auxiliary path variables to extend local neighborhoods while maintaining gradient-based advantages. However, these methods still perform sequential updates and are limited by local neighborhood constraints. Inspired by Langevin dynamics in continuous space (Roberts and Stramer, 2002), Zhang et al. (2022) introduced a parallel update scheme for discrete sampling using a Langevin-like proposal. Follow-up works further extend this line by handling non-differentiable objectives (Xiang et al., 2023) and introducing adaptive scaling (Sun et al., 2023a, 2022a; Pynadath et al., 2024). In contrast, our work aims to accelerate gradient-based discrete sampling by more fully exploiting the underlying continuous structure, rather than relying solely on gradients. This introduces a new way to utilize continuous information, bridging the gap between discrete and continuous sampling approaches.

Continuous Relaxation Incorporating gradients in the proposal has been a great success in continuous space, such as the Langevin algorithm, Hamiltonian Monte Carlo (HMC), and their variants (Duane et al., 1987; Neal et al., 2011). To take advantage of this success, continuous relaxation is developed, which samples in a continuous space and then transforms the collected samples to the original discrete space (Han et al., 2020; Nishimura et al., 2020). As shown in previous work, this type of method usually does not scale to high-dimensional discrete distributions (Grathwohl et al., 2021). Relaxation-based methods conduct most sampling steps in the continuous space, with discretization applied only at the end. In contrast, our approach only incorporates a continuous intermediate step within each discrete update. This design preserves the discrete nature of the target space throughout the sampling process. The continuous intermediates in cDLS are not used to replace discrete sampling but instead act as local exploratory guides that improve discrete proposal construction then accelerate convergence.

B. Theoretical Analysis

In this section, we provide a theoretical analysis of the proposed cDLS framework. Our analysis proceeds in two stages. We first study a simple and widely used log-quadratic setting, which admits an exact characterization and serves as a baseline for understanding the effect of continuous intermediates. We then extend the analysis to general target distributions whose energy functions deviate from quadratic structure, and quantify the resulting bias in terms of local linearization error. Finally, we analyze the Metropolis-adjusted variant cDMALA and establish non-asymptotic convergence guarantees.

B.1. A Simple and Widely Used Log-Quadratic Case

We begin with a simple yet widely used setting in which the target distribution $\pi(\theta)$ admits a log-quadratic form,

$$\pi(\theta) \propto \exp(\theta^\top W\theta + b^\top\theta), \quad \theta \in \Theta, \tag{11}$$

where $W \in \mathbb{R}^{d\times d}$ is symmetric (otherwise replaced by $(W + W^\top)/2$) and $b \in \mathbb{R}^d$. Such distributions arise naturally in structured discrete models, including Ising models and Gaussian graphical models (Mezard and Montanari, 2009; Wainwright et al., 2008). In this setting, the gradient of the energy function is exactly linear which leads to no intrinsic linearization error.

Theorem 1. *Assume the target distribution π is log-quadratic as in Equation 11. Let $\lambda_{\min}$ denote the smallest eigenvalue of W, and assume $\alpha_0 = o(\alpha)$ with $\alpha > 0$, $\alpha_0 \geq 0$. Then:*

i *The Markov chain defined by the proposal $q(\cdot|\theta^c)$ in Equation 8 is reversible.*

ii *The stationary distribution π_{α,α_0} converges weakly to π as $\alpha \to 0$. Moreover,*

$$\|\pi_{\alpha,\alpha_0} - \pi\|_1 \leq cZ \exp\left(-\frac{1}{2\alpha} - \frac{(\alpha+\alpha_0)\lambda_{\min}}{2\alpha}\right),$$

where Z is the normalizing constant of π and $c = \exp(\alpha_0 M/4)$ with $M = 2Da(2a + \|b\|_\infty)$, $D = \max_{\theta,\theta'\in\Theta} \|\theta - \theta'\|_\infty$, and $a = \max_i \sum_j |W_{ij}|$.

Theorem 1 shows that the bias of cDULA vanishes exponentially as $\alpha \to 0$, implying asymptotic correctness. Moreover, for any fixed $\alpha > 0$ and $\lambda_{\min} \geq 0$, the bound is strictly tighter than in the baseline case $\alpha_0 = 0$ (i.e., DULA), by an additional factor of $\exp\big(-\frac{\alpha_0\lambda_{\min}}{2\alpha}\big)$, up to a mild constant. Intuitively, the continuous intermediates steer the state toward high-probability regions before the discrete update, reducing the local linearization error that drives discretization bias under quadratic energies.

While the above result is derived under an exact log-quadratic assumption, a similar qualitative behavior persists for more general target distributions. In particular, when the gradient of the energy function is approximately linear over the discrete support, the stationary bias can be decomposed into a discretization-induced term analogous to the quadratic case and an additional approximation error due to nonlinearities (see Appendix B.2 for a formal statement).

In practice, this structure helps explain why cDULA remains well-behaved at moderate stepsizes where DULA may become unstable or divergent, and why it achieves lower bias over a finite number of iterations without excessive fine-tuning, as observed empirically in the Ising model experiments in Section 4.1.

B.2. Extension to General Target Distributions

We briefly discuss how the preceding analysis extends to general target distributions whose energy functions deviate from exact quadratic structure. In such cases, discretization bias arises not only from finite step sizes but also from local nonlinearities of the energy landscape.

To quantify this effect, we measure the deviation of the gradient $\nabla U(\theta)$ from a linear function. Specifically, we assume that there exist $W \in \mathbb{R}^{d\times d}$, $b \in \mathbb{R}^d$, and $\epsilon > 0$ such that

$$\|\nabla U(\theta) - (2W\theta + b)\|_1 \leq \epsilon, \quad \forall \theta \in \Theta. \tag{12}$$

This condition formalizes the notion that the energy function is approximately log-quadratic in a gradient sense, with ϵ controlling the local linearization error over the discrete domain. Let $\pi'(\theta) \propto \exp(\theta^\top W\theta + b^\top\theta)$ denote the corresponding log-quadratic distribution.

Theorem 2. *Under Assumption (12), the stationary distribution π_{α,α_0} of cDULA satisfies*

$$\|\pi_{\alpha,\alpha_0} - \pi\|_1 \leq 2c_1\big(\exp(c_2\epsilon) - 1\big) + Z' \exp\left(-\frac{1}{2\alpha} - \frac{(\alpha+\alpha_0)\lambda_{\min}}{2\alpha}\right),$$

where $c_1 = O(C_2)$ depends only on the distribution π' and the fixed stepsizes (α, α_0), and $c_2 = O(D + \max_\theta \|W\theta + b\|)$ depends only on geometric bounds of the finite discrete state space Θ

This result decomposes the total bias into two parts: The first term captures how far the true energy function U is from a quadratic form, with error scaling exponentially in ϵ. Thus, as $\epsilon \to 0$, the behavior of cDULA approaches that in Theorem 1; The second term is inherited from the previous theorem and quantifies the bias introduced by using non-zero stepsizes. Together, these results offer practical guidance: to reduce bias, one should choose a continuous extension whose gradient is close to linear over the discrete support Θ.

B.3. Non-asymptotic Convergence

Beyond asymptotic bias, we now analyze the non-asymptotic behavior of cDMALA. Our goal is to provide theoretical guarantees for the convergence rate and support statistical inference from finite samples. We establish these results by proving uniform ergodicity of the Markov chain P under standard smoothness and local concavity assumptions, which are common in the literature on optimization and sampling (Dalalyan, 2017; Bottou et al., 2018). We begin by stating the regularity conditions required for the analysis.

Assumption 3. *The energy function $U(\theta) \in \mathcal{C}^2(\mathbb{R}^d)$ has an L-Lipschitz gradient, i.e.,*

$$\|\nabla U(\theta) - \nabla U(\theta')\| \leq L\|\theta - \theta'\|, \qquad \forall \theta, \theta' \in \mathbb{R}^d.$$

Assumption 4. *For each $\theta \in \mathbb{R}^d$, there exists an open ball $B(\theta, r_\theta)$ such that $U(\cdot)$ is m_θ-strongly concave on $B(\theta, r_\theta)$ for some $m_\theta > 0$.*

These assumptions allow local control of both gradient variation and curvature, which are sufficient to establish a minorization condition for the Markov kernel.

Lemma 5. *Let Assumptions 3 and 4 hold, and suppose that $\alpha + \alpha_0 < \frac{2}{L}$. Then the Markov kernel P associated with* cDMALA *satisfies, for any $\theta, \theta' \in \Theta$,*

$$p(\theta \mid \theta') \;\geq\; \epsilon_{\alpha,\alpha_0} \frac{\exp(\tilde{\alpha} U(\theta))}{\sum_{\vartheta \in \Theta} \exp(\tilde{\alpha} U(\vartheta))},$$

where $\tilde{\alpha} = (\alpha + \alpha_0)/(2\alpha), a \in \arg\min_{\theta \in \Theta} \|\nabla U(\theta)\|$ and

$$\epsilon_{\alpha,\alpha_0} = \exp\left(-\left(\tfrac{2}{\alpha} + L - \tfrac{(\alpha+\alpha_0)m}{4\alpha}\right) D^2 - \tfrac{\alpha+\alpha_0}{2\alpha} \|\nabla U(a)\| D\right).$$

The minorization condition in Lemma 5 implies uniform ergodicity of the Markov chain and yields the following non-asymptotic convergence guarantees.

Theorem 6. *Under Assumptions 3 and 4, and for $\alpha + \alpha_0 < \frac{2}{L}$, the Markov chain P induced by* cDMALA *satisfies:*

i ***Uniform ergodicity.*** *For all $n \geq 1$,*

$$\|P^n - \pi\|_{\mathrm{TV}} \leq (1 - \epsilon_{\alpha,\alpha_0})^n.$$

ii ***Central Limit Theorem.*** *For any real-valued test function f,*

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}_\pi[f]\right) \xrightarrow{d} \mathcal{N}(0, \sigma_*^2),$$

for some $\sigma_*^2 > 0$.

Theorem 6 provides two complementary guarantees for cDMALA. First, uniform ergodicity ensures exponential convergence to the target distribution in total variation distance. Second, the Central Limit Theorem justifies valid statistical inference from finite samples, with asymptotic variances that can be consistently estimated using standard techniques such as batch means (Vats et al., 2019).

A key driver of this improvement lies in the modified parameter dependencies in the convergence bound. Compared to DMALA, the convergence coefficient $\epsilon_{\alpha,\alpha_0}$ of cDMALA features stronger contributions from both the gradient magnitude and the local curvature. Specifically, the gradient term is amplified by a factor of $(\alpha+\alpha_0)/2\alpha$, which enhances the directional bias of proposals toward high-probability regions when α_0 is chosen such that $0 < \alpha_0 = o(\alpha)$. Meanwhile, the effective curvature term becomes $(\alpha+\alpha_0)m/4\alpha$, allowing the proposal to better exploit strong local concavity.

Especially, under locally strongly concave potentials (large m) and when moderate stepsizes α are used, the effective lower bound $\epsilon_{\alpha,\alpha_0}$ of cDMALA is strictly larger than that of DMALA, which leads to faster convergence.

In summary, cDMALA is theoretically guaranteed to converge faster than DMALA whenever the target distribution exhibits sufficient local curvature and the auxiliary parameter α_0 is scaled appropriately with α. In these scenarios, the continuous intermediate yields more informed proposals and tighter non-asymptotic bounds.

Algorithm 1: Discrete Langevin Samplers via Continuous Intermediates

Input: Stepsizes α, α_0; Initial state θ_0.
for $k = 0, 1, 2, \ldots$ **do**
  `// Continuous intermediate`
  Compute $\theta_k^{c,\text{raw}} \leftarrow \theta_k + \frac{\alpha_0}{2}\nabla U(\theta_k)$;
  Set $\theta_k^c \leftarrow \text{Proj}_K(\theta_k^{c,\text{raw}})$;
  `// Discrete proposal`
  for $i = 1, 2, \ldots, d$ **do**
    Construct $q_i(\cdot \mid \theta_k^c)$ as in Equation 10;
    Sample $\theta'_{k,i} \sim q_i(\cdot \mid \theta_k^c)$;
  end
  Set $\theta'_k = (\theta'_{k,1}, \ldots, \theta'_{k,d})$;
  `// Optional MH correction`
  Compute $q(\theta'_k \mid \theta_k^c) = \prod_i q_i(\theta'_{k,i} \mid \theta_k^c)$;
  Compute $\theta_k^{c',\text{raw}} \leftarrow \theta'_k + \frac{\alpha_0}{2}\nabla U(\theta'_k)$;
  Set $\theta_k^{c'} \leftarrow \text{Proj}_K(\theta_k^{c',\text{raw}})$;
  Compute $q(\theta_k \mid \theta_k^{c'}) = \prod_i q_i(\theta_{k,i} \mid \theta_k^{c'})$;
  if *with MH step* **then**
    Set $\theta_{k+1} \leftarrow \theta'_k$ with probability in Equation 13;
  else
    Set $\theta_{k+1} \leftarrow \theta'_k$;
  end
end
Output: Samples $\{\theta_k\}$.

C. Metropolis–Hastings Correction

To ensure the resulting Markov chain has the desired stationary distribution $\pi(\theta)$, we optionally apply a Metropolis-Hastings (MH) correction (Metropolis et al., 1953; Hastings, 1970), which is usually combined with proposals to make the Markov chain reversible. Given the proposal $\theta' \sim q(\cdot|\theta^c)$, the acceptance probability is

$$A(\theta \to \theta') = \min\left(1, \exp\big(U(\theta') - U(\theta)\big)\frac{q(\theta \mid \theta^{c'})}{q(\theta' \mid \theta^c)}\right), \tag{13}$$

where θ^c and $\theta^{c'}$ denote the continuous intermediates associated with θ and θ', respectively.

While the MH correction enforces detailed balance, we find empirically that cDLS often performs effectively even without this adjustment. This observation is consistent with prior work showing that non-reversible Markov chains can exhibit improved mixing behavior and reduced asymptotic variance in Markov chain Monte Carlo methods (Diaconis et al., 2000; Duncan et al., 2016; Ma et al., 2015). Moreover, incorporating the MH correction offers a practical advantage by enabling adaptive stepsize selection based on the acceptance rate (Sun et al., 2022a), whereas samplers without MH correction typically rely on heuristic rules or grid-based tuning to identify suitable parameters.

C.1. Summary of the cDLS Framework

The cDLS framework introduces a two-stage update mechanism for discrete sampling: a gradient-guided continuous intermediate followed by a discrete update induced by this intermediate. By temporarily lifting the dynamics into continuous space, cDLS enables gradient information to be exploited more effectively while preserving the discrete nature of the Markov chain.

D. Algorithm with Binary Variables

cDLS could be simplified when the variable domain is binary $\Theta = \{0,1\}^d$. We obtain Algorithm 2 from 1 to help understand.

Algorithm 2: Discrete Langevin Samplers via Continuous Intermediates for Binary Variables

Input: The stepsize α, α^0
while *True* **do**
  Compute $\theta^c_{raw} = \theta + \frac{\alpha_0}{2}\nabla U(\theta)$;
  Project θ^c_{raw};
  Compute $P(\theta|\theta^c) = \frac{\exp\left(\frac{1}{2}\nabla U(\theta^c)\odot(1-\theta-\theta^c) - \frac{(1-\theta-\theta^c)^2}{2\alpha}\right)}{\exp(\frac{1}{2}\nabla U(\theta^c)\odot(1-\theta-\theta^c) - \frac{(1-\theta-\theta^c)^2}{2\alpha}) + \exp(\frac{1}{2}\nabla U(\theta^c)\odot(\theta-\theta^c) - \frac{(\theta-\theta^c)^2}{2\alpha})}$;
  Sample $\mu \sim \text{Unif}(0,1)^d$;
  $I \leftarrow \textbf{dim}(\mu \leq P(\theta|\theta^c))$;
  $\theta' \leftarrow \textbf{flipdim}(I)$;
  `// Optionally, do the MH step`
  Compute $q(\theta'|\theta) = \prod_i q_i(\theta_i^{'}|\theta) = \prod_{i\in I} P_i(\theta|\theta^c)\prod_{i\notin I}(1-P_i(\theta|\theta^c))$;
  Compute $\theta^{c'}_{raw} = \theta' + \frac{\alpha_0}{2}\nabla U(\theta')$;
  Project $\theta^{c'}_{raw}$;
  Construct $P(\theta'|\theta_c^{'}) = \frac{\exp\left(\frac{1}{2}\nabla U(\theta_c^{'})\odot(1-\theta'-\theta_c^{'}) - \frac{(1-\theta'-\theta_c^{'})^2}{2\alpha}\right)}{\exp(\frac{1}{2}\nabla U(\theta_c^{'})\odot(1-\theta'-\theta_c^{'}) - \frac{(1-\theta'-\theta_c^{'})^2}{2\alpha}) + \exp(\frac{1}{2}\nabla U(\theta_c^{'})\odot(\theta'-\theta_c^{'}) - \frac{(\theta'-\theta_c^{'})^2}{2\alpha})}$;
  Compute $q(\theta|\theta') = \prod_i q_i(\theta_i|\theta) = \prod_{i\in I} P_i(\theta'|\theta_c^{'})\prod_{i\notin I}(1-P_i(\theta'|\theta_c^{'}))$;
  Set $\theta \leftarrow \theta'$ with probability in Equation 13;
end
Output: samples$\{\theta_k\}$;

E. Algorithm with Categorical Variables

When using one-hot vectors to represent categorical variables, our algorithm becomes

$$\text{Categorical}\left(\text{Softmax}\left(\frac{1}{2}\nabla U(\theta^c)_i^\top(\theta_i^{'} - \theta_i^c) - \frac{\|\theta_i' - \theta_i^c\|_2^2}{2\alpha}\right)\right).$$

where $\theta_i^{'}$ is the one-hot vector, $\theta_c^{'}$ is the explored vector. Notably, if one-hot vectors are used, the gradient-explored vectors should extend on the coordinate with one. That is say, Equation 7 for $\theta = e_i$ should follow that

$$\theta^c = \text{argmin}_{y\in K}||y - \theta c|| = Proj_K(\theta^c)$$

Actually, if the variables are ordinal with clear ordering information, we can also use integer representation $\theta \in \{0, 1\cdots, K-1\}^d$ and follow Equation 6, Equation 7 and Equation 8 to sample from discrete distributions.

F. Necessity of the Projection

One direct method is without Equation 7. That is, the sample has explored its domain by gradient, then updates according to the extended continuous sample. However, it may meet some obstacles and be beyond the sample domain, which will result in a very low acceptance probability in practice. For example, in the RBM experiment, its acceptance probability is 0.001% while cDMALA's acceptance probability is 62% on average. We use a simple illustration in Figure 1.

G. Accelerating Other Gradient-Based Samplers via Continuous Intermediates

In the main text, we primarily focus on discrete Langevin samplers. In this section, we extend our framework to demonstrate how continuous intermediates can also accelerate other gradient-based discrete samplers.

Recall that in Grathwohl et al. (2021), Gibbs with Gradients (GWG) updates samples through several coordinate-wise steps. For simplicity, we consider the case of binary variables.

Algorithm 3: Gibbs With Gradients for Binary Variables

Input: energy function $U(\cdot)$, current sample θ
while *True* **do**
  Compute $\tilde{d}(\theta) = -(2\theta - 1) \odot \nabla U(\theta)$;
  Compute $q(i|\theta) = \text{Categorical}\left(\text{Softmax}\left(\frac{\tilde{d}(\theta)}{2}\right)\right)$;
  Sample $i \sim q(i|\theta)$;
  $\theta' = \texttt{flipdim}(\theta, i)$;
  Compute $q(i|\theta') = \text{Categorical}\left(\text{Softmax}\left(\frac{\tilde{d}(\theta')}{2}\right)\right)$;
  Accept with probability:;
  $\min\left(\exp(U(\theta') - U(\theta))\frac{q(i|\theta')}{q(i|\theta)}, 1\right)$;
end
Output: samples$\{\theta_k\}$;

When computing $\tilde{d}(\theta)$, introducing a continuous intermediate enables the sampler to construct a more informed proposal. Specifically, we replace this step with

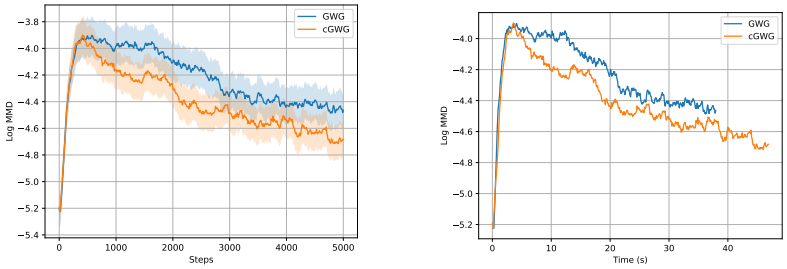
$$\theta^c = \text{Proj}_K(\theta + \frac{\alpha_0}{2}\nabla U(\theta)) = \arg\min_{y \in K} \|y - \theta - \frac{\alpha_0}{2}\nabla U(\theta)\|. \tag{14}$$

We refer to this variant as Gibbs with Gradients via Continuous intermediates (cGWG) and simply evaluate it on Restricted Boltzmann Machines (RBMs). Although cGWG updates only a single coordinate at each step, similar to the original GWG, it achieves consistently better performance, highlighting the effectiveness of incorporating continuous intermediates.

H. Proof of Theorem 1

Proof. Recall that,

$$\theta_c = \theta + \frac{\alpha_0}{2}\nabla U(\theta)$$
$$\theta' \sim \text{Categorical}\left(\text{Softmax}\left(\frac{1}{2}\nabla U(\theta_c)^\top(\theta' - \theta_c) - \frac{||\theta' - \theta_c||_2^2}{2\alpha}\right)\right).$$



(a) Log-MMD vs. Step (b) Log-MMD vs. Time

Figure 3. RBM Sampling for GWG and cGWG.

Expand the terms, we get

$$
\begin{aligned}
&\frac{1}{2}\nabla U(\theta_c)^\top(\theta'-\theta-\frac{\alpha_0}{2}\nabla U(\theta)) - \frac{\|\theta'-\theta-\frac{\alpha_0}{2}\nabla U(\theta)\|_2^2}{2\alpha} \\
\overset{*}{=}&\frac{1}{2}\nabla U(\theta)^\top(\theta'-\theta) - \frac{\|\theta'-\theta\|_2^2}{2\alpha} + (\frac{1}{2}\nabla U(\theta_c) - \frac{1}{2}\nabla U(\theta) + \frac{\alpha_0}{2\alpha}\nabla U(\theta))^\top(\theta'-\theta) \\
=&\frac{\alpha+\alpha_0}{2\alpha}\nabla U(\theta)^\top(\theta'-\theta) - \frac{\|\theta'-\theta\|_2^2}{2\alpha} + \frac{\alpha_0}{4}\nabla U(\theta)^\top\nabla^2 U(\theta)(\theta'-\theta) \\
=&\frac{\alpha+\alpha_0}{2\alpha}\nabla U(\theta)^\top(\theta'-\theta) - \frac{\|\theta'-\theta\|_2^2}{2\alpha} + \frac{\alpha_0}{4}\nabla U(\theta)^\top(\nabla U(\theta')-\nabla U(\theta)) \\
\overset{*}{=}&\frac{\alpha+\alpha_0}{2\alpha}\nabla U(\theta)^\top(\theta'-\theta) - \frac{\|\theta'-\theta\|_2^2}{2\alpha} + \frac{\alpha_0}{4}\nabla U(\theta)^\top(\nabla U(\theta'))
\end{aligned}
$$

Here, $*$ means we drop some terms that won't affect the softmax result, i.e, not relevant to θ'. Recall that the target distribution is $\pi(\theta) = \exp(\theta^\top W\theta + b^\top\theta)/Z$. We have that $\nabla U(\theta) = 2W^\top\theta + b, \nabla^2 U(\theta) = 2W$. Since $\nabla^2 U(\theta)$ is a constant, we can rewrite the proposal distribution as the following:

$$
\begin{aligned}
q_{\alpha,\alpha_0}(\theta'|\theta) &= \frac{\exp\left(\frac{\alpha+\alpha_0}{2\alpha}\nabla U(\theta)^\top(\theta'-\theta) - \frac{\|\theta'-\theta\|_2^2}{2\alpha} + \frac{\alpha_0}{4}\nabla U(\theta)^\top(\nabla U(\theta'))\right)}{\sum_x \exp\left(\frac{\alpha+\alpha_0}{2\alpha}\nabla U(\theta)^\top(x-\theta) - \frac{\|x-\theta\|_2^2}{2\alpha} + \frac{\alpha_0}{4}\nabla U(\theta)^\top(\nabla U(x))\right)} \\
&= \frac{\exp\left(\frac{\alpha+\alpha_0}{2\alpha}(U(\theta')-U(\theta)) - (\theta'-\theta)^\top(\frac{1}{2\alpha}I + \frac{\alpha+\alpha_0}{2\alpha}W)(\theta'-\theta) + \frac{\alpha_0}{4}\nabla U(\theta)^\top(\nabla U(\theta'))\right)}{\sum_x \exp\left(\frac{\alpha+\alpha_0}{2\alpha}(U(x)-U(\theta)) - (x-\theta)^\top(\frac{1}{2\alpha}I + \frac{\alpha+\alpha_0}{2\alpha}W)(x-\theta) + \frac{\alpha_0}{4}\nabla U(\theta)^\top(\nabla U(x))\right)}
\end{aligned}
$$

Let $Z_{\alpha,\alpha_0}(\theta) = \sum_x \exp\left(\frac{\alpha+\alpha_0}{2\alpha}(U(x)-U(\theta)) - (x-\theta)^\top(\frac{1}{2\alpha}I + \frac{\alpha+\alpha_0}{2\alpha}W)(x-\theta) + \frac{\alpha_0}{4}\nabla U(\theta)^\top(\nabla U(x))\right)$, and $\pi_{\alpha,\alpha_0} = \frac{Z_{\alpha,\alpha_0}(\theta)\pi(\theta)}{\sum Z_{\alpha,\alpha_0}(x)\pi(x)}$, now we will show that q_{α,α_0} is reversible w.r.t π_{α,α_0}. To simplify the notation, we use $\tilde{\alpha}$ to take place (α, α_0).

We have that

$$
\begin{aligned}
\pi_{\tilde{\alpha}} q_{\tilde{\alpha}}(\theta'|\theta) &= \frac{Z_{\tilde{\alpha}}(\theta)\pi(\theta)}{\sum Z_{\tilde{\alpha}}(x)\pi(x)} \cdot \frac{\exp\left(\frac{\alpha+\alpha_0}{2\alpha}(U(\theta')-U(\theta)) - (\theta'-\theta)^\top(\frac{1}{2\alpha}I + \frac{\alpha+\alpha_0}{2\alpha}W)(\theta'-\theta) + \frac{\alpha_0}{4}\nabla U(\theta)^\top(\nabla U(\theta'))\right)}{Z_{\tilde{\alpha}}} \\
&= \frac{\exp\left(\frac{\alpha+\alpha_0}{2\alpha}U(\theta') + \frac{\alpha-\alpha_0}{2\alpha}U(\theta)) - (\theta'-\theta)^\top(\frac{1}{2\alpha}I + \frac{\alpha+\alpha_0}{2\alpha}W)(\theta'-\theta) + \frac{\alpha_0}{4}\nabla U(\theta)^\top(\nabla U(\theta'))\right)}{Z \cdot \sum_x Z_{\tilde{\alpha}}(x)\pi(x)}
\end{aligned}
$$

Only and only if $\alpha_0 = o(\alpha)$, it is clear that this expression is symmetric in θ and θ'. Therefore $q_{\tilde{\alpha}}$ is reversible and the stationary distribution is $\pi_{\tilde{\alpha}}$.

Now we will prove that $\pi_{\tilde{\alpha}}$ converges weakly to π as $\alpha, \alpha_0 \to 0$ ($\alpha_0 = o(\alpha)$). Notice that for any θ,

$$\begin{aligned} Z_{\tilde{\alpha}}(\theta) &= \sum_x \exp\left(\frac{\alpha+\alpha_0}{2\alpha}(U(x)-U(\theta)) - (x-\theta)^\top(\frac{1}{2\alpha}I + \frac{\alpha+\alpha_0}{2\alpha}W)(x-\theta) + \frac{\alpha_0}{4}\nabla U(\theta)^\top(\nabla U(x))\right) \\ &\overset{\tilde{\alpha}\downarrow 0}{=} \sum_x \exp\left(\frac{1}{2}(U(x)-U(\theta))\right)\delta_\theta(x) \\ &= 1. \end{aligned}$$

where $\delta_\theta(x)$ is a Dirac delta. Due to $Z_{\tilde{\alpha}}(\theta)$ is a continuous function when θ is fixed on $\mathcal{A} = \{\alpha > 0, \alpha_0 \geq 0\}$, so the limitation is well-defined. It follows that $\pi_{\tilde{\alpha}}$ converges pointwisely to $\pi(\theta)$. By Scheffe's Lemma, we attain $\pi_{\tilde{\alpha}}$ converges weakly to π.

Convergence Rate w.r.t Stepsize Let us consider the convergence rate in terms of L_1-norm.

$$\|\pi_{\tilde{\alpha}} - \pi\|_1 = \sum_\theta \left|\frac{Z_{\tilde{\alpha}}(\theta)\pi(\theta)}{\sum_x Z_{\tilde{\alpha}}(x)\pi(x)} - \pi(\theta)\right|.$$

We write out each absolute value term

$$\begin{aligned} \left|\frac{Z_{\tilde{\alpha}}(\theta)\pi(\theta)}{\sum_x Z_{\tilde{\alpha}}(x)\pi(x)} - \pi(\theta)\right| &= \pi(\theta)\left|\frac{Z_{\tilde{\alpha}}(\theta)}{\sum_x Z_{\tilde{\alpha}}(x)\pi(x)} - 1\right| \\ &= \pi(\theta)\left|\frac{1+\sum_{x\neq\theta}\exp(\frac{\alpha+\alpha_0}{2\alpha}(U(x)-U(\theta)) - (x-\theta)^\top(\frac{1}{2\alpha}I + \frac{\alpha+\alpha_0}{2\alpha}W)(x-\theta) + \frac{\alpha_0}{4}\nabla U(\theta)^\top(\nabla U(x)-\nabla U(\theta)))}{1+\sum_y \frac{1}{Z}\exp(U(y))\sum_{x\neq y}\exp(\frac{\alpha+\alpha_0}{2\alpha}(U(x)-U(\theta)) - (x-\theta)^\top(\frac{1}{2\alpha}I + \frac{\alpha+\alpha_0}{2\alpha}W)(x-\theta) + \frac{\alpha_0}{4}\nabla U(\theta)^\top(\nabla U(x)-\nabla U(\theta)))} - 1\right| \end{aligned}$$

where $\tilde{\alpha} = \frac{\alpha+\alpha_0}{2\alpha}$. Since $\lambda_{min}(W)\|x\|^2 \leq x^\top W x, \forall x$, it follows that

$$(x-\theta)^\top(\frac{1}{2\alpha}I + \frac{\alpha+\alpha_0}{2\alpha}W)(x-\theta) \geq \|x-\theta\|^2(\frac{1+(\alpha+\alpha_0)\lambda_{min}}{2\alpha})$$

We also notice that $\min_{x\neq\theta}\|x-\theta\|^2 = 1$, and apply Hölder's inequality yields

$$\left|\nabla U(\theta)^\top\big(\nabla U(x) - \nabla U(\theta)\big)\right| = \left|(2W^\top\theta + b)^\top(2W^\top(x-\theta)\right| \leq 2\,\|2W\theta + b\|_\infty\,\|W(x-\theta)\|_1.$$

Let

$$D := \max_{\theta,\theta'\in\Theta}\|\theta-\theta'\|_\infty, \qquad a := \max_i \sum_j |W_{ij}|.$$

Since $\|x-\theta\|_\infty \leq D$, we have

$$\|W(x-\theta)\|_1 \leq \|W\|_{\infty\to 1}\,\|x-\theta\|_\infty \leq Da,$$

and

$$\|2W\theta + b\|_\infty \leq 2a + \|b\|_\infty.$$

Combining the above bounds, we obtain

$$\left|\nabla U(\theta)^\top\big(\nabla U(x) - \nabla U(\theta)\big)\right| \leq 2Da\big(2a + \|b\|_\infty\big) = M.$$

Thus when $\frac{Z_{\tilde{\alpha}}(\theta)}{\sum_x Z_{\tilde{\alpha}}(x)\pi(x)} - 1 > 0$, we get

$$\begin{aligned} \left|\frac{Z_{\tilde{\alpha}}(\theta)\pi(\theta)}{\sum_x Z_{\tilde{\alpha}}(x)\pi(x)} - \pi(\theta)\right| &= \pi(\theta)\left|\frac{Z_{\tilde{\alpha}}(\theta)}{\sum_x Z_{\tilde{\alpha}}(x)\pi(x)} - 1\right| \\ &= \pi(\theta)\left(\frac{1+\sum_{x\neq\theta}\exp(\frac{\alpha+\alpha_0}{2\alpha}(U(x)-U(\theta)) - (x-\theta)^\top(\frac{1}{2\alpha}I + \frac{\alpha+\alpha_0}{2\alpha}W)(x-\theta) + \frac{\alpha_0}{4}\nabla U(\theta)^\top(\nabla U(x)-\nabla U(\theta)))}{1+\sum_y \frac{1}{Z}\exp(U(y))\sum_{x\neq y}\exp(\frac{\alpha+\alpha_0}{2\alpha}(U(x)-U(y)) - (x-y)^\top(\frac{1}{2\alpha}I + \frac{\alpha+\alpha_0}{2\alpha}W)(x-y) + \frac{\alpha_0}{4}\nabla U(y)^\top(\nabla U(x)-\nabla U(y)))} - 1\right) \\ &\leq \pi(\theta)\left(1 + \sum_{x\neq\theta}\exp(\frac{\alpha+\alpha_0}{2\alpha}(U(x)-U(\theta)) - (\frac{1}{2\alpha}I + \frac{\alpha+\alpha_0}{2\alpha}W)\|x-\theta\|^2 + \frac{\alpha_0}{4}\nabla U(\theta)^\top(\nabla U(x)-\nabla U(\theta))) - 1\right) \\ &\leq \pi(\theta)\left(c\sum_x \exp(U(x))\right)\cdot\exp\left(-\frac{1+(\alpha+\alpha_0)\lambda_{min}}{2\alpha}\right) \\ &= c\cdot\pi(\theta)Z\cdot\exp(-\frac{1+(\alpha+\alpha_0)\lambda_{min}}{2\alpha}) \end{aligned}$$

where $c = \exp(\alpha_0 M/4)$ is a constant. Similarly, when $\frac{Z_{\tilde{\alpha}(\theta)}}{\sum_x Z_{\tilde{\alpha}}(x)\pi(x)} - 1 < 0$, we have,

$$\begin{aligned}
\left|\frac{Z_{\tilde{\alpha}}(\theta)\pi(\theta)}{\sum_x Z_{\tilde{\alpha}}(x)\pi(x)} - \pi(\theta)\right| &= \pi(\theta)\left(1 - \frac{1 + \sum_{x\neq\theta}\exp(\frac{\alpha+\alpha_0}{2\alpha}(U(x)-U(\theta)) - (x-\theta)^\top(\frac{1}{2\alpha}I + \frac{\alpha+\alpha_0}{2\alpha}W)(x-\theta) + \frac{\alpha_0}{4}\nabla U(\theta)^\top(\nabla U(x) - \nabla U(\theta)))}{1 + \sum_y \frac{1}{Z}\exp(U(y))\sum_{x\neq y}\exp(\frac{\alpha+\alpha_0}{2\alpha}(U(x)-U(y)) - (x-y)^\top(\frac{1}{2\alpha}I + \frac{\alpha+\alpha_0}{2\alpha}W)(x-y) + \frac{\alpha_0}{4}\nabla U(y)^\top(\nabla U(x) - \nabla U(y)))}\right) \\
&\leq \pi(\theta)\left(1 - \frac{1}{1 + \sum_y \frac{1}{Z}\exp(U(y))\sum_{x\neq y} c\exp(\frac{\alpha+\alpha_0}{2\alpha}(U(x)-U(\theta)) - \frac{1+(\alpha+\alpha_0)\lambda_{min}}{2\alpha})}\right) \\
&= \pi(\theta)\left(\frac{\sum_y \frac{1}{Z}\exp(U(y))\sum_{x\neq y} c\exp(\frac{\alpha+\alpha_0}{2\alpha}(U(x)-U(\theta)) - \frac{1+(\alpha+\alpha_0)\lambda_{min}}{2\alpha})}{1 + \sum_y \frac{1}{Z}\exp(U(y))\sum_{x\neq y} c\exp(\frac{\alpha+\alpha_0}{2\alpha}(U(x)-U(\theta)) - \frac{1+(\alpha+\alpha_0)\lambda_{min}}{2\alpha})}\right) \\
&\leq \pi(\theta)\left(\sum_y \frac{1}{Z}\exp(U(y))\sum_{x\neq y} c\exp(\frac{\alpha+\alpha_0}{2\alpha}(U(x)-U(\theta)))\right)\cdot\exp\left(-\frac{1+(\alpha+\alpha_0)\lambda_{min}}{2\alpha}\right) \\
&\leq \pi(\theta)\left(c\sum_x \exp(U(x))\right)\cdot\exp\left(-\frac{1+(\alpha+\alpha_0)\lambda_{min}}{2\alpha}\right) \\
&= c\cdot\pi(\theta)Z\cdot\exp(-\frac{1+(\alpha+\alpha_0)\lambda_{min}}{2\alpha})
\end{aligned}$$

Therefore, the difference between $\pi_{\tilde{\alpha}}$ and π can be bounded as follows.

$$\|\pi_{\tilde{\alpha}} - \pi\| \leq c\sum_\theta \pi(\theta)Z\cdot\exp\left(-\frac{1+(\alpha+\alpha_0)\lambda_{min}}{2\alpha}\right) = cZ\cdot\exp\left(-\frac{1+(\alpha+\alpha_0)\lambda_{min}}{2\alpha}\right).$$

I. Proof of Theorem 2

Proof. We use a log-quadratic distribution that is close to π as an intermediate term to bound the bias of cDULA. Recall that π is the target distribution, $\pi^{'}$ is the log-quadratic distribution that is close to π and $\pi_{\tilde{\alpha}}$ is the stationary distribution of cDULA. We let $\pi^{'}_{\tilde{\alpha}}$ be the stationary distributions of cDULA targetting $\pi^{'}$, then by triangle inequality,

$$\|\pi_{\tilde{\alpha}} - \pi\|_1 \leq \|\pi_{\tilde{\alpha}} - \pi'_{\tilde{\alpha}}\|_1 + \|\pi'_{\tilde{\alpha}} - \pi'\|_1 + \|\pi' - \pi\|_1.$$

Bound of $\|\pi' - \pi\|_1$. Let the energy function of π' be $V(\theta) = \theta^\top W\theta + b^\top\theta$. Since Θ is a discrete space, there exists a bounded subset $\Omega \in \mathrm{R}^d$ such that Θ is a subset of Ω. By Poincaré inequality, we get

$$|U(\theta) - V(\theta)| \leq C_1\cdot\|\nabla U(\theta) - (W\theta + b)\|_1 \leq C_1\epsilon, \qquad \forall\theta\in\Theta$$

where the constant C_1 depends on Ω.

Recall that $\pi(\theta) = \frac{\exp(U(\theta))}{Z}$. Let $\pi'(\theta) = \frac{\exp(V(\theta))}{Z'}$ where Z' is the normalizing constant to make π' a distribution. Then

$$\|\pi - \pi'\|_1 = \sum_{\theta\in\Theta}\left|\frac{\exp(U(\theta))}{Z} - \frac{\exp(V(\theta))}{Z'}\right|. \tag{15}$$

We notice that $\forall\theta$,

$$\frac{\exp(U(\theta))}{\exp(V(\theta))} = \exp(U(\theta) - V(\theta)) \leq \exp(|U(\theta) - V(\theta)|) \leq \exp(C\epsilon),$$

and similarly

$$\frac{\exp(V(\theta))}{\exp(U(\theta))} \leq \exp(C\epsilon).$$

Therefore we have $\forall\theta$,

$$\begin{aligned}
\left|\frac{\exp(U(\theta))}{Z} - \frac{\exp(V(\theta))}{Z'}\right| &= \left|\frac{\exp(U(\theta))\cdot Z' - \exp(V(\theta))\cdot Z}{Z\cdot Z'}\right| \\
&\leq (\exp(2C\epsilon) - 1)\min\left\{\frac{\exp(U(\theta))}{Z}, \frac{\exp(V(\theta))}{Z'}\right\} \\
&\leq (\exp(2C\epsilon) - 1)\left(\frac{\exp(U(\theta))}{Z} + \frac{\exp(V(\theta))}{Z'}\right).
\end{aligned} \tag{16}$$

Plugging the above in Equation 15, we obtain

$$\|\pi - \pi'\|_1 \leq 2\left(\exp(2C\epsilon) - 1\right).$$

Bound of $\|\pi'_{\tilde{\alpha}} - \pi'\|_1$. By Theorem 1, we know

$$\|\pi'_{\tilde{\alpha}} - \pi'\|_1 \leq cZ' \cdot \exp\left(-\frac{1 + (\alpha + \alpha_0)\lambda_{min}}{2\alpha}\right).$$

Bound of $\|\pi_{\tilde{\alpha}} - \pi'_{\tilde{\alpha}}\|_1$. Now we will bound $\|\pi_{\tilde{\alpha}} - \pi'_{\tilde{\alpha}}\|_1$ by perturbation bounds. Let $T_{\tilde{\alpha}}$, $T'_{\tilde{\alpha}}$ be the transition matrices of cDLS on π and π' respectively. We consider the difference between these two matrices.

$$\left\|T_{\tilde{\alpha}} - T'_{\tilde{\alpha}}\right\|_\infty = \max_\theta \sum_{\theta'} \left| \frac{\exp\left(\frac{\alpha+\alpha_0}{2\alpha}\nabla U(\theta)^\top(\theta' - \theta) - \frac{\|\theta'-\theta\|_2^2}{2\alpha} + \frac{\alpha_0}{4}\nabla U(\theta)^\top(\nabla U(\theta'))\right)}{Z_\alpha(\theta)} - \frac{\exp\left(\frac{\alpha+\alpha_0}{2\alpha}\nabla V(\theta)^\top(\theta' - \theta) - \frac{\|\theta'-\theta\|_2^2}{2\alpha} + \frac{\alpha_0}{4}\nabla V(\theta)^\top(\nabla V(\theta'))\right)}{Z'_\alpha(\theta)} \right|$$

where

$$Z_\alpha(\theta) = \sum_x \exp\left(\frac{\alpha + \alpha_0}{2\alpha}\nabla U(\theta)^\top(\theta' - \theta) - \frac{\|\theta' - \theta\|_2^2}{2\alpha} + \frac{\alpha_0}{4}\nabla U(\theta)^\top(\nabla U(\theta'))\right),$$

$$Z'_\alpha(\theta) = \sum_x \exp\left(\frac{\alpha + \alpha_0}{2\alpha}\nabla V(\theta)^\top(\theta' - \theta) - \frac{\|\theta' - \theta\|_2^2}{2\alpha} + \frac{\alpha_0}{4}\nabla V(\theta)^\top(\nabla V(\theta'))\right).$$

We denote $D = \max_{\theta,\theta' \in \Theta} \|\theta' - \theta\|_\infty$. By the assumption $\|\nabla U(\theta) - \nabla V(\theta)\|_1 \leq \epsilon$, we get

$$\begin{aligned}
\exp\left(\nabla U(\theta)^\top \nabla U(\theta') - \nabla V(\theta)^\top \nabla V(\theta')\right) &\leq \exp\left(\left|\nabla U(\theta)^\top \nabla U(\theta') - \nabla V(\theta)^\top \nabla V(\theta')\right|\right) \\
&= \exp\left(\left|\nabla U(\theta)^\top \nabla U(\theta') - \nabla U(\theta)^\top V(\theta') + \nabla U(\theta)^\top V(\theta') - \nabla V(\theta)^\top \nabla V(\theta')\right|\right) \\
&\leq \exp\left(\left|\nabla U(\theta)^\top \nabla U(\theta') - \nabla U(\theta)^\top V(\theta')\right| + \left|\nabla U(\theta)^\top V(\theta') - \nabla V(\theta)^\top \nabla V(\theta')\right|\right) \\
&\leq \exp(M\epsilon)
\end{aligned}$$

where $M = 2\max_\theta\{|\nabla V(\theta)|\}$. It gets from

$$|\nabla U(\theta)| = |\nabla U(\theta) - \nabla V(\theta) + \nabla V(\theta)| \leq C_1\epsilon + |\nabla V(\theta)|.$$

Similar to Equation (16), we have

$$\begin{aligned}
&\left| \frac{\exp\left(\frac{\alpha+\alpha_0}{2\alpha}\nabla U(\theta)^\top(\theta' - \theta) - \frac{\|\theta'-\theta\|_2^2}{2\alpha} + \frac{\alpha_0}{4}\nabla U(\theta)^\top(\nabla U(\theta'))\right)}{Z_\alpha(\theta)} - \frac{\exp\left(\frac{\alpha+\alpha_0}{2\alpha}\nabla V(\theta)^\top(\theta' - \theta) - \frac{\|\theta'-\theta\|_2^2}{2\alpha} + \frac{\alpha_0}{4}\nabla V(\theta)^\top(\nabla V(\theta'))\right)}{Z'_\alpha(\theta)} \right| \\
&\leq (\exp(N\epsilon) - 1) \\
&\quad \cdot \min\left\{ \frac{\exp\left(\frac{\alpha+\alpha_0}{2\alpha}\nabla U(\theta)^\top(\theta' - \theta) - \frac{\|\theta'-\theta\|_2^2}{2\alpha} + \frac{\alpha_0}{4}\nabla U(\theta)^\top(\nabla U(\theta'))\right)}{Z_\alpha(\theta)}, \frac{\exp\left(\frac{\alpha+\alpha_0}{2\alpha}\nabla V(\theta)^\top(\theta' - \theta) - \frac{\|\theta'-\theta\|_2^2}{2\alpha} + \frac{\alpha_0}{4}\nabla V(\theta)^\top(\nabla V(\theta'))\right)}{Z'_\alpha(\theta)} \right\} \\
&\leq (\exp(N\epsilon) - 1) \\
&\quad \cdot \left(\frac{\exp\left(\frac{\alpha+\alpha_0}{2\alpha}\nabla U(\theta)^\top(\theta' - \theta) - \frac{\|\theta'-\theta\|_2^2}{2\alpha} + \frac{\alpha_0}{4}\nabla U(\theta)^\top(\nabla U(\theta'))\right)}{Z_\alpha(\theta)} + \frac{\exp\left(\frac{\alpha+\alpha_0}{2\alpha}\nabla V(\theta)^\top(\theta' - \theta) - \frac{\|\theta'-\theta\|_2^2}{2\alpha} + \frac{\alpha_0}{4}\nabla V(\theta)^\top(\nabla V(\theta'))\right)}{Z'_\alpha(\theta)} \right).
\end{aligned}$$

where $N = \frac{\alpha+\alpha_0}{\alpha}D + \frac{M\alpha_0}{2}$. Now we substitute it to $\left\|T_\alpha - \tilde{T}_\alpha\right\|_\infty$,

$$\left\|T - \tilde{T}\right\|_\infty \leq 2\left(\exp(N\epsilon) - 1\right).$$

By the perturbation bound in Schweitzer (1968),

$$\|\pi_{\tilde{\alpha}} - \pi'_{\tilde{\alpha}}\|_1 \leq C_2 \cdot \|T_{\tilde{\alpha}} - T'_{\tilde{\alpha}}\|_\infty = 2C_2 \left(\exp\left(N\epsilon\right) - 1\right)$$

where C_2 is a constant depending on π' and $\tilde{\alpha}$. Please note that it is also possible to use other perturbation bounds (Cho and Meyer, 2001).

Combining these three bounds, we get

$$\begin{aligned}\|\pi_\alpha - \pi\|_1 &\leq \|\pi_\alpha - \pi'_\alpha\|_1 + \|\pi'_\alpha - \pi'\|_1 + \|\pi' - \pi\|_1 \\ &\leq 2C_2 \left(\exp\left(N\epsilon\right) - 1\right) + cZ' \cdot \exp\left(-\frac{1 + (\alpha + \alpha_0)\lambda_{min}}{2\alpha}\right) + 2\left(\exp(2C_1\epsilon) - 1\right).\end{aligned}$$

We define $c_1 := 2\max(2, 2C_2)$ and $c_2 := \max(2C_1, N)$, then we reach the the final result

$$\begin{aligned}\|\pi_\alpha - \pi\|_1 &\leq \|\pi_\alpha - \pi'_\alpha\|_1 + \|\pi'_\alpha - \pi'\|_1 + \|\pi' - \pi\|_1 \\ &\leq 2c_1 \cdot \left(\exp\left(c_2\epsilon\right) - 1\right) + Z' \cdot \exp\left(-\frac{1 + (\alpha + \alpha_0)\lambda_{min}}{2\alpha}\right).\end{aligned}$$

J. Proof of Lemma 5 and Theorem 6

Proof. Recalling cDLS, we consider the proposal kernel as

$$q_{\tilde{\alpha}}(\theta'|\theta) \propto \exp\left\{\tilde{\alpha}\nabla U(\theta)^\top(\theta' - \theta) - \frac{\|\theta - \theta'\|^2}{2\alpha} + \frac{1}{2}(\nabla U(\theta_c) - \nabla U(\theta))^\top(\theta' - \theta)\right\}$$

and consider the transition kernel as

$$p(\theta'|\theta) = \min\left\{\frac{\pi(\theta')q_{\tilde{\alpha}}(\theta|\theta')}{\pi(\theta)q_{\tilde{\alpha}}(\theta'|\theta)}, 1\right\} q_{\tilde{\alpha}}(\theta'|\theta) + (1 - L(\theta))\delta_\theta(\theta')$$

where $\delta_\theta(\theta')$ is the Kronecker delta function and $L(\theta)$ is the total acceptable probability from the point θ with

$$L(\theta) = \sum_{\theta' \in \Theta} \min\left\{\frac{\pi(\theta')q_{\tilde{\alpha}}(\theta|\theta')}{\pi(\theta)q_{\tilde{\alpha}}(\theta'|\theta)}, 1\right\} q_{\tilde{\alpha}}(\theta'|\theta).$$

We also define

$$Z_{\tilde{\alpha}}(\theta) = \sum_{x \in \Theta} \exp\left\{\tilde{\alpha}\nabla U(\theta)^\top(x - \theta) - \frac{\|x - \theta'\|^2}{2\alpha} + \frac{1}{2}(\nabla U(\theta_c) - \nabla U(\theta))^\top(x - \theta)\right\}$$

which is the normalizing constant for the proposal kernel.

Consider the term,

$$\tilde{\alpha}\nabla U(\theta)^\top(\theta' - \theta) - \frac{1}{2\alpha}\|\theta - \theta'\|^2 = \tilde{\alpha}(-U(\theta) + U(\theta')) - \frac{1}{2}(\theta - \theta')^\top\left(\tilde{\alpha}\int_0^1 \nabla^2 U[(1-s)\theta + s\theta']\mathrm{d}s + \frac{1}{\alpha}I\right)(\theta - \theta').$$

From Assumption 3, we have

$$\tilde{\alpha}\int_0^1 \nabla^2 U[(1-s)\theta + s\theta']\mathrm{d}s + \frac{1}{\alpha}I \geq (\frac{1}{\alpha} - \tilde{\alpha}L)I.$$

Since $\alpha + \alpha_0 < \frac{2}{L}$, the matrix $(\frac{1}{\alpha} - \tilde{\alpha}L)I$ is positive define. We denote that

$$\begin{aligned} p(\theta'|\theta) &= \min\left\{\frac{\pi(\theta')q_{\tilde{\alpha}}(\theta|\theta')}{\pi(\theta)q_{\tilde{\alpha}}(\theta'|\theta)}, 1\right\} q_{\tilde{\alpha}}(\theta'|\theta) + (1 - L(\theta))\delta_\theta(\theta') \\ &\geq \min\left\{\frac{\pi(\theta')q_{\tilde{\alpha}}(\theta|\theta')}{\pi(\theta)q_{\tilde{\alpha}}(\theta'|\theta)}, 1\right\} q_{\tilde{\alpha}}(\theta'|\theta) \\ &\geq \min\left\{\frac{Z_{\tilde{\alpha}}(\theta)\exp(\frac{\alpha_0}{4}L\left\|\nabla U(\theta')\right\|)}{Z_{\tilde{\alpha}}(\theta')\exp(\frac{\alpha_0}{4}LD\left\|\nabla U(\theta)\right\|)}, 1\right\} q_{\tilde{\alpha}}(\theta'|\theta) \end{aligned}$$

where $Z_{\tilde{\alpha}}$ could also be rewritten as

$$\begin{aligned} Z_{\tilde{\alpha}}(\theta) &= \sum_{x\in\Theta}\exp\left\{\tilde{\alpha}\nabla U(\theta)^\top(x-\theta) - \frac{\|x-\theta\|^2}{2\alpha} + \frac{1}{2}(\nabla U(\theta_c) - \nabla U(\theta))^\top(x-\theta)\right\} \\ &= \sum_{x\in\Theta}\exp\left\{\tilde{\alpha}(-U(\theta)+U(x)) - \frac{1}{2}(\theta-x)^\top\left(\tilde{\alpha}\int_0^1\nabla^2 U[(1-s)\theta+sx]\mathrm{d}s + \frac{1}{\alpha}I\right)(\theta-x) + \frac{1}{2}(\nabla U(\theta_c)-\nabla U(\theta))^\top(x-\theta)\right\} \end{aligned}$$

This can be seen as

$$\pi(\theta)q_{\tilde{\alpha}}(\theta'|\theta) = \frac{1}{ZZ_{\tilde{\alpha}}(\theta)}\exp\left\{\tilde{\alpha}(U(\theta)+U(\theta')) - \frac{1}{2}(\theta-\theta')^\top\left(\tilde{\alpha}\int_0^1\nabla^2 U[(1-s)\theta+s\theta']\mathrm{d}s + \frac{1}{\alpha}I\right)(\theta-\theta') + \frac{1}{2}(\nabla U(\theta_c)-\nabla U(\theta))^\top(\theta'-\theta)\right\}.$$

Since Assumption 4 holds true in this setting, we have an $m > 0$ such that for any $\theta \in conv(\Theta)$

$$-\nabla^2 U(\theta) \geq mI.$$

From this, one notes that

$$\exp\left(-\tilde{\alpha}U(\theta) - \frac{1}{2}(\frac{1}{\alpha} - \tilde{\alpha}m)D^2 + \frac{\alpha_0}{4}L\left\|\nabla U(\theta)\right\|\right)\sum_{x\in\Theta}\exp(\tilde{\alpha}U(x)) \leq Z_{\tilde{\alpha}}(\theta) \leq \\ \exp(-\tilde{\alpha}U(\theta) + \frac{\alpha_0}{4}LD\left\|\nabla U(\theta)\right\|)\sum_{x\in\Theta}\exp(\tilde{\alpha}U(x))$$

where the right-hand side follows from the fact that $\alpha + \alpha_0 < \frac{2}{L}$. Therefore,

$$\frac{Z_{\tilde{\alpha}}(\theta)}{Z_{\tilde{\alpha}}(\theta')} \geq \frac{\exp\{\tilde{\alpha}(-U(\theta)+U(\theta'))\}\cdot\exp\{\frac{\alpha_0}{4}L\left\|\nabla U(\theta)\right\|\}}{\exp\{\frac{1}{2}(\frac{1}{\alpha}-\tilde{\alpha}m)D^2\}\cdot\exp\{\frac{\alpha_0}{4}LD\left\|\nabla U(\theta')\right\|\}}$$

Also note that

$$\begin{aligned} q_{\tilde{\alpha}}(\theta'|\theta) =& \frac{\exp\left\{\tilde{\alpha}(-U(\theta)+U(\theta')) - \frac{1}{2}(\theta-\theta')^\top\left(\tilde{\alpha}\int_0^1\nabla^2 U[(1-s)\theta+s\theta']\mathrm{d}s + \frac{1}{\alpha}I\right)(\theta-\theta') + \frac{1}{2}(\nabla U(\theta_c)-\nabla U(\theta))^\top(\theta'-\theta)\right\}}{\sum_{x\in\Theta}\exp\left\{\tilde{\alpha}(-U(\theta)+U(x)) - \frac{1}{2}(\theta-x)^\top\left(\tilde{\alpha}\int_0^1\nabla^2 U[(1-s)\theta+sx]\mathrm{d}s + \frac{1}{\alpha}I\right)(\theta-x) + \frac{1}{2}(\nabla U(\theta_c)-\nabla U(\theta))^\top(x-\theta)\right\}} \\ \geq& \frac{\exp\left\{\tilde{\alpha}\nabla U(\theta)^\top(\theta'-\theta) - \frac{\|\theta-\theta'\|^2}{2\alpha} + \frac{1}{2}(\nabla U(\theta_c)-\nabla U(\theta))^\top(\theta'-\theta)\right\}}{\exp(-\tilde{\alpha}U(\theta) + \frac{1}{2}LD^2)\sum_{x\in\Theta}\exp(\tilde{\alpha}U(x))} \end{aligned}$$

and

$$\begin{aligned} -\tilde{\alpha}\nabla U(\theta)^\top(\theta'-\theta) + \frac{\|\theta-\theta'\|^2}{2\alpha} - \frac{1}{2}(\nabla U(\theta_c)-\nabla U(\theta))^\top(\theta'-\theta) =& \tilde{\alpha}(-\nabla U(\theta)+\nabla U(a))^\top(\theta'-\theta) - \nabla U(a)^\top(\theta'-\theta) \\ &+ \frac{\|\theta-\theta'\|^2}{2\alpha} - \frac{1}{2}(\nabla U(\theta_c)-\nabla U(\theta))^\top(\theta'-\theta) \\ &\leq \tilde{\alpha}\left\|\nabla U(a)-\nabla U(\theta)\right\|D + \tilde{\alpha}\left\|\nabla U(a)\right\|D + \frac{1}{2\alpha}D^2 \\ &\leq (\tilde{\alpha}L + \frac{1}{2\alpha})D^2 + \tilde{\alpha}\left\|\nabla U(a)\right\|D + \frac{1}{2}LD^2 \end{aligned}$$

Combining the inequities, we get

$$p(\theta'|\theta) \geq \epsilon_{\tilde{\alpha}} \frac{\exp\{\tilde{\alpha}U(\theta')\}}{\sum_{\theta'} \exp\{\tilde{\alpha}U(\theta')\}}$$

where

$$\epsilon_{\tilde{\alpha}} = \exp\left\{-(\frac{1}{\alpha} + \tilde{\alpha}L + L - \frac{\tilde{\alpha}m}{2})D^2 - \tilde{\alpha}\left\|\nabla U(a)\right\| D\right\}$$

To simplify, it also could be bound with $\alpha + \alpha_0 < \frac{2}{L}$,

$$\begin{aligned}\epsilon_{\tilde{\alpha}} &= \exp\left\{-(\frac{1}{\alpha} + \tilde{\alpha}L + L - \frac{\tilde{\alpha}m}{2})D^2 - \tilde{\alpha}\left\|\nabla U(a)\right\| D\right\} \\ &> \exp\left\{-(\frac{1}{\alpha} + \frac{2}{2\alpha \cdot L}L + L - \frac{\tilde{\alpha}m}{2})D^2 - \tilde{\alpha}\left\|\nabla U(a)\right\| D\right\} \\ &= \exp\left\{-(\frac{2}{\alpha} + L - \frac{\tilde{\alpha}m}{2})D^2 - \tilde{\alpha}\left\|\nabla U(a)\right\| D\right\}.\end{aligned}$$

As for Theorem 6, the proof directly follows from Lemma 5 and Cor.5 in Jones (2004).

K. Additional Experiments Results and Setting Details

K.1. Hyperparameter Settings

The key hyperparameters used for the compared methods are summarized as follows:

Table 1. Hyperparameter settings for all experiments.

Sampler	**GWG**	**DULA**	**DMALA**	**ACS**	**cDULA**	**cDMALA**
Table 1. General setup						
Hyperparameter	-	α	α	$[\beta_{\min}, \beta_{\max}]$	α, α_0	α, α_0
Table 2. Ising sample						
Hyperparameter	-	0.2	0.4	[0.5, 0.95]	0.4, 0.2	0.4, 0.2
Table 3. RBM sample and learning						
Hyperparameter	-	0.1	0.2	[0.5, 0.95]	0.2, 0.04	0.2, 0.04
Table 4. Ising learning						
Hyperparameter	-	0.1	0.2	[0.5, 0.95]	0.2, 0.01	0.2, 0.01
Table 5. EBM learning						
Hyperparameter	-	0.1	0.15	[0.5, 0.95]	0.15, 0.01	0.15, 0.01

We briefly introduce the hyperparameters used in different methods. GWG requires no additional hyperparameters, as it follows the Gibbs principle by updating one coordinate at a time based on gradient information. The parameter α in DLP and cDLS denotes the stepsize, analogous to that in continuous Langevin algorithms: a larger α leads to higher discretization error and potential divergence, whereas a smaller α results in more stable updates but requires more iterations for convergence. The parameter β in ACS serves as a balancing parameter, where a cyclical schedule is employed to generate a sequence of $\{\beta_k\}$. Correspondingly, a heuristic procedure is used to determine the associated stepsizes $\{\alpha_k\}$. Further implementation details can be found in Pynadath et al. (2024). The α_0 in cDLS controls the magnitude of exploration and is typically set not to exceed α.

K.2. Sampling from Synthetic Distribution

We use the experiment setting of Pynadath et al. (2024). We need to define the space between the modes, the total number of modes, and the variance of each mode σ. For convenience, we take the same mode number in Pynadath et al. (2024), the space between modes as 15, and the variance for each mode σ^2 as .3. Based on this, we can calculate the maximum value for each coordinate as follows:

$$MaxVal = (\sqrt{NumModes} + 1) * SpaceBetweenModes.$$

The mode's center could be calculated as follows:

$$\mu_{i,j} = \frac{MaxVal}{\sqrt{NumModes} + 2}(i + 1)$$

$$\mu_{i,j} = \frac{MaxVal}{\sqrt{NumModes} + 2}(j + 1)$$

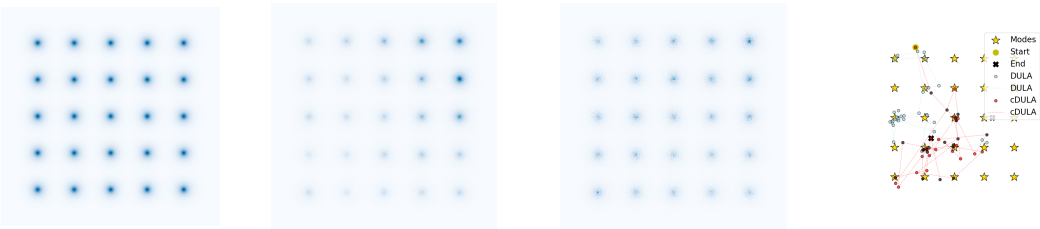
K.3. Sampling From a Synthetic Distribution

To illustrate how continuous intermediates can facilitate exploration in discrete multimodal landscapes, we construct a two-dimensional synthetic energy function defined as

$$U(\theta) = \log\left(\sum_{i=1}^{l} \exp\left(\frac{\|\theta - \mu_i\|^2}{2\sigma}\right)\right), \tag{17}$$

where the discrete domain is $\Theta = \{0, 1, \ldots, N\}^2$, and $\{\mu_1, \ldots, \mu_l\}$ specifies multiple well-separated modes.

This construction yields a multimodal target distribution with smooth geometry in the ambient continuous space but a constrained discrete support, making it a challenging test case for purely discrete gradient-based samplers. We compare DULA and cDULA in terms of both empirical total variation (TV) distance and qualitative sampling behavior, as shown in Figure 4. Additional experimental details are provided in Appendix K.2.



(a) Ground Truth (b) DULA,TV = 5.23 (c) cDULA,TV = 3.12 (d) Trajectory

Figure 4. Synthetic Distribution Sampling. We run the DULA and cDULA samplers to compare how continuous exploration helps sample from the target distribution.

Figure 4(b,c) shows that DULA tends to concentrate on a subset of modes and exhibits slow mixing across the landscape. In contrast, cDULA achieves a lower TV distance and more balanced coverage of the target distribution. The sampling trajectories in Figure 4(d) further illustrate this difference: by leveraging continuous intermediates (black circles), cDULA is able to move efficiently toward high-probability regions and reach dominant modes rapidly, whereas DULA often requires substantially more iterations to escape local neighborhoods.

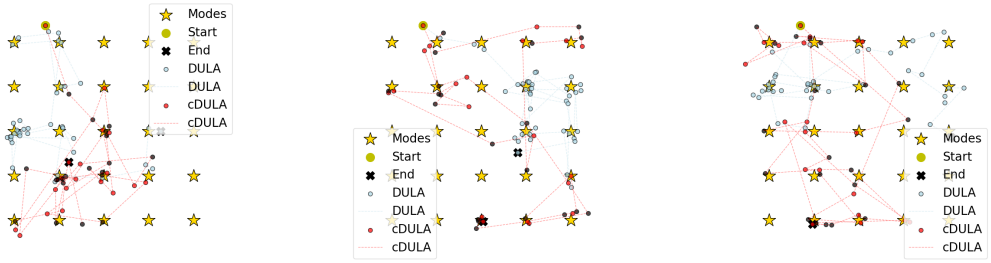
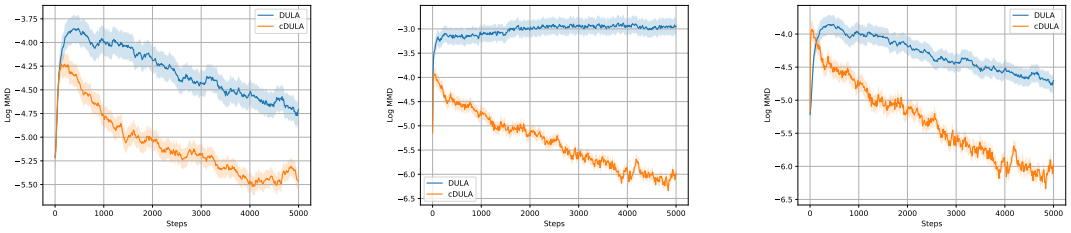


Figure 5. Sampling Trajectory with different random seeds.



(a) Set $\alpha = 0.1$ (b) Set $\alpha = 0.2$ (c) Set $\alpha = 0.1$ and 0.2.

Figure 6. Additional stability comparisons under different stepsizes.

K.4. Sampling from Ising Models

We adopt the experiment setting of Zhang et al. (2022). The energy of the Ising model is defined as,

$$\log(p(x)) = ax^\top Jx + b^\top x$$

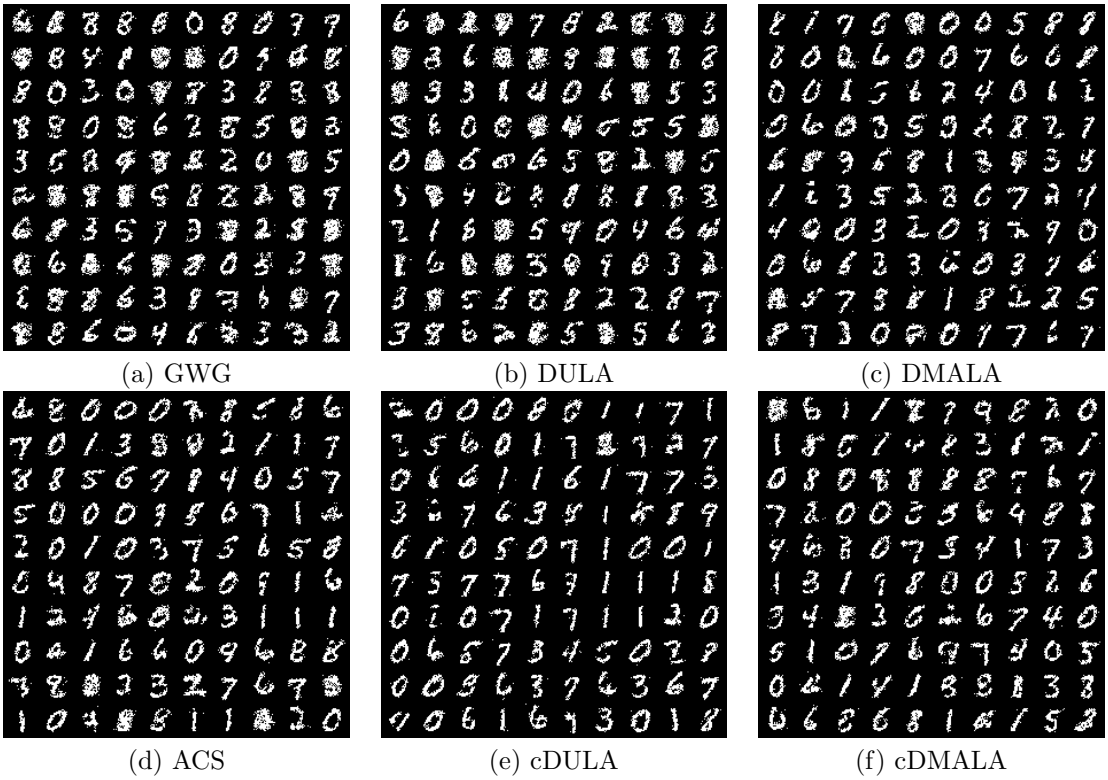
where a controls the connectivity strength and J is an adjacency matrix whose elements are either 0 or 1. If $J = 0$, then the model becomes a factorized Bernoulli distribution. In our experiments, J is the adjacency matrix of a 2D lattice graph. cDULA and cDMALA use the same stepsize tuple (0.4,0.2). We run all methods for the same number of iterations. The results are shown in Figure 2. We can clearly observe that our methods outperform other baselines on all connectivity strengths. These three coupling strengths respectively correspond to real scenarios within the Ising model dominated by a strong external field, strong internal interactions, and the absence of an external field.

K.5. Sampling from RBMs

RBM Overview We will give a brief overview of the Blocks-Gibbs sampler used to represent the ground truth of the RBM distribution. For a more in-depth explanation, see Grathwohl et al. (2021). Given the hidden units h and the sample x, we define the RBM distribution as follows:

$$\log(p(x, h)) = h^\top Wx + b^\top x + c^\top - \log(Z) \tag{18}$$

where Z is the normalizing constant for the distribution. The sample x is represented by the visible layer with units corresponding to the sample space dimension, and h represents the model capacity.



(a) GWG (b) DULA (c) DMALA

(d) ACS (e) cDULA (f) cDMALA

Figure 7. Generated samples from an RBM trained on MNIST. Our methods generate images that are closer to ground-truth samples.

It can be shown that the marginal distributions are as follows:

$$
\begin{aligned}
p(x|h) &= \text{Bernoulli}(Wx + c) \\
p(h|x) &= \text{Bernoulli}(W^\top h + b) \\
\log(p(x)) &= \sum_i \text{Softplus}(Wx + a)_i + b^\top x - \log(z).
\end{aligned}
$$

The Block-Gibbs sampler updates x and h alternatively, allowing for many of the coordinates to get changes at the same time, due to utilizing the specific structure of the RBM model.

Experiment Setup Similar to the experimental setup of Zhang et al. (2022); Pynadath et al. (2024), we use RBM models with 500 hidden units and 784 visible units. We adopt the same training protocol, except that we fine-tune the learning rate of Omniglot. From simpler MNIST-like data sets, we only train for a single pass through the dataset, while for more complicated datasets, we train for 3000 iterations to better represent the character and generate more realistic samplers. We include the generated images in Figure 11 to demonstrate that these models have learned the dataset reasonably well.

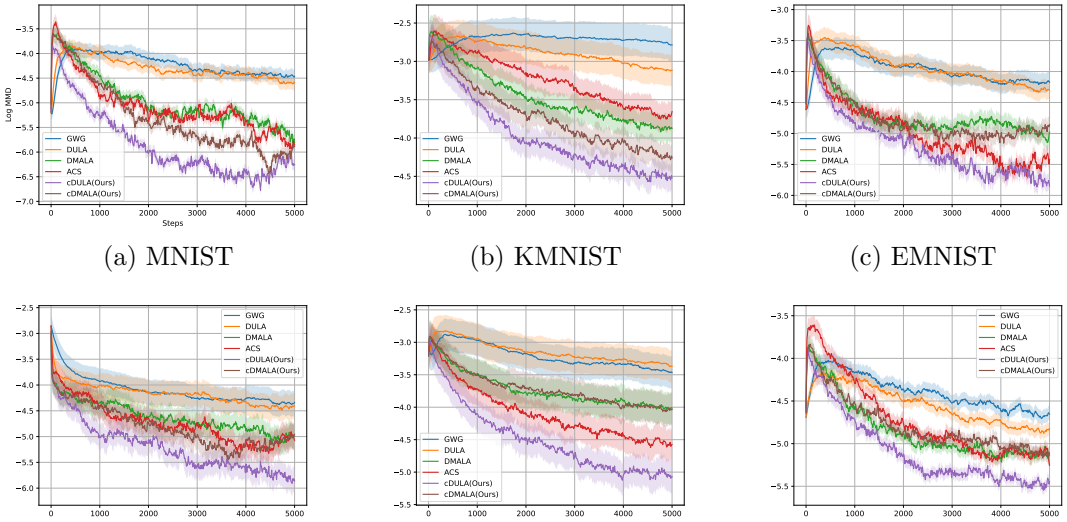
Sampler Configuration For GWG, we use the same setting as Grathwohl et al. (2021). For DLP, we set the stepsize to 0.2 and 0.1 alternatively with MH correction and without MH correction. For cDLS, we use $\alpha = 0.2, \alpha_0 = 0.04$.

Sampling Speed While the run time can vary depending on the specific implementation of a given sampling algorithm, we illustrate the efficiency of cDULA in Figure 9. cDULA is able to capture

the efficiency of DMALA without Metropolis-Hastings correction, which significantly improves the convergence speed and generates more diverse samples.

Effective sample size We find that the continuous exploration allows us to get closer to the high probability region before discrete sampling, allowing for larger steps of cDLS. In addition to accelerating the speed of convergence of the distribution, another significant benefit is the improvement in effective sample size (ESS). We run 100 chains in parallel with 5,000 iterations. Table 2 shows the ESS of the sampling of the RBM model, and it can be seen that our approach significantly improves the effective sample size.

Compared to Ising models, RBMs induce more complex, non-quadratic energy landscapes through latent-variable interactions, posing a more challenging sampling task. Following prior work (Grathwohl et al., 2021; Zhang et al., 2022), all samplers are initialized randomly. The RBM parameters $\{W, a, b\}$ are trained using contrastive divergence (Hinton, 2002). Sampling quality is evaluated using the Maximum Mean Discrepancy (MMD) between samples generated by each method and reference samples obtained via Block-Gibbs sampling, which exploits the conditional independence structure of RBMs.



(a) MNIST (b) KMNIST (c) EMNIST

(d) Fashion (e) Caltech (f) Omniglot

Figure 8. RBM sampling results measured by MMD to Block-Gibbs reference samples. cDLS, particularly cDULA, achieves competitive or superior convergence across all datasets.

Results Figure 8 reports the MMD convergence curves across six benchmark datasets. cDLS, and in particular cDULA, consistently attains lower MMD values or faster convergence compared to baseline discrete samplers. The improvements are especially pronounced on datasets with more complex or multimodal structure, such as MNIST-like datasets, Caltech, and Omniglot.

These results indicate that incorporating continuous intermediates enables more effective exploration of the non-quadratic energy landscapes induced by RBMs, leading to improved finite-sample performance. Additional qualitative results, including generated samples and runtime comparisons, are provided in Appendix K.5.

Table 2. Effective sample size for RBM sampling. Our method outperforms gradient-based baselines across all datasets.

Dataset	GWG	DULA	DMALA	ACS	cDULA	cDMALA
MNIST	151.22±553.28	397.86±1051.08	676.07±1269.37	878.55±768.21	1948.68±1868.10	989.63±1518.10
eMNIST	143.49±513.91	312.11±939.11	262.66±755.76	478.32±372.79	1305.44±1802.20	587.28±1088.37
kMNIST	143.08±554.69	381.70±1023.72	518.35±1059.64	799.86±873.49	1601.81±1829.77	721.82±1218.88
Fashion	43.76±159.14	184.00±630.67	711.49±1232.28	977.23±1029.32	1357.86±1555.12	781.11±1387.46
Caltech	123.20±446.86	222.62±703.02	501.88±796.30	789.21±891.79	860.05±862.92	705.36±932.25
Omniglot	83.33±288.10	101.73±524.26	428.52±982.70	573.49±807.91	709.58±1395.64	504.40±1104.62

Table 3. Improvement on average acceptance probability.

Dataset	DMALA	cDMALA	Dataset	DMALA	cDMALA
MNIST	0.5307	0.6595	Fashion	0.5752	0.7025
eMNIST	0.5415	0.6706	Caltech	0.5339	0.7095
kMNIST	0.5307	0.6595	Omniglot	0.7430	0.8191

K.6. Learning Energy-based Models

Energy-based models (EBMs) have demonstrated remarkable success across various machine learning domains (LeCun et al., 2006). An EBM defines a probability distribution as

$$f_\theta(x) = \frac{1}{Z_\theta} \exp(-E_\theta(x)) \tag{19}$$

where E_θ is a function parameterized by θ and Z_θ is the normalizing constant.[2] Training EBMs typically involves maximizing the log-likelihood,

$$\mathcal{L}(\theta) \triangleq \mathbb{E}_{x\sim p_{data}}[\log f_\theta(x)]\,. \tag{20}$$

Since computing Z_θ directly is usually infeasible, the optimization instead relies on estimating the gradient of the log-likelihood (Song and Kingma, 2021):

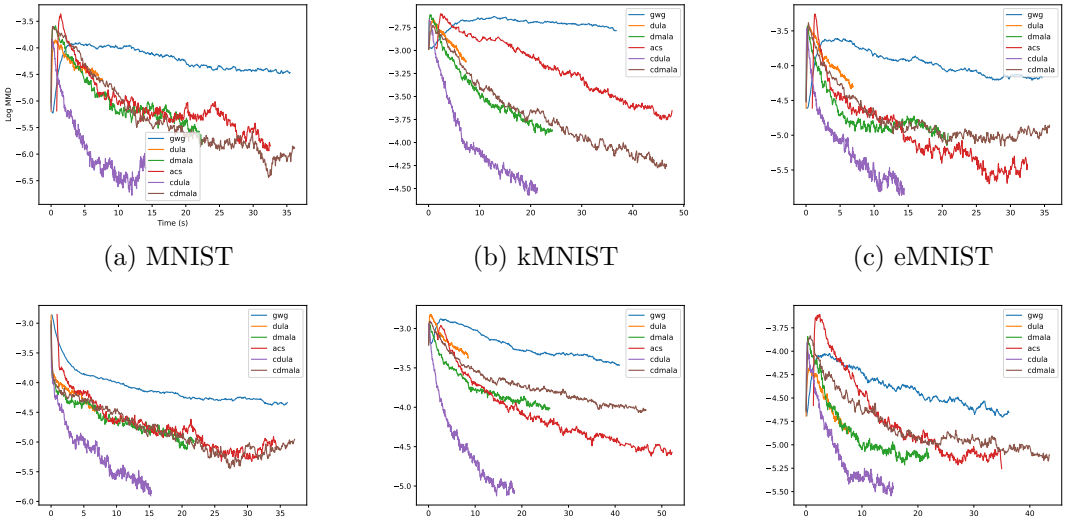
$$\nabla\mathcal{L}(\theta) = \mathbb{E}_{x\sim p_\theta}[\nabla_\theta E_\theta(x)] - \mathbb{E}_{x\sim p_{data}}[\nabla_\theta E_\theta(x)]\,. \tag{21}$$

While the second expectation can be easily computed from data samples, the first requires drawing samples from the model distribution p_θ, which is often the computational bottleneck. Consequently, the efficiency and quality of the sampler directly influence the training dynamics and the overall performance of EBMs. For all the experiments in this section, we use the stepsize α to be 0.1 for cDULA and 0.15 for cDMALA, α_0 to be 0.01 for the two.

K.6.1. Learning Ising Models

Following Grathwohl et al. (2021); Zhang et al. (2022), we consider a 25×25 Ising model and generate data by running the Gibbs sampler. In this setting, the energy function E_θ corresponds

2. In the general formulation, the energy function appears as $\exp(-\beta E_\theta(x))$, where β denotes the inverse temperature. In this work, following common practice in discrete sampling literature (Grathwohl et al., 2021; Zhang et al., 2022), we set $\beta = 1$ for simplicity without loss of generality.



(a) MNIST (b) kMNIST (c) eMNIST

(d) Fashion (e) Caltech (f) Omniglot

Figure 9. RBM Sampling with 5,000 sampling steps with results measured against time in seconds. cDULA is competitive both in terms of accuracy and efficiency.

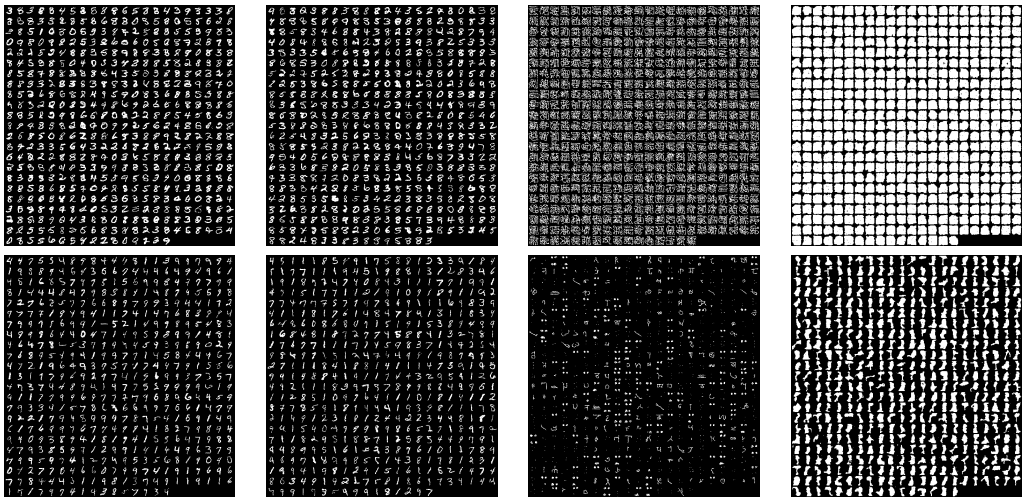


Figure 10. Left to right: Static MNIST, Dynamic MNIST, Omniglot, Caltech Silhouettes. Top: cDULA. Bottom: cDMALA.

to an Ising model parameterized by a learnable adjacency matrix $\hat{W}$. To evaluate the quality of different samplers, we compute the root mean squared error (RMSE) between the estimated $\hat{W}$ and the ground-truth W.

Experiment Design We construct a training dataset of 2,000 instances by running 1,000,000 steps of Gibbs sampling for each instance. The model is trained by Persistent Constastive Divergence Tieleman (2008) with a buffer size of 256 samples. We also use the Adam optimizer with a learning rate of 0.001. The batchsize is 256. We train all models with an l_1 penalty with a penalty coefficient

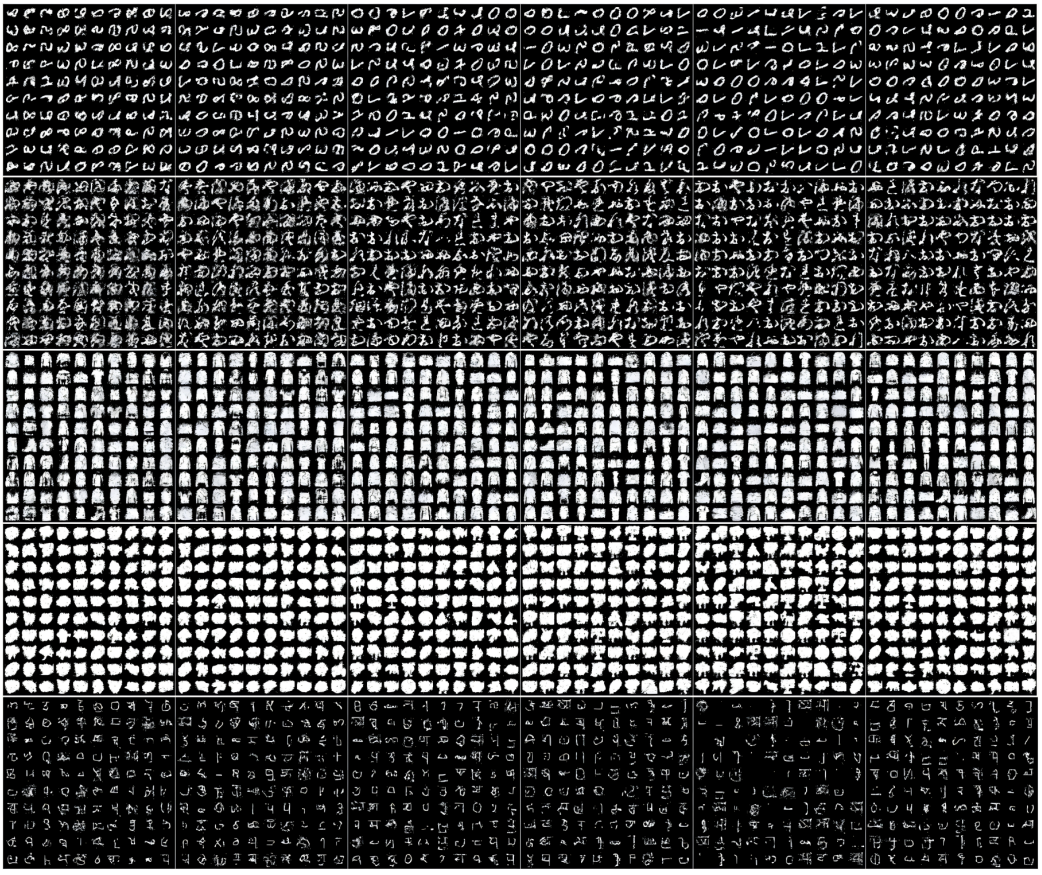
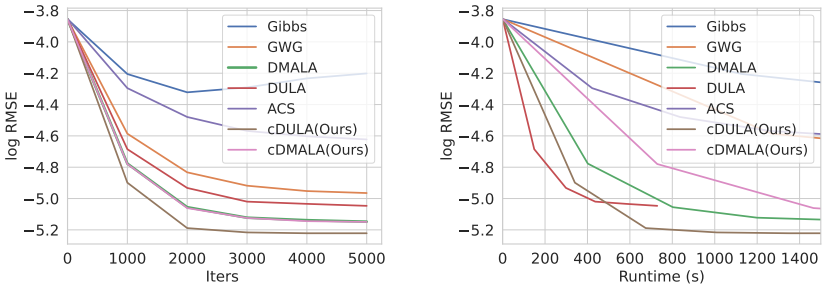


Figure 11. Images generated from RBMs trained by Constactive-Divergence with Block-Gibbs. Top to Bottom: eMNIST, kMNIST, Fashion, Caltech, Omniglot; Left to right: GWG, DULA, DMALA, ACS, cDULA and cDMALA.



(a) (b)

Figure 12. Ising model learning results. cDULA outperforms all baselines, finding the true W in a short time and getting a better convergence speed.

of 0.01 to encourage sparsity. The experiment setting is basically the same as Grathwohl et al. (2021); Zhang et al. (2022).

Results Our results are summarized in Figure 12. cDULA outperforms all the compared methods, having the smallest RMSE among the baselines given the same number of iterations. Compared to the running time, cDULA also achieves the same log-RMSE in less time, except DULA as cDULA need time to explore the continuous space.

K.6.2. Learning RBM

We demonstrate the benefits of cDLS on RBMs. We evaluate the learned model using Annealed Importance Sampling (AIS) (Neal, 2001). We compare the sampling methods of interest to Block-Gibbs (BG), which utilizes the structure of RBMs well.

Table 4. Log likelihood scores for RBM learning on test data as estimated by AIS. Our method outperforms gradient-based baselines across most datasets. ACS results are mainly taken from Pynadath et al. (2024).

Dataset	BG	GWG	DULA	DMALA	ACS	cDULA	cDMALA
MNIST	*-208.76*	-372.59	-271.17	-234.43	-249.55	**-211.24**	-346.90
eMNIST	*-337.90*	-445.68	-312.12	-349.85	-304.96	**-289.91**	-297.70
kMNIST	*-347.72*	-519.72	-381.27	-452.43	-407.39	**-380.12**	-436.04
Fashion	*-307.77*	-671.03	-471.35	-476.21	-452.37	**-406.30**	-460.93
Omniglot	*-177.83*	-338.34	**-180.66**	-267.83	-220.71	-190.47	-281.47
Caltech	*-524.34*	-593.47	-538.50	-731.18	-396.04	**-390.88**	-518.73

Experiment Design Follow Pynadath et al. (2024), We use the same RBM structure as the sampling task, with 500 hidden units and 784 visible units. However, we apply the samplers of interest to the PCD algorithm introduced by Tieleman (2008). The model parameters are tuned via the Adam optimizer Kingma and Ba (2015) with a learning rate of .001.

In order to evaluate the learned RBMs, we run AIS with Block-Gibbs as the sampler to calculate the log likelihood values for the models Neal (2001). We run AIS for 100,000 steps, which is adequate given the efficiency of Block Gibbs for this specific model. The results are shown in Table 4.

Results From Table 4, we note that our methods outperform the baselines on most of the datasets. Since the structure of Omniglot is quite different from the other datasets (illustrated in Figure 11), it leads to insufficiency of exploration when using the same α_0. This reveals that there should be a more adaptive approach to parameter tuning for different datasets, and we leave this for future work.

Table 5. Experiment results with binary Bayesian neural networks on different datasets.

Dataset	Training Log-likelihood (↑)						
	Gibbs	GWG	DULA	DMALA	ACS	cDULA	cDMALA
COMPAS	**-0.3102**$\pm$0.0029	-0.3104$\pm$0.0022	-0.3295$\pm$0.0049	-0.3122$\pm$0.0011	-0.3163$\pm$0.0002	-0.3431$\pm$0.0001	-0.3132$\pm$0.0033
News	-0.2132$\pm$0.0013	-0.2121$\pm$0.0023	-0.2116$\pm$0.0013	-0.2109$\pm$0.0001	-0.2099$\pm$0.0071	-0.2107$\pm$0.0001	**-0.2098**$\pm$0.0033
Adult	-0.3631$\pm$0.0007	-0.3249$\pm$0.0010	-0.3051$\pm$0.0001	-0.2950$\pm$0.0042	-0.3040$\pm$0.0014	**-0.2900**$\pm$0.0002	-0.3008$\pm$0.0030
Blog	-0.3746$\pm$0.0021	-0.3247$\pm$0.0004	-0.2705$\pm$0.0023	-0.2603$\pm$0.0031	-0.2654$\pm$0.0043	**-0.2601**$\pm$0.0006	-0.2607$\pm$0.0009
Dataset	**Test RMSE (↓)**						
	Gibbs	GWG	DULA	DMALA	ACS	cDULA	cDMALA
COMPAS	0.4795$\pm$0.0034	0.4774$\pm$0.0037	0.4848$\pm$0.0013	0.4775$\pm$0.0029	0.4750$\pm$0.0039	**0.4674**$\pm$0.0029	0.4789$\pm$0.0036
News	0.0964$\pm$0.0022	0.0975$\pm$0.0046	0.0994$\pm$0.0011	0.0946$\pm$0.0030	0.0992$\pm$0.0099	0.0932$\pm$0.0001	**0.0905**$\pm$0.0009
Adult	0.4284$\pm$0.0022	0.4044$\pm$0.0004	0.3900$\pm$0.0016	0.3853$\pm$0.0003	0.3919$\pm$0.0072	**0.3793**$\pm$0.0002	0.3869$\pm$0.0035
Blog	0.4040$\pm$0.0009	0.3564$\pm$0.0057	0.3186$\pm$0.0037	0.3130$\pm$0.0019	0.3168$\pm$0.0030	**0.3077**$\pm$0.0040	0.3099$\pm$0.0004

Table 6. Log likelihood scores for EBM learning on test data as estimated by AIS. cDLS is able to achieve better results than the baselines. Previous results are mainly taken from Pynadath et al. (2024).

Dataset	Gibbs	GWG	DULA	DMALA	ACS	cDULA	cDMALA
MNIST	-117.17	-80.01	-81.20	-79.93	-79.76	-80.42	**-79.55**
Dynamic MNIST	-121.19	-80.51	-80.06	-80.13	-79.70	**-79.50**	-79.62
Omniglot	-142.06	-94.72	-127.68	-100.08	-91.32	-111.72	**-91.01**
Caltech Silhouettes	-163.50	-96.20	-114.82	-99.35	-88.34	-102.56	**-87.82**

K.6.3. Learning EBM

We further evaluate our methods on deep convolutional energy-based models (EBMs) to demonstrate their efficiency in large-scale settings. In these experiments, the energy function E_θ is parameterized by a ResNet (He et al., 2016), trained using Persistent Contrastive Divergence (PCD) (Tieleman, 2008; Tieleman and Hinton, 2009) with a replay buffer (Du and Mordatch, 2019), following the setups of Grathwohl et al. (2021) and Zhang et al. (2022). After training, we employ Annealed Importance Sampling (AIS) to estimate the model likelihood. Guided by the findings of Pynadath et al. (2024), we use 10 sampling steps per training iteration for all datasets, except for Caltech, where we adopt 30 steps to ensure sufficient mixing.

Experiment Design We use the same EBM model architecture as Grathwohl et al. (2021); Zhang et al. (2022) and follow the same experimental design, where we implement ResNet with 8 residual blocks of 64 feature maps. We evaluate the models every 5,000 iterations by running AIS for 10, 000 steps. The reported results are from the model that performs the best on the validation set. The final reported numbers are generated by running 300, 000 iterations of AIS. All the models are trained with Adam with a learning rate of 0.0001 for 50, 000 iterations. We show the generated images with cDULA and cDMALA in Figure 10.

Results Table 6 demonstrates that our method achieves superior performance over GWG and DLP. This advantage comes from our sampler's enhanced capacity to explore a broader range of modes per iteration, which accelerates the discovery of high probability regions. Such efficient exploration directly translates into higher-quality gradient estimates during training, driving more stable and effective model updates. It is reasonable to hypothesize that cDLS with multi-step or adaptive exploration steps will be more advantageous in such complex experiments (Loshchilov and Hutter, 2017; Zhang et al., 2020; Pynadath et al., 2024).

Additional experiments on learning energy-based models, including Ising models, Restricted Boltzmann Machines, and Bayesian neural networks, are presented in Appendix K.6. To better understand the source of the empirical gains observed in the previous sections, we next conduct a comparison of algorithmic variants.

K.7. Binary Bayesian Neural Networks

Since MCMC-based methods are naturally suited for Bayesian inference, we further apply our approach to Bayesian neural networks (BNNs), where sampling efficiency plays a crucial role in model performance. Bayesian neural networks are known to provide reliable uncertainty estimation and strong predictive performance in deep learning (Hernández-Lobato and Adams, 2015; Zhang et al., 2020; Liu et al., 2021a). Meanwhile, binary neural networks (Courbariaux et al., 2016; Rastegari et al., 2016; Liu et al., 2021b), in which weights are restricted to $\{-1, 1\}$, offer significant computational and memory advantages. To combine the strengths of both paradigms, we study binary Bayesian neural networks trained via discrete sampling. We perform regression experiments on four

UCI datasets (Dua and Graff, 2017), where the energy function is defined as

$$U(\theta) = -\sum_{i=1}^{N} \|f_\theta(x_i) - y_i\|_2^2,$$

with $D = \{(x_i, y_i)\}_{i=1}^N$ denoting the training dataset, and f_θ representing a two-layer neural network with `Tanh` activation and 500 hidden units.

Details of the Datasets (1) COMPAS: COMPAS J. Angwin and Kirchner (2016) is a dataset containing the criminal records of 6,172 individuals arrested in Florida. The task is to predict whether the individual will commit a crime again in 2 years. We use 13 attributes for prediction. (2) News: Online News Popularity Data Set [3] contains 39,797 instances of the statistics on the articles published by a website. The task is to predict the number of clicks in several hours after the articles are published. (3) Adult: Adult Income Dataset [4] is a dataset containing the information of US individuals from the 1994 census. The prediction task is to predict whether an individual makes more than 50K dollars per year. The dataset contains 44,920 data points. (4) Blog: Blog Feedback Buza (2014) is a dataset containing 54,270 data points from blog posts. The raw HTML documents of the blog posts were crawled and processed. The prediction task associated with the data is the prediction of the number of comments in the upcoming 24 hours. The feature of the dataset has 276 dimensions.

Details on Training We run 10 chains in parallel and collect the samples at the end of training. All datasets are randomly partitioned into 80% for training and 20% for testing. The features and the predictive targets are normalized to (0,1). We set $\alpha = 0.1, \alpha_0 = 0.01$ for all datasets. We use a uniform prior over the weights. We train the Bayesian neural network for 2,000 steps. We use the full-batch training for the results in Section 4 so that Gibbs, GWG and DLP are also applicable. Beyond this, we could also get a stochastic variant for cDLS.

Results From Table 5, we observe that our methods significantly outperform other gradient-based discrete samplers on all datasets except COMPAS, as there are fewer data samples. This phenomenon could also be found in Zhang et al. (2022). These results demonstrate that our methods converge fast for high-dimensional distributions, due to the ability to explore the parameter space using gradient, and suggest that our methods are compelling for training low-precision Bayesian neural networks where weights are discrete. Additionally, our method shows superior generalization for data, as RMSE on test datasets is significantly decreased while the training likelihood is almost equal.

K.8. Comparison of Variants

The primary advantage of cDULA over DULA comes from the continuous intermediates, which allows for maintaining the comparable error with a larger stepsize (under certain conditions, the discretization error will increase the stepsize), shown in Figure 6. cDULA achieves a significant performance gain, remaining stable even with a doubled stepsize (i.e., $0.1 \to 0.2$), whereas DULA diverges quickly under the same stepsize.

When comparing cDULA and cDMALA, we observe that cDULA performs better in most of our experiments (Figure 8, Figure 12, Table 4). This observation is consistent with prior work showing that non-reversible Markov chains can exhibit improved mixing behavior (Diaconis et al., 2000; Duncan et al., 2016; Ma et al., 2015). For training complex EBMs, however, this adventure action of cDULA can lead to divergence on challenging datasets (e.g., Caltech in Table 6), while cDMALA remains stable and converges more reliably. We present this trade-off between efficiency and robustness without extensive hyperparameter tuning to emphasize the inherent characteristics of cDULA and cDMALA.

3. https://archive.ics.uci.edu/ml/datasets/online+news+popularity
4. https://archive.ics.uci.edu/ml/datasets/adult

The difference between cDMALA and DMALA entirely arises from the introduction of the continuous intermediates. Although MH correction in cDMALA occasionally leads to conservative updates, it still converges faster and yields larger effective sample sizes across several experiments (Figure 8, Table 2).

L. Discussion and Future works

There are still directions for further improvement of this domain. Currently, we adopt a fixed parameter α_0 across all datasets for simplicity in a single experiment. However, energy landscapes often vary in smoothness and curvature, suggesting that an adaptive strategy for tuning α_0 could better balance exploration and stability across different regimes. Although our methods already outperform existing baselines, designing more informed proposal mechanisms tailored to discrete domains remains a valuable direction, as such proposals could leverage structural priors to guide sampling more efficiently.