Slithering Through Gaps: Capturing Discrete Isolated Modes via Logistic Bridging

Pinaki Mohanty
Department of Computer Science
College of Science & College of Engineering
Purdue University, West Lafayette, IN, USA

Ruqi Zhang
Department of Computer Science
College of Science & College of Engineering
Purdue University, West Lafayette, IN, USA

Abstract

High-dimensional and complex discrete distributions often exhibit multimodal behavior due to inherent discontinuities, posing significant challenges for sampling. Gradient-based discrete samplers, while effective, frequently become trapped in local modes when confronted with rugged or disconnected energy landscapes. This limits their ability to achieve adequate mixing and convergence in high-dimensional multimodal discrete spaces. To address these challenges, we propose *Hyperbolic Secant-squared Gibbs-Sampling (HiSS)*, a novel family of sampling algorithms that integrates a *Metropolis-within-Gibbs* framework to enhance mixing efficiency. HiSS leverages a logistic convolution kernel to couple the discrete sampling variable with the continuous auxiliary variable in a joint distribution. This design allows the auxiliary variable to encapsulate the true target distribution while facilitating easy transitions between distant and disconnected modes. We provide theoretical guarantees of convergence and demonstrate empirically that HiSS outperforms many popular alternatives on a wide variety of tasks, including Ising models, binary neural networks, and combinatorial optimization.

1 INTRODUCTION

Gradient-based sampling methods, such as Langevin and Hamiltonian Monte Carlo (HMC) (Roberts and Rosenthal, 2002; Neal et al., 2011), have achieved remarkable success in continuous spaces by using gradient information to guide transitions between states. However, these methods face significant limitations when sampling from multimodal distributions with low-density barriers separating distinct modes. In such scenarios, gradients often fail to provide the global context for effective navigation across disconnected regions (Betancourt, 2017; Livingstone et al., 2019; Pompe et al., 2020). Techniques like cyclical step size (Zhang et al., 2020a), PT(parallel tempering) (Swendsen and Wang, 1986), and flat-histogram approaches (Berg and Neuhaus, 1992) have been developed to address this issue, improving mixing efficiency in continuous spaces by facilitating mode transitions.

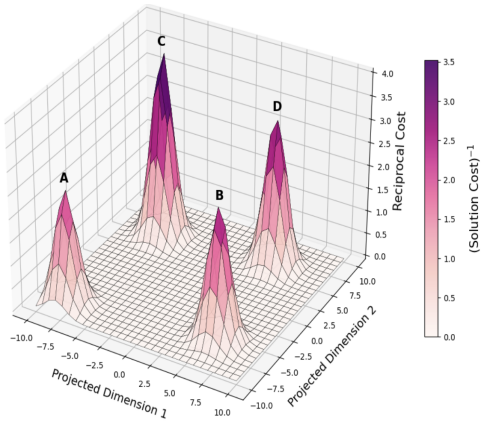

Figure 1: A visualization of the multimodal cost landscape in the Traveling Salesman Problem, showcasing solutions separated by low energy barriers.

In discrete spaces, these challenges are even more pronounced due to the inherent discontinuities in the landscape, which exacerbate ruggedness and multimodality. These challenges are particularly pervasive in applications such as text generation in natural language processing (NLP) (Gu et al., 2018; Devlin et al., 2019; Lewis et al., 2020), protein coupling prediction (Lapedes et al., 1999), low-precision neural networks (Courbariaux et al., 2016), and combinatorial optimization (Applegate et al.,

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

2006). For instance, in Figure 1 mode A is thoroughly isolated making it difficult for gradient-based samplers to escape and discover lower-cost alternatives such as modes B, C, or D.

While gradient-based discrete samplers, such as the Discrete Langevin Proposal (DLP) (Zhang et al., 2022) and Gibbs-With-Gradient (GWG) (Grathwohl et al., 2021b), have improved sampling efficiency, they face limitations similar to their continuous counterparts. These methods rely heavily on local gradient information, which renders them *myopic* in their exploration, often failing to identify paths to promising yet disconnected regions in the landscape. Recent advancement using cyclical schedules (Pynadath et al., 2024) has attempted to address these limitations. However, the challenge of effectively sampling from *disconnected* modes in discrete spaces separated by near-zero-probability regions remains unresolved.

In this paper, we propose ***H**yperboli**c** **S**ecant-squared Gibbs-**S**ampling* (HiSS), a novel Metropolis-within-Gibbs sampling framework. Our method introduces a continuous auxiliary variable, $\boldsymbol{\theta}_a$, alongside the discrete primary variable, $\boldsymbol{\theta}$, modeled under a joint distribution. $\boldsymbol{\theta}_a$ encapsulates the multimodal discrete target distribution, allowing $\boldsymbol{\theta}$ to escape local modes and transition to more promising states while ensuring detailed balance with respect to the target distribution. This makes exploring the state space effective and improves mixing between states. We summarize our contributions as follows:

- We propose Hyperbolic Secant-Squared Gibbs Sampling (HiSS), a gradient-based algorithm for sampling multimodal discrete distributions. The core of HiSS is a logistic convolution kernel that bridges isolated modes, enabling the discrete sampler to efficiently traverse disconnected regions.
- We prove that HiSS satisfies detailed balance with respect to the discrete target distribution, ensuring asymptotic correctness. Furthermore, we establish a non-asymptotic convergence guarantee for HiSS with Metropolis-Hastings correction in locally log-concave discrete distributions.
- We present extensive experimental results, demonstrating the superiority of HiSS over standard gradient-based methods and other sampling techniques in multimodal settings. Our evaluations span diverse tasks, including Bernoulli distributions, Ising models, binary Bayesian neural networks, and combinatorial optimization. In addition to convergence to the true target distributions, we also analyze runtime performance and sample diversity. To promote reproducibility, we release the code at `https://github.com/pinakirm/HiSS`.

2 RELATED WORKS

Gradient-Based Discrete Sampling. Gradient-based discrete sampling methods have significantly advanced, enhancing efficiency and applicability. Locally informed proposals (Zanella, 2020) proposals using probability ratios improved convergence, while later gradient-based approaches (Grathwohl et al., 2021a) to approximate the probability ratio, boosted efficiency . Subsequent works proposed various gradient-based techniques (Rhodes and Gutmann, 2022; Sun et al., 2022a,b, 2023; Xiang et al., 2023). Discrete Langevin Proposal (DLP) (Zhang et al., 2022) extended the Langevin algorithm from continuous to discrete spaces, enabling parallel updates for all coordinates.

Multimodal Sampling in Continuous Spaces. In continuous spaces, techniques like simulated tempering (Marinari and Parisi, 1992), cyclical step sizes (Zhang et al., 2020a), parallel tempering (Swendsen and Wang, 1986), flat histograms (Berg and Neuhaus, 1992; Deng et al., 2020), and Wolff algorithm (Wolff, 1989) have been widely employed to facilitate mode transitions.

More recently, Diffusive Gibbs (DiGs) (Chen et al., 2024) used Gaussian Convolution on continuous multimodal distributions. HiSS and DiGS enhance Markov chain mixing efficiency through auxiliary structures, but extending DiGS to discrete distributions is not straightforward: DiGS utilizes Gaussian proposals($q(x|\tilde{x}^{i-1})$) centered around scaled noisy samples and a denoising posterior($p(x|\tilde{x}^{i-1})$). To tackle multimodality in discrete spaces, one may think of combating these problems using DLP, but because the former does not involve gradient information, DLP in its true form cannot be directly applied. HiSS accelerates convergence in discrete distributions using a joint-hybrid distribution (Lemma 4.1), while DiGS's joint distribution is purely continuous. In our hybrid setting, noising and denoising steps involve state-switching strategies, which are non-trivial and require careful design. HiSS introduces a logistic convolutional kernel, unlike DiGS's Gaussian convolutional kernel. This structural shift is detailed in Section A. HiSS also distinguishes itself by simplicity. Unlike DiGS, which relies on a complex hyperparameter schedule (the VP schedule), HiSS uses a static hyperparameter design. This makes HiSS easier to tune and offers competitive performance across various applications. Unlike DiGS, we provide convergence guarantees(Theorem 5.5) to substantiate the reliability of our proposed approach.

Multimodal Sampling in Discrete Spaces. Most prior works on multimodal discrete sampling rely on combinatorial or swap-based proposals like parallel tempering (Swendsen and Wang, 1986), Wang-Landau Sampling (Wang and Landau, 2001), Swedson-Wang Algo-

rithm (Swendsen and Wang, 1987), and early Mode-Jumping MCMC (Madras and Zheng, 2003). Based on gradient-based discrete samplers, Pynadath et al. (2024) introduced Automatic Cyclical Scheduling (ACS), combining automatic tuning of cyclical step sizes and balancing schedules to encourage dynamic transitions between global exploration and localized moves within each cycle.

Hybrid Coupling MCMC Samplers. While both EDLP (Mohanty et al., 2025) and HiSS use a coupling mechanism between a discrete primary variable and an auxiliary continuous variable, while retaining only the discrete variable their objectives and principles differ: In EDMALA, $\boldsymbol{\theta}_a$ guides $\boldsymbol{\theta}$ to target and enhance sampling from flat modes while ensuring detailed balance. In HiSS, $\boldsymbol{\theta_a}$'s role is to fling $\boldsymbol{\theta}$ when stuck at a local mode while ensuring detailed balance with respect to the marginal. Another fundamental distinction lies in their theoretical guarantees. EDLP's convergence bounds pertain to the joint distribution (Theorems 5.5 and 5.6 in Mohanty et al. (2025)), while Theorem 5.5 establishes convergence strictly for the marginal distribution.

Diffusion Models. Diffusion models, a powerful class of generative models, operate by iteratively injecting Gaussian noise into data and gradually denoising it to generate new samples (Sohl-Dickstein et al., 2015). This core paradigm has been extended to discrete spaces, often requiring a trained neural network to approximate the reverse "scoring" process (Lou et al., 2024; Avdeyev et al., 2023; Sahoo et al., 2024). Our work adapts the noise-denoise mechanism as a means of dislodging and exploring the neighborhood under a pure MCMC framework.

Further, HiSS as a pure MCMC sampler, does not require training or a model-specific score network. It samples from a known (unnormalized) target distribution for probabilistic inference. In contrast, diffusion models learn an unknown data distribution for generative tasks or optimization. HiSS uses a heavy-tailed logistic kernel to bridge disconnected modes, unlike the standard Gaussian transition kernels in discrete diffusion models (Sahoo et al., 2024). HiSS uses a simple single noise-denoise cycle per Gibbs sweep, unlike the multi-step denoising schedules in many diffusion models.

3 PRELIMINARIES

Target Distribution. We define a target distribution over a discrete space using an energy function. The target distribution is given by $\pi(\boldsymbol{\theta}) = \frac{1}{Z}\exp(U(\boldsymbol{\theta}))$, where $\boldsymbol{\theta}$ is a d-dimensional discrete variable within domain $\boldsymbol{\Theta}$, $U(\boldsymbol{\theta})$ represents the energy function, and Z is the normalizing constant ensuring $\pi(\boldsymbol{\theta})$ is a proper probability distribution. We make the following assumptions consistent with the literature on gradient-based discrete sampling (Grathwohl et al., 2021a; Sun et al., 2022a; Zhang et al., 2022): 1. *Coordinate-wise Factorization*: The domain $\boldsymbol{\Theta}$ is factorized such that $\boldsymbol{\Theta} = \Pi_{i=1}^{d}\Theta_i$. 2. *Differentiable Energy Function*: The energy function U can be extended to a differentiable function in $\mathbb{R}^d$. This extension is crucial for applying gradient-based sampling methods, as it allows the use of gradient information.

Gradient Computation in Discrete Spaces. To enable gradient-based sampling in discrete spaces, we rely on the Functional Extension framework, discussed in Section 3 of Grathwohl et al. (2021a). Per this framework, discrete distributions defined by energy functions sometimes possess a 'natural differentiable extension': $U : \mathbb{R}^d \to \mathbb{R}$ to compute $\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$ at discrete points via standard automatic differentiation. This extension provides the necessary local geometry to guide various discrete gradient-based samplers like GWG (Grathwohl et al., 2021a), DMALA (Zhang et al., 2022), and ACS (Pynadath et al., 2024).

Discrete Langevin Proposal. The DLP is an extension of the Langevin algorithm tailored for discrete spaces (Zhang et al., 2022). At a given position $\boldsymbol{\theta}$, the proposal distribution $q(\cdot|\boldsymbol{\theta})$ determines the next position. The proposal distribution is formulated as:

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

where $Z_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ is the normalizing constant.

Logistic Convolution. In density estimation, the logistic convolutional kernel(sometimes referred to as the sech-squared convolutional kernel) smoothly facilitates with optimal bandwidth selection and enhances the modeling of complex data distributions (Aboelhadid et al., 2018). Thus, if the conditional distribution is of the form $p(\boldsymbol{\theta}_a \mid \boldsymbol{\theta}) \sim \textit{Logistic}(\boldsymbol{\theta}, \eta)$, where $\eta > 0$ is the scaling parameter. Then, for original distribution $p(\boldsymbol{\theta})$, the convolved distribution $p(\boldsymbol{\theta}_a)$ can be expressed as:

$$p(\boldsymbol{\theta}_a) = \sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} p(\boldsymbol{\theta}_a \mid \boldsymbol{\theta})p(\boldsymbol{\theta}) \quad (2)$$

By ensuring non-negligible density paths between distant modes, the kernel improves connectivity in $p(\boldsymbol{\theta}_a)$ relative to $p(\boldsymbol{\theta})$, capturing both sharp transitions and smooth variations in the probability landscape.

4 HYPERBOLIC SECANT-SQUARED GIBBS SAMPLING

In this section, we present our sampling framework to address the limitations of gradient-based methods in

traversing disconnected modes. To achieve this, we introduce an auxiliary variable $\boldsymbol{\theta}_a$ and explain how it smooths the disconnected energy landscape through a well-defined joint probability distribution. We then detail our sampling strategy, outlining each sub-step of the algorithm.

4.1 Joint Distribution

To facilitate efficient exploration between modes, we aim to derive a smoothed version of the target discrete distribution, $p(\boldsymbol{\theta}) \propto \exp(U(\boldsymbol{\theta}))$. Inspired by the logistic kernel in (2), we define the smoothed target distribution:

$$p(\boldsymbol{\theta}_a) \propto \sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \exp\left\{U(\boldsymbol{\theta}) - 2\ln\left(\cosh\left(\frac{\boldsymbol{\theta}_a - \boldsymbol{\theta}}{2\eta}\right)\right)\right\} \tag{3}$$

$p(\boldsymbol{\theta}_a)$ represents a continuous version of $p(\boldsymbol{\theta})$, connecting otherwise isolated modes.

Why Logistic Convolutional Kernel? Our choice of a logistic convolutional kernel is a core design decision driven by both theoretical and practical considerations. Unlike Gaussian kernels, which rapidly decay and concentrate mass near the current mode, the logistic kernel's slower tail decay and broader spread facilitate better bridging across disconnected regions. Compared to a Gaussian kernel, the logistic kernel is also less sensitive to hyperparameter tuning, ensuring stable $p(\boldsymbol{\theta}_a)$. Notably, it retains support over $\mathbb{R}^d$, similar to the Gaussian kernel, but avoids pathologies observed in other heavy-tailed distributions (e.g., undefined moments in Cauchy, non-differentiability in Laplace). Furthermore, generating logistic noise is computationally efficient, due to closed-form inverse CDF sampling. These attributes make it a robust and scalable choice for discrete MCMC (Kingma and Welling, 2013; Maddison et al., 2017). We empirically demonstrate the advantages of logistic kernel in Section 6.5 and Appendix A.

Inspired by the coupling method used in Mohanty et al. (2025), we couple $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_a$ as follows:

Lemma 4.1. *Given $\widetilde{\boldsymbol{\theta}} = [\boldsymbol{\theta}^T, \boldsymbol{\theta}_a^T]^T$, the joint distribution $p(\widetilde{\boldsymbol{\theta}})$ is:*

$$p(\widetilde{\boldsymbol{\theta}}) = p(\boldsymbol{\theta}, \boldsymbol{\theta}_a) \propto \exp\left\{U(\boldsymbol{\theta}) - 2\ln\left(\cosh\left(\tfrac{\boldsymbol{\theta}_a - \boldsymbol{\theta}}{2\eta}\right)\right)\right\} \tag{4}$$

By construction, the marginal distributions of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_a$ are the original distribution $p(\boldsymbol{\theta})$ and the smoothed distribution $p(\boldsymbol{\theta}_a)$(Eq. 3).

One notices, the joint hybrid variable $\widetilde{\boldsymbol{\theta}}$ lies in a product space where first d coordinates are discrete-valued and the remaining d coordinates lie in $\mathbb{R}^d$ i.e. $\widetilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta} \times \mathbb{R}^d$ and its energy function can be expressed as

$$U(\widetilde{\boldsymbol{\theta}}) = U(\boldsymbol{\theta}) - 2\ln\left(\cosh\left(\frac{\boldsymbol{\theta}_a - \boldsymbol{\theta}}{2\eta}\right)\right).$$

4.2 Gibbs-like Update Procedure

Since $p(\boldsymbol{\theta}_a)$ is a smoothed version of the target distribution, samples from $p(\boldsymbol{\theta}_a)$ can effectively serve as launchpads for exploring the target discrete distribution $p(\boldsymbol{\theta})$. However, sampling directly from $p(\boldsymbol{\theta}_a)$ is intractable due to its summation form. In contrast, the conditional distribution $p(\boldsymbol{\theta}|\boldsymbol{\theta}_a)$ remains tractable and is given by:

$$p(\boldsymbol{\theta}|\boldsymbol{\theta}_a) \propto \frac{1}{Z_{\boldsymbol{\theta}_a}} \exp\left\{U(\boldsymbol{\theta}) - 2\ln\left(\cosh\left(\frac{\boldsymbol{\theta}_a - \boldsymbol{\theta}}{2\eta}\right)\right)\right\}, \tag{5}$$

where $Z_{\boldsymbol{\theta}_a}$ is the associated normalization constant. Moreover, its gradient also becomes easy to compute, i.e.

$$\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\boldsymbol{\theta}_a) = \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) + \frac{1}{\eta}\tanh\left(\frac{\boldsymbol{\theta}_a - \boldsymbol{\theta}}{2\eta}\right)$$

. This allows us to use gradient-based discrete samplers like DMALA (Zhang et al., 2022). Similarly, the conditional distribution for the auxiliary variable $\boldsymbol{\theta}_a$ is, $p(\boldsymbol{\theta}_a|\boldsymbol{\theta}) \sim \mathit{Logistic}(\boldsymbol{\theta}, \eta)$, where we can perform sampling easily as follows,

$$\boldsymbol{\theta}_a = \boldsymbol{\theta} + \eta\xi, \quad \xi \sim \text{Logistic}(0,1)^d. \tag{6}$$

4.3 Sampling Algorithm

HiSS utilizes a Gibbs sampler to sample from the joint distribution specified in Equation (4). It alternates between (1) perturbing the discrete state $\boldsymbol{\theta}$ via a logistic noising step to encourage exploration, (2) recovering a candidate discrete state $\boldsymbol{\theta}'_{\text{init}}$ through a denoising step grounded in a coordinate-wise logistic energy model, (3) accepting the proposal using a Metropolis-Hastings correction, and (4) refining the sample with a gradient-based sampler. This combination enables both mode hopping and local exploitation, facilitating efficient sampling in disconnected and multimodal discrete landscapes. We provide an ablation study for HiSS in Section 6.5.

4.3.1 Noising

To facilitate exploration of new cluster of modes beyond the current discrete mode, we perturb $\boldsymbol{\theta}$ by adding noise scaled by a factor η to get $\boldsymbol{\theta}_a$, encouraging the sampler to escape local modes and explore diverse regions of the state space. Mathematically, the noising process is defined in Equation (6).

4.3.2 Denoising

To recover a discrete proposal state from the previously noised sample $\boldsymbol{\theta}_a$, we will map $\boldsymbol{\theta}_a$ back to a discrete state. To ensure that this mapping favors discrete points close to $\boldsymbol{\theta}_a$ in a probabilistically consistent way, we define an energy function based on the negative log-density of a logistic distribution. For each coordinate i, we compute the energy for every candidate discrete state $\theta_i' \in \Theta_i$. To see this, we write $q_{\text{denoise}}(\boldsymbol{\theta}'|\boldsymbol{\theta}_a) = \prod_{i=1}^{d} q_{\text{denoise}_i}(\theta_i'|\boldsymbol{\theta}_a)$, where $q_{\text{denoise}_i}(\theta_i'|\boldsymbol{\theta}_a)$ has the energy:

$$U(\theta_i') = -2\ln\left(\cosh\left(\frac{\theta_{a\,i} - \theta_i'}{2\eta}\right)\right)$$

Using these energies, we compute the probabilities for each candidate discrete state via the softmax function. These probabilities form a categorical distribution over $\boldsymbol{\Theta}$, and thus $\boldsymbol{\theta}'_{init}$ is sampled as:

$$\boldsymbol{\theta}'_{init} \sim \text{Categorical}(\text{Softmax}(U(\theta_i'))). \tag{7}$$

It is worth noting that the denoising is independent of the original sample $\boldsymbol{\theta}$.

4.3.3 Metropolis-Hastings Acceptance for Denoised Proposal

We apply the Metropolis-Hastings (MH) step for the newly proposed discrete state:

$$a_{\text{init}}(\boldsymbol{\theta}'_{\text{init}}|\boldsymbol{\theta}^{(i-1)}) = \min\left(1, \frac{\pi(\boldsymbol{\theta}'_{\text{init}})q_{\text{noise}}(\boldsymbol{\theta}_a^{(i-1)}|\boldsymbol{\theta}'_{\text{init}})q_{\text{denoise}}(\boldsymbol{\theta}^{(i-1)}|\boldsymbol{\theta}_a^{(i-1)})}{\pi(\boldsymbol{\theta}^{(i-1)})q_{\text{noise}}(\boldsymbol{\theta}_a^{(i-1)}|\boldsymbol{\theta}^{(i-1)})q_{\text{denoise}}(\boldsymbol{\theta}'_{\text{init}}|\boldsymbol{\theta}_a^{(i-1)})}\right) \tag{8}$$

The MH step ensures the proposed state is from a high-density region $p(\boldsymbol{\theta})$, thereby providing a good initialization for the next step. We refer to this step as the MwG (Metropolis-within-Gibbs) step.

4.3.4 Gradient-based Denoising

After the MH step, HiSS runs a discrete gradient-based sampler Φ (e.g., GWG, DMALA), initialized from the accepted state $\boldsymbol{\theta}_{\text{init}}^{(i)}$, to sample from $p(\boldsymbol{\theta}|\boldsymbol{\theta}_a)$ in (5). For this work, we use DMALA for theoretical results (Section 5) and experiments (Section 6). Detailed implementation specifics are provided in Appendix B. We will collect samples of $\boldsymbol{\theta}$, as the marginal distribution of $p(\widetilde{\boldsymbol{\theta}})$ over $\boldsymbol{\theta}$ yields our desired target distribution.

Comparison to DiGS. HiSS is specifically designed for discrete multimodal distributions with isolated modes. Unlike DiGS (Chen et al., 2024), which operates solely in continuous spaces, HiSS targets a hybrid joint distribution where the primary variable $\boldsymbol{\theta}$ is discrete and the auxiliary variable $\boldsymbol{\theta}_a$ is continuous. To couple the two, HiSS replaces DiGS's Gaussian kernel and sensitive VP schedule with a logistic convolution kernel with static tuning, leading to a simpler formulation with theoretical guarantees.

Algorithm 1 Hyperbolic Secant-squared Gibbs-Sampling (HiSS)

Inputs:
Main variable $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, Auxiliary variable $\boldsymbol{\theta}_a \in \mathbb{R}^d$, Main stepsize α, Scale parameter η, Number of Gibbs Sweeps G, Discrete gradient-based Sampler Φ, Number of denoising steps L
Initialize: $\mathcal{S} \leftarrow \emptyset$
loop
  for $i \leftarrow 1$ to G **do**
    1. Sample $\boldsymbol{\theta}_a^{i-1} \sim q_{\text{noise}}(\boldsymbol{\theta}_a|\boldsymbol{\theta})$ using proposal in Equation 6.
    2. Propose $\boldsymbol{\theta}'_{init} \sim q_{\text{denoise}}(\boldsymbol{\theta}|\boldsymbol{\theta}_a^{i-1})$ using proposal in Equation 7.
    3. Accept $\boldsymbol{\theta}_{init}^{(i)} \leftarrow \boldsymbol{\theta}'_{init}$ with probability in (8). Otherwise, set $\boldsymbol{\theta}_{init}^{(i)} \leftarrow \boldsymbol{\theta}^{i-1}$.
    4. Sample $\boldsymbol{\theta}^{(i)}$ using Φ for L steps from the initial point $\boldsymbol{\theta}_{init}^{(i)}$ conditioned on $\boldsymbol{\theta}_a^{i-1}$.
  end for
  Update $\mathcal{S} \leftarrow \mathcal{S} \cup \{\boldsymbol{\theta}^{(G)}\}$
end loop
Output: $\mathcal{S}$

5 THEORETICAL ANALYSIS

In this section, we provide asymptotic and non-asymptotic convergence guarantees for HiSS. We make similar assumptions as in Pynadath et al. (2024). Those are as follows,

Assumption 5.1. *The function $U(\cdot) \in C^2(\mathbb{R}^d)$ has M-Lipschitz gradient. Note that it implicitly assumes that the set in domain $\boldsymbol{\Theta}$ is finite. We define conv($\boldsymbol{\Theta}$) as the convex hull of the set $\boldsymbol{\Theta}$.*

Assumption 5.2. *For each $\boldsymbol{\theta} \in \mathbb{R}^d$, there exists an open ball containing $\boldsymbol{\theta}$ of some radius $r_{\boldsymbol{\theta}}$, denoted by $B(\boldsymbol{\theta}, r_{\boldsymbol{\theta}})$, such that the function $U(\cdot)$ is $m_{\boldsymbol{\theta}}$-strongly concave in $B(\boldsymbol{\theta}, r_{\boldsymbol{\theta}})$ for some $m_{\boldsymbol{\theta}} > 0$.*

We define $\text{diam}(\boldsymbol{\Theta}) = \sup_{\boldsymbol{\theta},\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$, and $\Delta(\boldsymbol{\Theta}) = \sup_{\boldsymbol{\theta},\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|_1$. Finally, we define $a \in \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\nabla U(\boldsymbol{\theta})\|$ as the set of values which minimizes the energy function in $\boldsymbol{\Theta}$. Assumptions 5.1 ,5.2 are standard in optimization and sampling literature (Bottou et al., 2018; Dalalyan, 2017; Durmus and Moulines, 2017). Under Assumption 5.2, $U(\cdot)$ is m-strongly concave on conv($\boldsymbol{\Theta}$), following Lemma C.3 from Pynadath et al. (2024). The total variation distance between two probability measures μ and ν, defined on some space $\boldsymbol{\theta} \subset \mathbb{R}^d$ is $\|\mu - \nu\|_{TV} = \sup_{A \subseteq B(\boldsymbol{\theta})} |\mu(A) - \nu(A)|$, where $B(\boldsymbol{\theta})$ is the set of all measurable sets in $\boldsymbol{\theta}$.

Proposition 5.3. *HiSS (Algorithm 1) generates an irreducible and recurrent Markov chain for the joint distribution $\pi(\widetilde{\boldsymbol{\theta}})$.*

This theoretical guarantee for HiSS holds true for any choice of Φ.

5.1 Convergence Analysis for HiSS (with DMALA)

The following results hold true for DMALA as Φ in Algorithm 1.

Proposition 5.4. *HiSS (Algorithm 1) ensures detailed balance with respect to the target distribution $\pi(\boldsymbol{\theta})$.*

Based on Proposition 5.4, we focus on the marginal distribution $\pi(\boldsymbol{\theta})$, we establish a non-asymptotic convergence guarantee for HiSS using a uniform minorization argument.

Theorem 5.5. *Under Assumptions 5.1 ,5.2 and $\alpha < \frac{2}{M}$ in Algorithm 1, the marginal Markov chain P is uniformly ergodic under,*

$$\|P^k(x,\cdot) - \pi\|_{TV} \leq (1-\epsilon_\alpha)^k$$

where,

$$\epsilon_\alpha = \exp\left\{\begin{array}{l}\left(-M\left(\frac{LG}{2}+G\right) - \frac{LG}{\alpha} + \frac{mLG}{4}\right)\operatorname{diam}(\boldsymbol{\Theta})^2 \\ + \left(\frac{G\sqrt{d}(3L-2)+LG}{\eta} - \left(\frac{LG}{2}+G\right)\|\nabla U(a)\|\right)\operatorname{diam}(\boldsymbol{\Theta})\end{array}\right\}$$

As $\alpha \to 0$, $\epsilon_\alpha \to 0$, causing the convergence factor $1-\epsilon_\alpha$ to approach 1. This slows the convergence rate, as the chain takes longer to approach the stationary distribution. As dimension d grows, we may need to increase η (propose even broader jumps) to maintain adequate level of mixing and faster convergence. Refer to Appendix C for detailed proofs.

Theorem 5.6. *Under Assumptions 5.1 ,5.2 and $\alpha < \frac{2}{M}$ in Algorithm 1, for Markov chain P, for any real-valued function f and samples $X_1, X_2, X_3, \cdots, X_n$ from P, one has*

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \sum_{\theta\in\Theta} f(\theta)\pi(\theta)\right) \xrightarrow{d} N(0, \tilde{\sigma}_*^2)$$

for some $\tilde{\sigma}_* > 0$ as $n \to \infty$.

Proof. Theorem 5.6 is true due to direct consequence of using Theorem 5.5 and Jones (2004)[Corollary 5]. □

6 EXPERIMENTS

We evaluated HiSS empirically, showing it mixes faster and more efficiently than existing methods in various discrete multimodal settings. Our setups are inspired by Zhang et al. (2022); Mohanty et al. (2025). We compare HiSS against some popular pure gradient-based baselines like Gibbs with Gradient (GWG) (Grathwohl et al., 2021a), Discrete Metropolis-Adjusted Langevin Algorithm (DMALA) (Zhang et al., 2022). Other baselines targeting discrete multimodal distributions such as, Automatic Cyclical Sampler (ACS) (Pynadath et al., 2024), and Parallel Tempering(PT) (Swendsen and Wang, 1986) are also included. Being consistent with Chen et al. (2024), HiSS and PT both employ DMALA as their base sampler. For PT, for most tasks, for every original chain we use 5 temperature chains with geometric temperature scaling, employing consecutive swaps between adjacent chains (Earl and Deem, 2005; Kone and Kofke, 2005). We use the same stepsize α for DMALA, ACS, PT+DMALA, and HiSS. We ensure the number of iterations per sample for all samplers across all tasks is the same for fair runtime comparison.

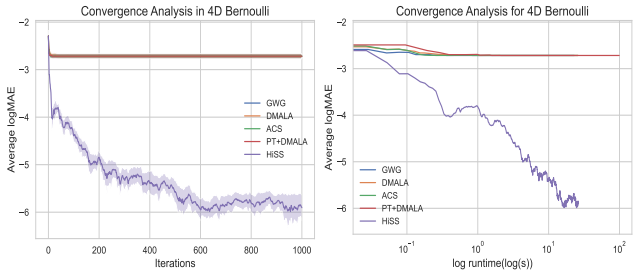

Figure 2: 4D Joint Bernoulli

6.1 Motivational Synthetic Example

We consider sampling from a Joint Quadrivariate Bernoulli Distribution, a multivariate distribution where each of the four binary random variables can take on the value 0 or 1. Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)$ be a 4-dimensional binary random vector. The joint probability distribution is specified by $p_{\boldsymbol{\theta}}$, which represents the probability of the vector $(\theta_1, \theta_2, \theta_3, \theta_4)$. For a given state $\boldsymbol{\theta}$ the energy function is given by :

$$U(\boldsymbol{\theta}) = \sum_{a\in\{0,1\}^4}\left(\prod_{n=1}^{4}\theta_n^{a_n}(1-\theta_n)^{1-a_n}\right)\ln p_a,$$

Details of the target distribution, featuring isolated modes separated by low-density energy, and its visualization can be found in Figure 11 of the Appendix. For each sampler, we ran 10 parallel chains for 1000 iterations. Even in this simple setup, gradient-based samplers like GWG and DMALA struggle to converge effectively. ACS and PT, methods designed to handle multimodal distributions, also perform poorly. In contrast, HiSS, demonstrates impressive performance, achieving the lowest log Mean Absolute Error (MAE). PT, as a high-resource sampling method, takes the longest due to inter-temperature chain communication overhead. HiSS, however, achieves remarkable convergence with respect

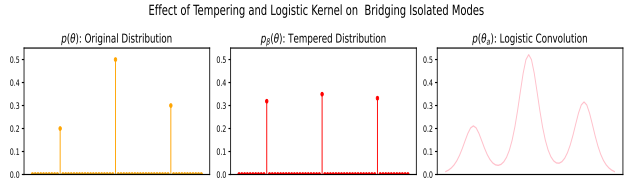

Figure 3: Comparison of the target distribution, tempered distribution, and $p(\boldsymbol{\theta}_a)$ in HiSS.

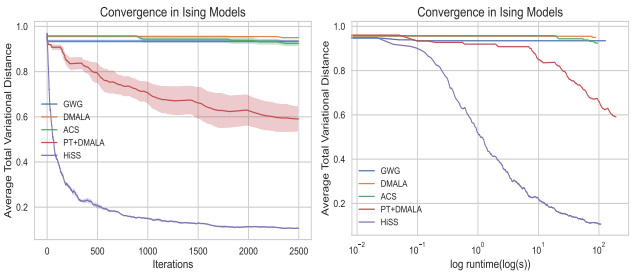

Figure 4: Ising Model

to runtime, demonstrating efficiency and accuracy (Figure 2). We provide hyperparameter settings, additional results, and diagnostics in the Appendix D.1.

Why HiSS Outperforms PT? HiSS significantly outperforms PT because PT adjusts the inverse temperature $\beta = \frac{1}{T}$ to enhance exploration, but disconnected modes still remain inaccessible since $p(\boldsymbol{\theta})^\beta = 0 \quad \forall \beta > 0$ when $p(\boldsymbol{\theta}) = 0$. In contrast, the Logistic Convolution kernel assigns strictly positive mass across the target distribution (see Figure 3).

6.2 Sampling from Ising Models

We consider a 3 by 3 lattice Ising model with random variable $\boldsymbol{\theta} \in \{-1, 1\}^d$, and $d = 3 \times 3 = 9$. The energy function is,

$$U(\boldsymbol{\theta}) = a\boldsymbol{\theta}^\top \mathbf{W}\boldsymbol{\theta} + b\boldsymbol{\theta},$$

where $\mathbf{W}$ is the interaction matrix, $a = 0.5$ is the connectivity strength and $b = 0.1$ is the bias.

We run 5 chains independently for each sampler for 2500 iterations and report the average total variational distance (TVD) between the estimated and true distribution with standard error against iterations and runtime in Figure 4. The convergence analysis reveals that HiSS significantly outperforms other samplers, achieving the lowest TVD (≈ 0.15) with consistent and rapid convergence. In contrast, PT+DMALA shows delayed convergence, while GWG, DMALA, and ACS fail to effectively navigate the distribution, remaining stuck at higher TVD levels. HiSS achieves convergence faster than other methods, while PT+DMALA has slower runtime performance due to high communication overhead.

Table 1: Performance comparison of different samplers on the `eil14` dataset.

Sampler	Cost ↓	PMC ↑	Jaccard ↓	Unique Solns. ↑
GWG	370.7086 ± 19.7404	27.0000 ± 0.0000	0.5555 ± 0.0000	2
DMALA	339.0111 ± 43.5156	84.2444 ± 23.4995	0.1483 ± 0.1481	10
ACS	382.8639 ± 2.5501	**105.0000** ± 0.0000	**0.0370** ± 0.0000	2
PT+DMALA	337.1584 ± 44.9881	80.4167 ± 19.7419	0.1402 ± 0.1357	9
HiSS	**277.9008** ± 19.8467	103.0727 ± 9.1807	0.0990 ± 0.0964	**11**

GWG, DMALA, and ACS show limited improvement. We provide hyperparameter settings, additional results, and diagnostics in the Appendix D.2.

6.3 Traveling Salesman Problem

In the Traveling Salesman Problem (TSP), the objective is to determine the shortest route that visits n cities exactly once before returning to the starting location. $U(\boldsymbol{\theta})$ is designed to capture the total cost of a particular route configuration $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$. The expression of $U(\boldsymbol{\theta})$ is:

$$U(\boldsymbol{\theta}) = -\Big(\sum_{i=1}^{n-1} \sqrt{(x_i - x_{i+1})^2 + (y_i - y_{i+1})^2} + \sqrt{(x_n - x_1)^2 + (y_n - y_1)^2}\Big)$$

where each city location is given by $\theta_i = (x_i, y_i) \in \mathbb{R}^2$. The final term, $\sqrt{(x_n - x_1)^2 + (y_n - y_1)^2}$, ensures that the route forms a closed loop, thereby completing the tour.

We use the *eil14* dataset, a 14-city problem extracted from the 51-city Christofides/Eilon instance, available in the TSPLIB95 benchmark repository, first introduced in Reinelt (1991). The dataset consists of 2D Euclidean coordinates of cities, and the objective is to find the shortest tour visiting all cities exactly once. The dataset is publicly available.[1] Each solution is represented as a binary square matrix of size n, where each row corresponds to a position in the tour and each column corresponds to a city. If a proposed solution violates the uniqueness of city visits, we reject the sample and retain the current solution. In TSP, the solution space grows exponentially with the number of cities, reaching $\mathcal{O}(n!)$. We execute each sampler for 10000 iterations.

Table 1 highlights HiSS's superior balance between solution quality, consistency, and solution exploration quality in combinatorial optimization. HiSS achieves the lowest cost with minimal variance, ensuring both optimality and stability, unlike PT+DMALA and DMALA, which generate competitive solution counts but at significantly higher costs. HiSS produces structurally diverse solutions, as measured by Pairwise Mismatch Count (PMC), inspired by Pairwise Order Discrepancy (Zaefferer et al., 2014; Liao et al., 2012; Vinyals et al., 2015) and Kendall's Tau (Fagin et al., 2003), and Jaccard similarity (Li et al., 2022) while ensuring samples dis-

[1] GitHub Repository: `https://github.com/jam7/tsp/tree/master`; TSPLIB Archive: `http://comopt.ifi.uni-heidelberg.de/software/TSPLIB95/`.

Table 2: Experiment results with binary Bayesian neural networks on four datasets.

Dataset	Average Test Log-likelihood (↑)					Average Test Root-Mean Square Error (↓)				
	GWG	DMALA	ACS	PT+DMALA	HiSS	GWG	DMALA	ACS	PT+DMALA	HiSS
Breast Cancer	-0.0241 ± 0.0030	-0.0240 ± 0.0014	-0.0280 ± 0.0022	-0.0246 ± 0.0018	$\mathbf{-0.0237} \pm 0.0014$	0.1553 ± 0.0082	0.1550 ± 0.0047	0.1673 ± 0.0065	0.1568 ± 0.0060	$\mathbf{0.1541 \pm 0.0047}$
COMPAS	-0.2265 ± 0.0025	-0.2271 ± 0.0025	-0.2284 ± 0.0025	-0.2265 ± 0.0026	$\mathbf{-0.2237} \pm 0.0039$	0.4759 ± 0.0027	0.4766 ± 0.0026	0.4779 ± 0.0026	0.4759 ± 0.0027	$\mathbf{0.4729 \pm 0.0042}$
HIV	-0.7025 ± 0.1127	-0.7446 ± 0.0000	-0.7446 ± 0.0000	-0.7446 ± 0.0000	$\mathbf{-0.2551} \pm 0.0009$	0.8341 ± 0.0082	0.8629 ± 0.0000	0.8629 ± 0.0000	0.8629 ± 0.0000	$\mathbf{0.5050 \pm 0.0089}$
Blog	-0.2799 ± 0.0351	-0.2919 ± 0.0000	-0.2919 ± 0.0000	-0.2919 ± 0.0000	$\mathbf{-0.2136} \pm 0.0129$	0.5277 ± 0.0376	0.5403 ± 0.0000	0.5403 ± 0.0000	0.5403 ± 0.0000	$\mathbf{0.4620 \pm 0.0137}$

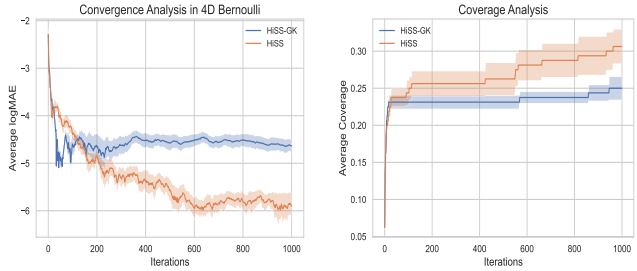

Figure 5: Gaussian Kernel vs Logistic Kernel

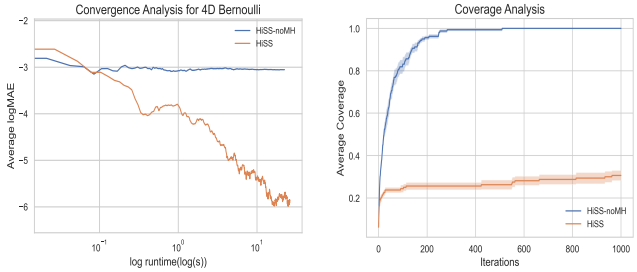

Figure 6: MH vs no-MH

covered are low-cost solutions. We provide additional insights and hyperparameter settings in Appendix D.4.

6.4 Binary Bayesian Neural Networks

The posterior distribution of binary neural networks (BNNs) (Courbariaux et al., 2016; Rastegari et al., 2016; Liu et al., 2021) is highly multimodal, characterized by disconnected or isolated modes (Zhang et al., 2020b; Izmailov et al., 2021). To investigate this, we perform regression tasks on four UCI datasets (Dua and Graff, 2017), defining the energy function as:

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

where $D = \{x_i, y_i\}_{i=1}^N$ is the training dataset, and f_θ represents a two-layer neural network with `Tanh` activation and 100 hidden neurons. To ensure a robust evaluation, we train 50 networks in parallel. We report the log-likelihood and RMSE on the test set along with their standard deviation (see Table 2). Notably, HiSS outperforms other baseline methods in terms of generalization to unseen data. Consistent with the findings of Chen et al. (2024) on their synthetic Bayesian Neural Networks Setup (Section 4.2), we hypothesize that this performance gain arises from HiSS's ability to capture a broader range of modes in the posterior distribution. We provide the dataset details in the Appendix D.5.

6.5 Ablation Study

We conduct an Ablation Study for HiSS using 4D Joint Bernoulli task under the same conditions as Section 6.

Gaussian Kernel vs Logistic Kernel. To empirically assess the benefits of the Logistic Kernel over the Gaussian Kernel, we replace HiSS's logistic convolution with Gaussian convolution and call it 'HiSS with Gaussian Kernel' (Hiss-GK). Setting $\sigma^2 = 0.9$ for Hiss-GK, HiSS converges faster than HiSS-GK, as shown by the rapid decrease in log mean absolute error (logMAE) in Figure 5. HiSS also achieves higher coverage of the target distribution across iterations. The random walk acceptance ratios for HiSS and HiSS-GK are similar, with values of 0.136 ± 0.109 and 0.144 ± 0.111, respectively. This improvement is due to the logistic kernel's superior mode-bridging capability, which facilitates better mixing and exploration of disconnected modes.

MH Step for Denoised Proposal.

From Figure 6, skipping this step reduces overall execution time, but it causes markedly early poor convergence to the marginal distribution. The MH step for the denoised proposal ensures convergence to the marginal distribution (Proposition 5.4). Since we accept the proposed state without MH correction, coverage quickly reaches 1, unlike the sampler that uses MH correction to select states and evaluate exploration potential. We use DMALA as the base sampler for both methods.

Effect of η. The scaling parameter η is arguably the most crucial hyperparameter for HiSS. We report the average logMAE, MwG acceptance probability, and coverage (defined in Appendix D) across various η values (Figure 7). We observe that for smaller η, HiSS exhibits high acceptance rates, poor coverage, and higher logMAE, as the sampler tends to retain the current state after the noise-denoise procedure, limiting exploration. As η increases, HiSS explores the state space more effectively, leading to improved coverage and con-

vergence (lower logMAE), albeit at the cost of reduced acceptance probability due to more targeted, aggressive proposals. Notably, the standard deviation in logMAE also increases with larger η, reflecting more dynamic movement through the state space rather than stagnation. Beyond a certain threshold of η (approx 1 here), the performance across all metrics saturates. We provide insights into tuning G and L in the Appendix D.2.

Impact of gradient refinement.

We investigate the behavior of HiSS when the gradient-based refinement step is omitted (i.e., setting $L = 0$). Theoretically, the resulting sampler, now consisting solely of the Noising, Denoising, and MH correction remains a valid MCMC kernel that satisfies detailed balance with respect to the marginal distribution $\pi(\boldsymbol{\theta})$. However, removing the gradient step fundamentally alters the sampler's exploration dynamics. Without gradient-informed updates (e.g., DMALA), the sampler loses its ability to exploit the local geometry of the energy landscape. The exploration now reduces to a random-walk behavior governed solely by the logistic kernel. While the kernel facilitates inter-mode jumps (global exploration), it is inefficient at intra-mode mixing (local exploration), particularly in high-dimensional spaces where the volume of the mode is large. We observed that the $L = 0$ variant exhibits rapid early convergence due to the reduced computational overhead (no gradient evaluations). However, this advantage is quickly negated as the sampler struggles to thoroughly explore the discovered modes as seen in Figure 8.

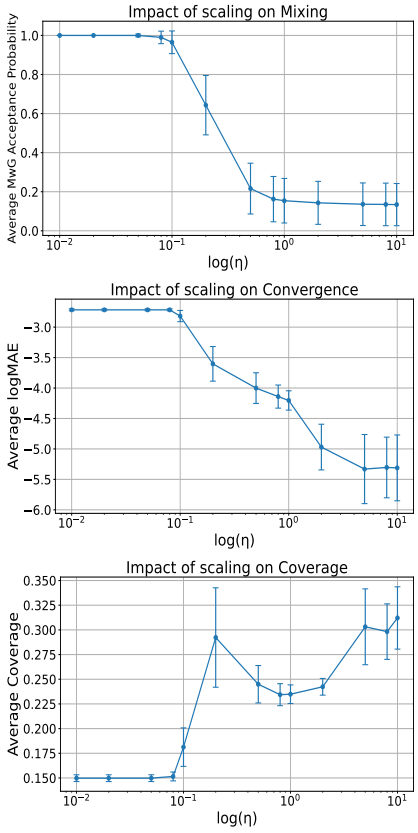

Figure 7: Sensitivity analysis of HiSS.

Limitations. While effective, HiSS faces several challenges. First, compared to any gradient-based sampler ran for LG steps, HiSS requires additional G MH steps, thereby slightly increasing runtime. Second, the denoised sample's MH acceptance rate, after tuning, remains low 13-14%. Designing asymmetric (e.g., Gumbel or skewed distributions), mode-aware intelligent proposals through landscape scouting can enhance efficiency by directing proposals to promising regions, potentially bypassing the additional MH step and reducing runtime. Third, our convergence analysis and theoretical guarantees are based solely on DMALA as the base refinement step. Performing a comparative analysis using other gradient-informed samplers (e.g., ACS, DULA, GWG) could provide additional insights.

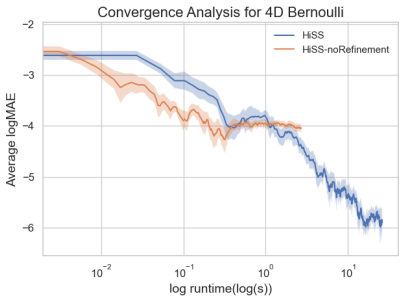

Figure 8: No Gradient Refinement for 4D Bernoulli

Most importantly, HiSS is designed on the premise that gradient-based methods are effective for local refinement, not global exploration. By introducing an auxiliary variable, we decouple these roles: the Logistic Kernel facilitates global mode-hopping, while the inner gradient sampler (e.g., DMALA) handles intra-mode mixing. This allows the system to escape isolated modes while maintaining high acceptance rates within basins. However, this hierarchical structure creates a dependency on the inner sampler. As analyzed by Chehab and Korba (2024), if the local basins exhibit extreme ruggedness or lack local log-concavity, making the local problem as difficult as the global one, the difficulty is shifted to the inner sampler, failing our two-step construction. HiSS assumes that the target landscape consists of basins that are locally amenable to gradient-based sampling.

Discussion. We introduce Hyperbolic Secant-Squared Gibbs Sampling (HiSS), a novel approach for exploring multimodal distributions in discrete spaces, especially those with distant modes. HiSS improves mixing efficiency through a logistic convolution kernel, enhancing the characterization of complex distributions. We provide both asymptotic and non-asymptotic convergence guarantees. Extensive experiments across spin glass systems, Bayesian inference, and combinatorial optimization show HiSS consistently outperforms existing sampling techniques. These findings highlight HiSS's potential for studying complex discrete distributions and its applications in various scientific fields.

7 CONCLUSION

Acknowledgment

The authors thank Dr. Rajiv Khanna for his helpful discussions, constructive feedback, and suggestions that improved the clarity and presentation of this work. RZ acknowledges support from NSF IIS-2508145 and Amazon Research Award.

References

Aboelhadid, M. et al. (2018). Logistic kernel density estimation: Statistical properties and optimal bandwidth selection. *International Journal of Computational Mathematical Sciences*.

Applegate, D. L., Bixby, R. E., Chvatal, V., and Cook, W. J. (2006). *The traveling salesman problem: A computational study*. Princeton University Press.

Avdeyev, P., Shi, C., Tan, Y., Dudnyk, K., and Zhou, J. (2023). Dirichlet diffusion score model for biological sequence generation. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 1276–1301. PMLR.

Banterle, M., Grazian, C., Lee, A., and Robert, C. P. (2019). Accelerating metropolis-hastings algorithms by delayed acceptance. *Foundations of Data Science*, 1(2):103–128.

Berg, B. A. and Neuhaus, T. (1992). Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Physical Review Letters*, 68(1):9–12.

Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*.

Bishop, C. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

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

Buza, K. (2014). BlogFeedback. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C58S3F.

Chaikin, P. M. and Lubensky, T. C. (1995). *Principles of Condensed Matter Physics*. Cambridge University Press.

Chehab, O. and Korba, A. (2024). A practical diffusion path for sampling.

Chen, W., Zhang, M., Paige, B., Hernández-Lobato, J. M., and Barber, D. (2024). Diffusive gibbs sampling.

Courbariaux, M., Hubara, I., Soudry, D., El-Yaniv, R., and Bengio, Y. (2016). Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830*.

Dalalyan, A. (2017). Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. In *Conference on Learning Theory*, pages 678–689. PMLR.

Deng, W., Lin, G., and Liang, F. (2020). A contour stochastic gradient langevin dynamics algorithm for simulations of multi-modal distributions. *Advances in neural information processing systems*, 33:15725–15736.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Durmus, A. and Moulines, E. (2017). Nonasymptotic convergence analysis for the unadjusted langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587.

Earl, D. J. and Deem, M. W. (2005). Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916.

Edwards, S. F. and Anderson, P. W. (1975). Theory of spin glasses. *Journal of Physics F: Metal Physics*, 5(5):965.

Fagin, R., Kumar, R., and Sivakumar, D. (2003). Comparing top k lists. *SIAM Journal on Discrete Mathematics*, 17(1):134–160.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, volume 2. Wiley.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Grathwohl, W., Swersky, K., Hashemi, M., Duvenaud, D., and Maddison, C. J. (2021a). Oops i took a gradient: Scalable sampling for discrete distributions. *International Conference on Machine Learning*.

Grathwohl, W., Wang, M. H., Liao, R., Swersky, K., and Duvenaud, D. (2021b). Oops i took a gradient: Scalable sampling for discrete distributions. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Gu, J., Bradbury, J., Xiong, C., Li, V. O., and Socher, R. (2018). Non-autoregressive neural machine translation. In *International Conference on Learning Representations*.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1):73–101.

Izmailov, P., Vikram, S., Hoffman, M. D., and Wilson, A. G. G. (2021). What are bayesian neural network posteriors really like? In *International conference on machine learning*, pages 4629–4640. PMLR.

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

Jones, G. L. (2004). On the Markov chain central limit theorem. *Probability Surveys*, 1(none):299 – 320.

Khanna, R., Hodgkinson, L., and Mahoney, M. W. (2021). Geometric rates of convergence for kernel-based sampling algorithms. In de Campos, C. and Maathuis, M. H., editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 2156–2164. PMLR.

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

Kone, A. and Kofke, D. A. (2005). Selection of temperature intervals for parallel-tempering simulations. *The Journal of Chemical Physics*, 122(20):206101.

Koyejo, O. O., Khanna, R., Ghosh, J., and Poldrack, R. (2014). On prior distributions and approximate inference for structured variables. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, pages 172–180. Curran Associates, Inc.

Lapedes, A. S., Giraud, B. G., Liu, L., and Stormo, G. D. (1999). Correlated mutations in models of protein sequences: phylogenetic and structural effects. In *Lecture Notes–Monograph Series*, pages 236–256. Institute of Mathematical Statistics.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, H., Zhang, M., and Zeng, C. (2022). Circular jaccard distance-based multi-solution optimization for traveling salesman problems. *Mathematical Biosciences and Engineering*, 19(5):4458–4480.

Liao, Y.-F., Yau, D.-H., and Chen, C.-L. (2012). Evolutionary algorithm to traveling salesman problems. *Computers & Mathematics with Applications*, 64(5):788–797. Advanced Technologies in Computer, Consumer and Control.

Liu, X., Ye, M., Zhou, D., and Liu, Q. (2021). Post-training quantization with multiple points: Mixed precision without mixed precision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8697–8705.

Livingstone, S., Betancourt, M., Byrne, S., and Girolami, M. (2019). Geometric foundations of hamiltonian monte carlo. *Bernoulli*, 25(4A):2257–2298.

Lou, A., Meng, C., and Ermon, S. (2024). Discrete diffusion modeling by estimating the ratios of the data distribution. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F., editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 32819–32848. PMLR.

Louizos, C., Welling, M., and Kingma, D. P. (2018). Learning sparse neural networks through l0 regularization. In *International Conference on Learning Representations*.

Maddison, C. J., Mnih, A., and Teh, Y. W. (2017). The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*.

Madras, N. and Zheng, D. (2003). Markov chain importance sampling. *Journal of Statistical Physics*, 112(1–2):293–312.

Marinari, E. and Parisi, G. (1992). Simulated tempering: a new monte carlo scheme. *Europhysics letters*, 19(6):451.

Meyn, S. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability*. Cambridge University Press, 2nd edition.

Mohanty, P., Bhattacharya, R., and Zhang, R. (2025). Entropy-guided sampling of flat modes in discrete spaces. In *NeurIPS 2025 Workshop on Structured Probabilistic Inference & Generative Modeling*.

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

Noé, F., Olsson, S., Köhler, J., and Wu, H. (2019). Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147.

Nummelin, E. (1984). *General Irreducible Markov Chains and Non-negative Operators*. Cambridge University Press.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American statistical association*, 103(482):681–686.

Pompe, E., Holmes, C. C., and Latuszy'nski, K. (2020). A framework for adaptive mcmc targeting multimodal distributions. *The Annals of Statistics*.

Pynadath, P., Bhattacharya, R., HARIHARAN, A. N., and Zhang, R. (2024). Gradient-based discrete sampling with automatic cyclical scheduling. In *ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling*.

Rastegari, M., Ordonez, V., Redmon, J., and Farhadi, A. (2016). Xnor-net: Imagenet classification using binary convolutional neural networks. In *European conference on computer vision*, pages 525–542. Springer.

Reinelt, G. (1991). TSPLIB—A Traveling Salesman Problem Library. *ORSA Journal on Computing*, 3(4):376–384.

Rgnvaldsson, T. (2015). HIV-1 protease cleavage. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5H03P.

Rhodes, B. and Gutmann, M. U. (2022). Enhanced gradient-based MCMC in discrete spaces. *Transactions on Machine Learning Research*.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer, New York.

Roberts, G. O. and Rosenthal, J. S. (2002). Langevin diffusions and metropolis-hastings algorithms. *Methodology and Computing in Applied Probability*, 4(4):337–357.

Roberts, G. O. and Tweedie, R. L. (1994). Simple conditions for the convergence of the metropolis-hastings algorithm. *Stochastic Processes and their Applications*, 49(2):207–216.

Sahoo, S. S., Arriola, M., Schiff, Y., Gokaslan, A., Marroquin, E., Chiu, J. T., Rush, A., and Kuleshov, V. (2024). Simple and effective masked diffusion language models. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C., editors, *Advances in Neural Information Processing Systems*, volume 37, pages 130136–130184. Curran Associates, Inc.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR.

Song, Y. and Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2021). Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.

Sun, H., Dai, H., Dai, B., Zhou, H., and Schuurmans, D. (2023). Discrete langevin samplers via wasserstein gradient flow. In *International Conference on Artificial Intelligence and Statistics*, pages 6290–6313. PMLR.

Sun, H., Dai, H., Xia, W., and Ramamurthy, A. (2022a). Path auxiliary proposal for mcmc in discrete space. In *International Conference on Learning Representations*.

Sun, H., Dai, H., Xia, W., and Ramamurthy, A. (2022b). Path auxiliary proposal for MCMC in discrete space. In *International Conference on Learning Representations*.

Swendsen, R. H. and Wang, J.-S. (1986). Replica monte carlo simulation of spin-glasses. *Physical Review Letters*, 57(21):2607.

Swendsen, R. H. and Wang, J.-S. (1987). Nonuniversal critical dynamics in monte carlo simulations. *Physical review letters*, 58(2):86.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1728.

Vinyals, O., Fortunato, M., and Jaitly, N. (2015). Pointer networks. In *Advances in Neural Information Processing Systems*, pages 2692–2700.

Wang, F. and Landau, D. P. (2001). Efficient, multiple-range random walk algorithm to calculate the density of states. *Physical review letters*, 86(10):2050.

Wolberg, W. H., Mangasarian, O. L., Street, N., and Street, W. (1993). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. DOI: `https://doi.org/10.24432/C5DW2B`.

Wolff, U. (1989). Collective monte carlo updating for spin systems. *Physical Review Letters*, 62(4):361.

Xiang, Y., Zhu, D., Lei, B., Xu, D., and Zhang, R. (2023). Efficient informed proposals for discrete distributions via newton's series approximation.

Zaefferer, M., Stork, J., and Bartz-Beielstein, T. (2014). Distance measures for permutations in combinatorial efficient global optimization. In Bartz-Beielstein, T., Branke, J., Filipič, B., and Smith, J., editors, *Parallel Problem Solving from Nature – PPSN XIII*, pages 373–383, Cham. Springer International Publishing.

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

Zhang, R., Li, C., Zhang, J., Chen, C., and Wilson, A. G. (2020a). Cyclical stochastic gradient mcmc for bayesian deep learning. In *International Conference on Learning Representations*.

Zhang, R., Li, C., Zhang, J., Chen, C., and Wilson, A. G. (2020b). Cyclical stochastic gradient mcmc for bayesian deep learning. In *International Conference on Learning Representations*.

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

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

All experiments were run on a single RTX A6000.

A Justification for Logistic Convolutional Kernel

A.1 Intuition

Our choice of using logistic convolution over Gaussian convolution stems from the need for a kernel with fatter tails, which ensures better exploration of multimodal energy landscapes. In high-dimensional problems, escaping local modes is essential, and the Gaussian kernel, with its thin tails, heavily centers the probability mass around the current mode, making mode escape less efficient (Bishop, 2006). The logistic distribution, on the other hand, exhibits slower tail decay ($e^{-|x|}$) than Gaussian (e^{-x^2}) (Feller, 1971), enabling it to retain intermediate mass in disconnected regions of the landscape. We specifically discard other heavy-tailed distributions like the Cauchy and Laplace distributions: the Cauchy distribution lacks finite moments, leading to numerical instabilities, and the Laplace distribution is non-differentiable at its mean, which can hinder gradient-based optimization (Huber, 1964).

A.2 Practical Implementation

Logistic noise can be generated in Python via the inverse transform sampling method, where the standard logistic distribution can be expressed as:

$$F^{-1}(u) = \ln\left(\frac{u}{1-u}\right),$$

with $u \sim \text{Uniform}(0, 1)$.

A.3 Auxiliary Distribution Characteristics

In Figure 9, we illustrate the auxiliary distributions for both the Gaussian kernel with a VP schedule and the Logistic kernel, alongside the target distribution, which in this case is $\mathit{Bernoulli}(0.7)$. For both kernels, as σ (Gaussian) and η (Logistic) increase, the coupling weakens, causing $p(\boldsymbol{\theta}_a)$ to become increasingly independent of $\boldsymbol{\theta}_a$. Notably, for the Gaussian kernel, when the signal strength diminishes ($\alpha \to 0$), $\sigma \to 1$, leading the auxiliary distribution to converge to the standard normal distribution ($p(\boldsymbol{\theta}_a) \to \mathcal{N}(0, 1)$).

While both kernels are capable of bridging the modes effectively, the Logistic kernel demonstrates a significant advantage due to its broader support. In contrast, the Gaussian kernel's support narrows sharply as its variance decreases, resulting in limited coverage of these discrete modes.

Additionally, the Logistic kernel provides a smooth and gradual transition as the scale parameter η changes, ensuring a consistent evolution of the auxiliary distribution. This consistency enables a balanced trade-off between exploration and exploitation. In comparison, the Gaussian kernel exhibits abrupt changes in the auxiliary distribution as the scale parameter σ shrinks (e.g., as α transitions from 0.95 to 0.99). These abrupt changes render the Gaussian kernel highly sensitive to hyperparameters.

A.4 Motivational Example: Mixture of Dirac Deltas

We consider a symmetric mixture of two Dirac delta distributions. This setting is theoretically significant, as kernel-based methods have been shown to achieve geometric convergence rates specifically when the target measure is atomic(sum of Dirac deltas) (Khanna et al., 2021). Thus,

$$p(x) = \frac{1}{2}\delta(x+\mu) + \frac{1}{2}\delta(x-\mu), \quad \mu > 0,$$

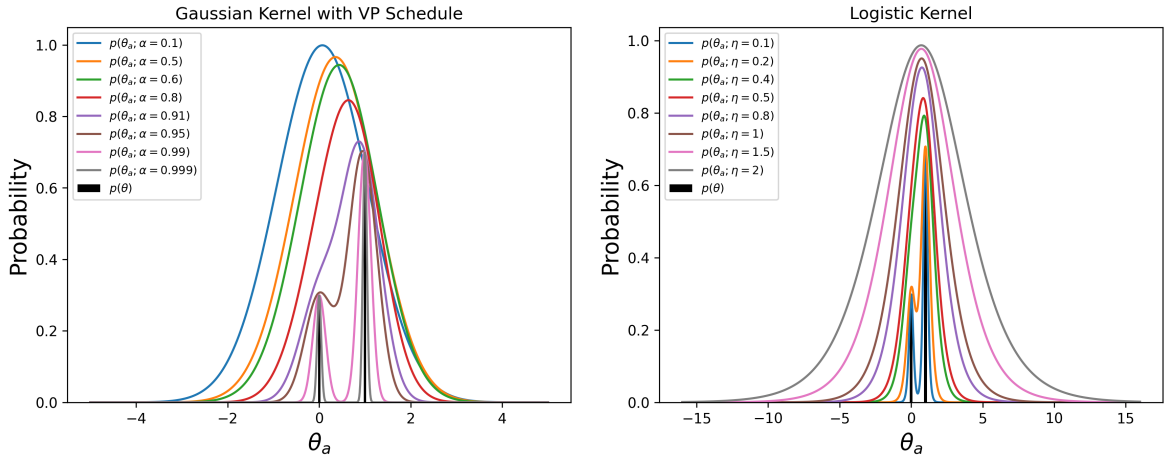

Figure 9: Mode Bridging in Bernoulli Distribution

where μ controls the separation between the two modes. Convolving $p(x)$ with a kernel $k(x-x')$ produces the smoothed distribution:

$$\tilde{p}(x) = \sum_{x' \in \{-\mu,\mu\}} p(x')k(x-x')$$

In order to measure the mode bridging tendency of the kernels, we wish to compute the **intermediate mass** in the ϵ-strip, defined as the probability mass within the region $|x| < \epsilon$ under $\tilde{p}(x)$(See Figure 10 for intuition).

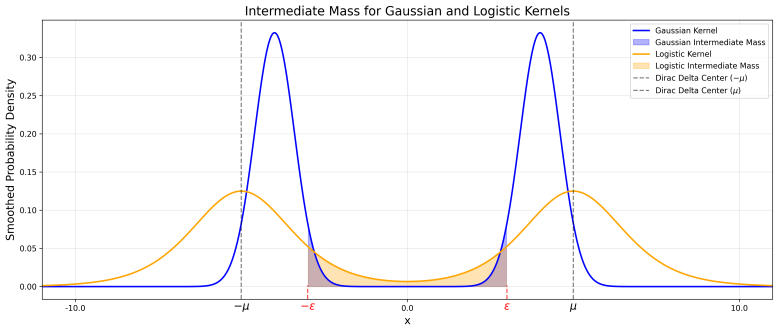

Figure 10: Intermediate Mass for Kernels

Mathematically,

$$\tilde{I}(\epsilon) = \int_{-\epsilon}^{\epsilon} \tilde{p}(x)\,dx, \quad \epsilon > 0,$$

Gaussian Kernel

Under the VP schedule inspired by Diffusion Models (Sohl-Dickstein et al., 2015; Song and Ermon, 2019; Ho et al., 2020), the Gaussian kernel is parameterized by α and σ, satisfying:

$$\alpha^2 + \sigma^2 = 1, \quad \alpha > 0, \sigma > 0,$$

The Gaussian kernel is given by:

$$k_{\mathrm{G}}(x-x') = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\alpha x')^2}{2\sigma^2}}.$$

The smoothed distribution becomes:

$$\tilde{p}_{\mathrm{G}}(x) = \frac{1}{2}\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x+\alpha\mu)^2}{2\sigma^2}} + \frac{1}{2}\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\alpha\mu)^2}{2\sigma^2}}$$

Logistic Kernel

The logistic kernel is parameterized by $\eta > 0$ and is defined as:

$$k_{\mathrm{L}}(x - x') = \frac{1}{4\eta}\mathrm{sech}^2\left(\frac{x - x'}{2\eta}\right),$$

where $\mathrm{sech}(z) = \frac{2}{e^z + e^{-z}}$.

The smoothed distribution becomes:

$$\tilde{p}_{\mathrm{L}}(x) = \frac{1}{2}\frac{1}{4\eta}\mathrm{sech}^2\left(\frac{x + \mu}{2\eta}\right) + \frac{1}{2}\frac{1}{4\eta}\mathrm{sech}^2\left(\frac{x - \mu}{2\eta}\right).$$

Now near $x = 0$,

$$\begin{aligned}\tilde{p}_{\mathrm{G}}(0) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(\alpha\mu)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-\mu^2}{2}\cdot\frac{1-\sigma^2}{\sigma^2}}\end{aligned}$$

$$\tilde{p}_{\mathrm{L}}(0) = \frac{1}{4\eta}\mathrm{sech}^2\left(\frac{\mu}{2\eta}\right) \tag{9}$$

For distant modes i.e. $\mu >>> \epsilon$, the intermediate mass can be approximated as,

$$\tilde{I}(\epsilon) \approx 2\epsilon.\tilde{p}(0)$$

Thus,

$$\begin{aligned}\tilde{I}_{\mathrm{G}}(\epsilon) &\approx 2\epsilon \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-\mu^2}{2}\cdot\frac{1-\sigma^2}{\sigma^2}} \\ &= \sqrt{\frac{2}{\pi}}\frac{\epsilon}{\sigma} e^{-\frac{\mu^2}{(\frac{2\sigma^2}{1-\sigma^2})}}\end{aligned}$$

$$\begin{aligned}\tilde{I}_{\mathrm{L}}(\epsilon) &\approx 2\epsilon \cdot \frac{1}{4\eta}\mathrm{sech}^2\left(\frac{\mu}{2\eta}\right) \\ &\approx 2\frac{\epsilon}{\eta} e^{-\frac{\mu}{\eta}}\end{aligned}$$

We notice, for $\mu \to \infty$, both kernels are highly sensitive to their respective parameters when the parameters are small, as the exponential decay dominates. The Logistic kernel is generally more robust to parameter variations compared to the Gaussian kernel, as its decay rate remains linear in μ, while the Gaussian kernel exhibits a steep quadratic decay that magnifies sensitivity to small σ^2.

The Gaussian kernel, due to its rapid decay ($e^{-\mu^2}$), heavily concentrates probability mass near the modes, making it inefficient for escaping local modes. In contrast, the logistic kernel, with its slower tail decay ($e^{-\mu}$), retains significantly more intermediate mass, allowing it to bridge distant isolated modes for effectively.

A.5 Empirical Assessment

See our discussion on Gaussian Kernel vs Logistic Kernel under Section 6.5, Figure 5.

B Conditional DLP

By incorporating the coupling between the variables, we refine the DLP proposal by replacing $\nabla U(\boldsymbol{\theta})$ with $\nabla_{\boldsymbol{\theta}} U_\eta(\widetilde{\boldsymbol{\theta}})$. This adjustment results in the modified proposal:

$$q^{\text{joint}}_{\text{DMALA}}(\widetilde{\boldsymbol{\theta}}^{(t)} \mid \widetilde{\boldsymbol{\theta}}^{(t-1)}) = \frac{1}{Z(\widetilde{\boldsymbol{\theta}}^{(t-1)})} \exp\left(\frac{1}{2}\nabla_{\boldsymbol{\theta}} U_\eta(\widetilde{\boldsymbol{\theta}}^{(t-1)})^\top(\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}) - \frac{1}{2\alpha}\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}\|^2\right) \tag{10}$$

and corresponding normalization constant is $Z(\widetilde{\boldsymbol{\theta}}^{(t-1)})$. To further simplify, we exploit *coordinate-wise factorization*, allowing us to express Eq. $q^{\text{joint}}_{\text{DMALA}}(\widetilde{\boldsymbol{\theta}}^{(t)} \mid \widetilde{\boldsymbol{\theta}}^{(t-1)}) = \prod_{i=1}^{d} q^{\text{joint}}_{\text{DMALA}_i}(\theta'_i|\widetilde{\boldsymbol{\theta}})$, where $q^{\text{joint}}_{\text{DMALA}_i}(\widetilde{\boldsymbol{\theta}}'_i|\widetilde{\boldsymbol{\theta}})$ is a categorical distribution:

$$\text{Categorical}\left(\text{Softmax}\left(\frac{1}{2}\nabla_\theta U_\eta(\widetilde{\boldsymbol{\theta}})_i(\theta'_i - \theta_i) - \frac{(\theta'_i - \theta_i)^2}{2\alpha}\right)\right). \tag{11}$$

We use the keyword *joint* specifically to denote that at the current joint position $\widetilde{\boldsymbol{\theta}}$, the proposal distribution $q^{\text{joint}}_{\text{DMALA}}(\cdot|\widetilde{\boldsymbol{\theta}})$ generates the next joint position, keeping $\boldsymbol{\theta}_a$ the same across both positions.

During the Metropolis-Hastings (MH) step, the conditional acceptance probability is calculated as:

$$a^{\text{joint}}_{\text{DMALA}}(\widetilde{\boldsymbol{\theta}}^{(t)}|\widetilde{\boldsymbol{\theta}}^{(t-1)}) = \min\left(1, \frac{q^{\text{joint}}_{\text{DMALA}}(\widetilde{\boldsymbol{\theta}}^{(t-1)})|\widetilde{\boldsymbol{\theta}}^{(t)})}{q^{\text{joint}}_{\text{DMALA}}(\widetilde{\boldsymbol{\theta}}^{(t)}|\widetilde{\boldsymbol{\theta}}^{(t-1)})} \cdot \frac{\pi(\widetilde{\boldsymbol{\theta}}^{(t)})}{\pi(\widetilde{\boldsymbol{\theta}}^{(t-1)})}\right). \tag{12}$$

where $\widetilde{\boldsymbol{\theta}}^{(t)} = [{\boldsymbol{\theta}^{(t)}}^T, {\boldsymbol{\theta}_a^{(i-1)}}^T]^T$and $\widetilde{\boldsymbol{\theta}}^{(t-1)} = [{\boldsymbol{\theta}^{(t-1)}}^T, {\boldsymbol{\theta}_a^{(i-1)}}^T]^T$.

C Proof of Theorems

C.1 Proof of Lemma 4.1

Proof. Assume $\widetilde{\boldsymbol{\theta}} = [\boldsymbol{\theta}^T, \boldsymbol{\theta}_a^T]^T$ is sampled from the joint posterior distribution:

$$p(\widetilde{\boldsymbol{\theta}}) = p(\boldsymbol{\theta}, \boldsymbol{\theta}_a) \propto \exp\left\{U(\boldsymbol{\theta}) - 2\ln\left(\cosh\left(\frac{\boldsymbol{\theta}_a - \boldsymbol{\theta}}{2\eta}\right)\right)\right\}. \tag{13}$$

Then the marginal distribution for $\boldsymbol{\theta}$ is:

$$\begin{aligned} p(\boldsymbol{\theta}) &= \int p(\boldsymbol{\theta}, \boldsymbol{\theta}_a) d\boldsymbol{\theta}_a \\ &= (4\eta)^{-d} Z^{-1} \int \exp\left\{U(\boldsymbol{\theta}) - 2\ln\left(\cosh\left(\frac{\boldsymbol{\theta}_a - \boldsymbol{\theta}}{2\eta}\right)\right)\right\} d\boldsymbol{\theta}_a \\ &= Z^{-1}\exp(U(\boldsymbol{\theta}))(4\eta)^{-d} \int \exp\left\{-2\ln\left(\cosh\left(\frac{\boldsymbol{\theta}_a - \boldsymbol{\theta}}{2\eta}\right)\right)\right\} d\boldsymbol{\theta}_a \\ &= Z^{-1}\exp(U(\boldsymbol{\theta})), \end{aligned} \tag{14}$$

where $Z = \sum_{\boldsymbol{\Theta}} \exp(U(\boldsymbol{\theta}))$ is the normalizing constant, and it is obtained by:

$$\sum_{\boldsymbol{\Theta}} \int \exp\left\{U(\boldsymbol{\theta}) - 2\ln\left(\cosh\left(\frac{\boldsymbol{\theta}_a - \boldsymbol{\theta}}{2\eta}\right)\right)\right\} d\boldsymbol{\theta}_a = (4\eta)^d \sum_{\boldsymbol{\Theta}} \exp(U(\boldsymbol{\theta})) := (4\eta)^d Z. \tag{15}$$

This verifies that the joint posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\theta}_a)$ is mathematically well-defined[2]. Similarly, the marginal distribution for $\boldsymbol{\theta}_a$ is:

$$\begin{aligned} p(\boldsymbol{\theta}_a) &= \sum_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}, \boldsymbol{\theta}_a) \\ &\propto \sum_{\boldsymbol{\Theta}} \exp\left\{U(\boldsymbol{\theta}) - 2\ln\left(\cosh\left(\frac{\boldsymbol{\theta}_a - \boldsymbol{\theta}}{2\eta}\right)\right)\right\} \end{aligned} \tag{16}$$

□

C.2 Proof of Proposition 5.3

Proof. To show irreducibility, we consider two states $\widetilde{\boldsymbol{\theta}}_1$ and $\widetilde{\boldsymbol{\theta}}_2$, such that, the Markov chain can move from the first to the second with positive probability. In particular we focus on one full gibbs sweep of Algorithm 1 consisting of: (a) updating $\boldsymbol{\theta}_a$ given the current $\boldsymbol{\theta}$, and (b) updating $\boldsymbol{\theta}$ given the new $\boldsymbol{\theta}_a$. Because $p(\boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta}$ (by assumption), and the logistic conditional $p(\boldsymbol{\theta}_a|\boldsymbol{\theta})$ has a density that is positive everywhere on $\mathbb{R}^d$, each sub-step has full support over its variable's domain. In practical terms, no matter what values $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_{a_1})$ the chain is in, the next $\boldsymbol{\theta}_a$-update can yield any $\boldsymbol{\theta}_a$ in $\mathbb{R}^d$ with some positive probability.Further, $\boldsymbol{\theta}_a$ is arbitrarily close to the desired $\boldsymbol{\theta}_{a_2}$. In the subsequent $\boldsymbol{\theta}$-update, it can then move $\boldsymbol{\theta}$ to (or near) $\boldsymbol{\theta}_2$ because $p(\boldsymbol{\theta} \mid \boldsymbol{\theta}_a) \propto p(\boldsymbol{\theta})p(\boldsymbol{\theta}_a|\boldsymbol{\theta})$, which remains positive for $\boldsymbol{\theta} = \boldsymbol{\theta}_2$ given $p(\boldsymbol{\theta}_2) > 0$ and $p(\boldsymbol{\theta}_{a_2}|\boldsymbol{\theta}_2) > 0$. Combining these two moves: starting at $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_{a_1})$, the probability of first picking $\boldsymbol{\theta}_a$ near $\boldsymbol{\theta}_{a_2}$ and then picking $\boldsymbol{\theta}$ near $\boldsymbol{\theta}_2$ is positive. By making the neighborhood small, we see there is a non-zero probability of landing exactly in state $(\boldsymbol{\theta}_2, \boldsymbol{\theta}_{a_2})$ after one full sweep (or in a few sweeps, if we proceed gradually). Thus, $\widetilde{\boldsymbol{\theta}}_2$ is reachable with some positive probability. Equivalently, the 2-step transition kernel has a density that is everywhere positive on the state space i.e. $p(\widetilde{\boldsymbol{\theta}}) > 0$(Roberts and Tweedie, 1994; Robert and Casella, 1999). Therefore, every state can reach with every other state in some number of steps, making the chain one communicating class.

Since the joint target $p(\widetilde{\boldsymbol{\theta}})$ is a proper probability distribution (i.e. it integrates to 1 over the whole state space), it serves as an invariant (stationary) distribution for the Markov chain. An irreducible chain(shown above) that possesses a finite invariant measure is positive recurrent(Meyn and Tweedie, 2009). Positive recurrence implies that the expected return time to any state is finite, and in fact each state (or any set of states with nonzero stationary probability) will be visited infinitely often over an infinite chain trajectory(Nummelin, 1984).

All in all, the Markov chain induced by Algorithm 1 is irreducible and recurrent (in particular, positive recurrent).

□

C.3 Proof of Proposition 5.4

Proof. To establish that HiSS satisfies detailed balance with respect to the target distribution $\pi(\boldsymbol{\theta})$, we aim to show that for every pair of states $\boldsymbol{\theta}^{(i)}$ and $\boldsymbol{\theta}^{(i-1)}$:

$$\pi(\boldsymbol{\theta}^{(i-1)})\kappa_{\text{HiSS}}(\boldsymbol{\theta}^{(i)} \mid \boldsymbol{\theta}^{(i-1)}) = \pi(\boldsymbol{\theta}^{(i)})\kappa_{\text{HiSS}}(\boldsymbol{\theta}^{(i-1)} \mid \boldsymbol{\theta}^{(i)}),$$

where $\kappa_{\text{HiSS}}(\boldsymbol{\theta}^{(i)} \mid \boldsymbol{\theta}^{(i-1)})$ is the marginal transition kernel for G Gibbs sweep.

We rigorously define the marginal transition kernel for a single Gibbs sweep (denoted as κ) in Algorithm 1 in (17). In particular, we discuss the process for the i^{th} Gibbs sweep, where $i = 1, 2, 3, \cdots, G$, from $\boldsymbol{\theta}^{(i-1)}$ to $\boldsymbol{\theta}^{(i)}$:

$$\kappa(\boldsymbol{\theta}^{(i)} \mid \boldsymbol{\theta}^{(i-1)}) = \int p_{\text{MwG}}(\boldsymbol{\theta}^{(i)}_{\text{init}}, \boldsymbol{\theta}^{(i-1)}_a \mid \boldsymbol{\theta}^{(i-1)}) \cdot p_{\text{DMALA}}(\boldsymbol{\theta}^{(i)} \mid [\boldsymbol{\theta}^{(i)\,T}_{\text{init}}, \boldsymbol{\theta}^{(i-1)T}_a]^T)\, d\boldsymbol{\theta}^{(i-1)}_a \tag{17}$$

where $p_{\text{MwG}}(\boldsymbol{\theta}^{(i)}_{\text{init}}, \boldsymbol{\theta}^{(i-1)}_a \mid \boldsymbol{\theta}^{(i-1)})$ represents the local transition kernel for the acceptance of the denoised proposal(Metropolis-within-Gibbs (MwG) style), and $p_{\text{DMALA}}(\boldsymbol{\theta}^{(i)} \mid \boldsymbol{\theta}^{(i)}_{\text{init}}, \boldsymbol{\theta}^{(i-1)}_a)$ represents the L-stepped score-based refinement using DMALA. We also sometimes use this notation $\widetilde{\boldsymbol{\theta}}^{(i)}_{\text{init}}$ for intermediate state $(\boldsymbol{\theta}^{(i)}_{\text{init}}, \boldsymbol{\theta}^{(i-1)}_a)$.

[2]The exact form of the joint posterior is $p(\boldsymbol{\theta}, \boldsymbol{\theta}_a) = (4\eta)^{-d} Z^{-1} \exp\left\{U(\boldsymbol{\theta}) - 2\ln\left(\cosh\left(\frac{\boldsymbol{\theta}_a - \boldsymbol{\theta}}{2\eta}\right)\right)\right\}$.

$p_{\mathbf{MwG}}(. \mid .)$

The MwG step transitions from $\boldsymbol{\theta}^{(i-1)}$ to $\widetilde{\boldsymbol{\theta}}^{(i)}_{\text{init}}$:

$$p_{\text{MwG}}(\widetilde{\boldsymbol{\theta}}^{(i)}_{\text{init}} \mid \boldsymbol{\theta}^{(i-1)}) = q_{\text{noise}}(\boldsymbol{\theta}^{(i-1)}_a \mid \boldsymbol{\theta}^{(i-1)}) \cdot q_{\text{denoise}}(\boldsymbol{\theta}'_{\text{init}} \mid \boldsymbol{\theta}^{(i-1)}_a) \cdot a_{\text{init}}(\boldsymbol{\theta}'_{\text{init}} | \boldsymbol{\theta}^{(i-1)}) + \left(1 - L(\boldsymbol{\theta}^{(i-1)})\right) \delta(\boldsymbol{\theta}^{(i)}_{\text{init}})$$

where $p_{\text{noise-denoise}}(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \int_{\boldsymbol{\theta}_a} q_{\text{noise}}(\boldsymbol{\theta}_a|\boldsymbol{\theta}) q_{\text{denoise}}(\boldsymbol{\theta}'|\boldsymbol{\theta}_a) d\boldsymbol{\theta}_a$, $\delta(.)$ is the Kronecker delta function and $L(\boldsymbol{\theta}^{(i-1)})$ is the total acceptance probability away from the point $\boldsymbol{\theta}^{(i-1)}$ with

$$L(\boldsymbol{\theta}^{(i-1)}) = \sum_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} a_{\text{init}}(\boldsymbol{\theta}' \mid \boldsymbol{\theta}^{(i-1)}) \cdot p_{\text{noise-denoise}}(\boldsymbol{\theta}' | \boldsymbol{\theta}^{(i-1)})$$

$p_{\mathbf{DMALA}}(. \mid .)$

The overall L-stepped kernel mapping $\widetilde{\boldsymbol{\theta}}^{(i)}_{\text{init}}$ to $\boldsymbol{\theta}^{(i)}$ such that $\boldsymbol{\theta}^{(t=0)} = \boldsymbol{\theta}^{(i)}_{\text{init}}, \boldsymbol{\theta}^{(t=L)} = \boldsymbol{\theta}^{(i)}$ is given by:

$$p_{\text{DMALA}}(\boldsymbol{\theta}^{(i)} \mid \widetilde{\boldsymbol{\theta}}^{(i)}_{\text{init}}) = \prod_{t=1}^{L} \left[q_{\text{DMALA}}(\boldsymbol{\theta}^{(t)} \mid \widetilde{\boldsymbol{\theta}}^{(t-1)}) \cdot a_{\text{DMALA}}(\boldsymbol{\theta}^{(t)} | \widetilde{\boldsymbol{\theta}}^{(t-1)}) + \left(1 - L(\widetilde{\boldsymbol{\theta}}^{(t-1)})\right) \delta(\boldsymbol{\theta}^{(t)}) \right]$$

where by incorporating the coupling between the variables, we refine the DLP proposal by replacing $\nabla U(\boldsymbol{\theta})$ with $\nabla_{\boldsymbol{\theta}} U_\eta(\widetilde{\boldsymbol{\theta}})$. This adjustment results in the modified proposal from (10). During the Metropolis-Hastings (MH) step, the conditional acceptance probability is calculated per (12). $L(\widetilde{\boldsymbol{\theta}}^{(t-1)})$ is the total acceptance probability away from the point $\widetilde{\boldsymbol{\theta}}^{(t-1)}$ with auxiliary variable fixed. Thus, $L(\widetilde{\boldsymbol{\theta}}^{(t-1)}) = \sum_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} a_{\text{DMALA}}([\boldsymbol{\theta}'^T, \boldsymbol{\theta}_a^{(i-1)^T}]^T \mid [\boldsymbol{\theta}^{(t-1)^T}, \boldsymbol{\theta}_a^{(i-1)^T}]^T) \cdot q_{\text{DMALA}}([\boldsymbol{\theta}'^T, \boldsymbol{\theta}_a^{(i-1)^T}]^T \mid [\boldsymbol{\theta}^{(t-1)^T}, \boldsymbol{\theta}_a^{(i-1)^T}]^T)$.

To show detailed balance for the marginal, we proceed in two steps. First, we show that both components of the HiSS transition : the Metropolis-within-Gibbs (MwG) proposal and the DMALA refinement, individually satisfy detailed balance with respect to π. Then, we show that their composition via a delayed acceptance framework preserves detailed balance in the marginal space over $\boldsymbol{\theta}$.

Step 1: Local detailed balance.

1.1 MwG Step

$$\begin{aligned} a_{\text{init}}(\boldsymbol{\theta}'_{\text{init}} | \boldsymbol{\theta}^{(i-1)}) &= \min\left(1, \frac{\pi(\boldsymbol{\theta}'_{\text{init}}) q_{\text{noise}}(\boldsymbol{\theta}^{(i-1)}_a | \boldsymbol{\theta}'_{\text{init}}) q_{\text{denoise}}(\boldsymbol{\theta}^{(i-1)} \mid \boldsymbol{\theta}^{(i-1)}_a)}{\pi(\boldsymbol{\theta}^{(i-1)}) q_{\text{noise}}(\boldsymbol{\theta}^{(i-1)}_a | \boldsymbol{\theta}^{(i-1)}) q_{\text{denoise}}(\boldsymbol{\theta}'_{\text{init}} \mid \boldsymbol{\theta}^{(i-1)}_a)}\right) \\ &= \min\left(1, \frac{\pi(\boldsymbol{\theta}'_{\text{init}}) p_{\text{noise-denoise}}(\boldsymbol{\theta}^{(i-1)} \mid \boldsymbol{\theta}'_{\text{init}})}{\pi(\boldsymbol{\theta}^{(i-1)}) p_{\text{noise-denoise}}(\boldsymbol{\theta}'_{\text{init}} \mid \boldsymbol{\theta}^{(i-1)})}\right) \end{aligned}$$

The above acceptance probability $a_{\text{init}}(\cdot|\cdot)$ explicitly matches the detailed balance condition for the marginal distribution $\pi(\boldsymbol{\theta})$. This is evident because the acceptance ratio is constructed precisely as the product of the target distribution and the marginalized noise-denoise kernels in both forward and reverse directions. Therefore, the Metropolis-within-Gibbs step preserves detailed balance with respect to the marginal distribution $\pi(\boldsymbol{\theta})$.

1.2 Denoising Step

Each refinement step of DMALA aims to sample from the conditional distribution of $\boldsymbol{\theta}$ keeping $\boldsymbol{\theta}_a$ fixed. From Equation (12), we notice, transition kernel respects detailed balance with respect to the joint distribution $\pi(\widetilde{\boldsymbol{\theta}})$. This guarantees that marginal transitions over $\boldsymbol{\theta}$ inherit detailed balance with respect to $\pi(\boldsymbol{\theta})$.(Robert and Casella, 1999; Tierney, 1994). Consequently, the *induced* marginal proposal $q^{\text{marg}}_{\text{DMALA}}(\boldsymbol{\theta}' \mid \boldsymbol{\theta})$ satisfies:

$$a_{\text{DMALA}}(\boldsymbol{\theta}^{(t)} | \boldsymbol{\theta}^{(t-1)}) = \min\left(1, \frac{\pi(\boldsymbol{\theta}^{(t)}) q^{\text{marg}}_{\text{DMALA}}(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{\theta}^{(t)})}{\pi(\boldsymbol{\theta}^{(t-1)}) q^{\text{marg}}_{\text{DMALA}}(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)})}\right).$$

From above, we see, the DMALA kernel is Metropolis-adjusted and preserves detailed balance. Since each step in DMALA retains detailed balance, applying it L times sequentially preserves detailed balance for the same stationary distribution π for the entire Gibbs loop.

Step 2: Composition via delayed acceptance.

While each component satisfies detailed balance individually, their composition is not automatically reversible unless structured carefully. HiSS addresses this through a *delayed-acceptance* Metropolis-Hastings construction (Banterle et al., 2019).

Although this is *not* how Algorithm 1 is executed in practice, we now *interpret* the overall transition between discrete states $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ as a standard Metropolis-Hastings (MH) algorithm with an *analytically induced effective proposal* $Q(\boldsymbol{\theta} \to \boldsymbol{\theta}')$, to encapsulate the two-stage process of sampling an auxiliary variable $\boldsymbol{\theta}_a$, proposing an intermediate state $\boldsymbol{\theta}_{\text{init}}$, and refining it into $\boldsymbol{\theta}'$ via L-step DMALA, and total acceptance probability $A(\boldsymbol{\theta} \to \boldsymbol{\theta}')$. This interpretation is used purely for the purpose of proving detailed balance.

The total acceptance probability then factors as:

$$A(\boldsymbol{\theta} \to \boldsymbol{\theta}') = a_{\text{init}}(\boldsymbol{\theta}_{\text{init}} \mid \boldsymbol{\theta}) \cdot a_{\text{DMALA}}(\boldsymbol{\theta}' \mid \boldsymbol{\theta}_{\text{init}}),$$

where $a_{\text{DMALA}}(\boldsymbol{\theta}' \mid \boldsymbol{\theta}_{\text{init}}) = \prod_{t=1}^{L} a_{\text{DMALA}}(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)})$.

In essence, we argue HiSS is a simple MH algorithm in disguise. The following decomposition solely serves as a condensed, conceptual MH abstraction to analyze the marginal chain:

$$\kappa(\boldsymbol{\theta}' \mid \boldsymbol{\theta}) = Q(\boldsymbol{\theta} \to \boldsymbol{\theta}')A(\boldsymbol{\theta} \to \boldsymbol{\theta}') + \left[1 - \sum_{u} Q(\boldsymbol{\theta} \to u)A(\boldsymbol{\theta} \to u)\right] \delta(\boldsymbol{\theta}' - \boldsymbol{\theta}).$$

By construction, the acceptance term is,

$$A(\boldsymbol{\theta} \to \boldsymbol{\theta}') = \min\left(1, \frac{\pi(\boldsymbol{\theta}')Q(\boldsymbol{\theta}' \to \boldsymbol{\theta})}{\pi(\boldsymbol{\theta})Q(\boldsymbol{\theta} \to \boldsymbol{\theta}')}\right)$$

This factorization corresponds to the delayed-acceptance MH framework described in (Banterle et al., 2019, Section 2.2, Lemma 2), where the acceptance is decomposed across multiple refinement steps.

Therefore, even though detailed balance does not necessarily hold jointly for the auxiliary-augmented chain, the marginal transition kernel $\kappa(\boldsymbol{\theta}' \mid \boldsymbol{\theta})$ satisfies detailed balance with respect to the target distribution $\pi(\boldsymbol{\theta})$:

$$\pi(\boldsymbol{\theta})\kappa(\boldsymbol{\theta}' \mid \boldsymbol{\theta}) = \pi(\boldsymbol{\theta}')\kappa(\boldsymbol{\theta} \mid \boldsymbol{\theta}').$$

By induction, applying G such Gibbs sweeps—each satisfying detailed balance—yields an overall kernel κ_{HiSS} that also satisfies detailed balance with respect to π:

$$\pi(\boldsymbol{\theta}^{(i-1)})\kappa_{\text{HiSS}}(\boldsymbol{\theta}^{(i)} \mid \boldsymbol{\theta}^{(i-1)}) = \pi(\boldsymbol{\theta}^{(i)})\kappa_{\text{HiSS}}(\boldsymbol{\theta}^{(i-1)} \mid \boldsymbol{\theta}^{(i)}).$$

Thus, HiSS satisfies detailed balance with respect to the marginal distribution $\pi(\boldsymbol{\theta})$, ensuring that the Markov chain converges to the desired discrete target distribution.

□

Lemma C.1. *Under Assumptions 5.1 and 5.2, and for step size $\alpha < \frac{2}{M}$ in Algorithm 1, for the i^{th} Gibbs Sweep, the L-step DMALA refinement kernel admits a uniform lower bound independent of the auxiliary variable $\boldsymbol{\theta}_a$. Specifically, for any starting state $\boldsymbol{\theta}_{init}^{(i)} \in \boldsymbol{\Theta}$ and any refined final state $\boldsymbol{\theta}^{(i)} \in \boldsymbol{\Theta}$,*

$$p_{DMALA}(\boldsymbol{\theta}^{(i)} \mid \widetilde{\boldsymbol{\theta}}_{init}^{(i)}) \geq \nu_i(A) \exp\left\{L\left((-\frac{M}{2} - \frac{1}{\alpha} + \frac{m}{4})diam^2(\boldsymbol{\Theta}) - (\frac{1}{2}\|\nabla U(a)\| + \frac{3\sqrt{d}+1}{\eta})diam(\boldsymbol{\Theta})\right)\right\}$$

where $\widetilde{\boldsymbol{\theta}}_{\text{init}}^{(i)} = [\boldsymbol{\theta}_{\text{init}}^{(i)}, \boldsymbol{\theta}_a^{(i-1)}]^T$ is the starting joint state and $A \subseteq \boldsymbol{\Theta}^L$ is a measurable set of length-L trajectories, with $\nu_i(A)$ denoting a probability measure over those discrete trajectories.

C.4 Proof of Lemma C.1

Proof. We follow a similar minorization proof style as of Lemma 5.3 from Pynadath et al. (2024). The DMALA step transitions from $\widetilde{\boldsymbol{\theta}}'_{\text{init}}$ to $\boldsymbol{\theta}^{(i)}$ conditionally over L iterative steps, for some arbitrary Gibbs-Sweep, each incorporating DLP with α as the step size for the t^{th} refinement.

$$
\begin{aligned}
q_{\text{DMALA}}(\boldsymbol{\theta}^{(t)} \mid \widetilde{\boldsymbol{\theta}}^{(t-1)}) &\propto \exp\left(\frac{1}{2}\nabla_{\boldsymbol{\theta}} U_\eta(\widetilde{\boldsymbol{\theta}}^{(t-1)})^\top(\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}) - \frac{1}{2\alpha}\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}\|^2\right) \\
&= \exp\left(\frac{1}{2}\nabla U(\boldsymbol{\theta}^{(t-1)})^\top(\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}) + \frac{1}{2\eta}\tanh\left(\frac{\boldsymbol{\theta}_a^{(t-1)} - \boldsymbol{\theta}^{(t-1)}}{2\eta}\right)^\top(\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)})\right) \\
&= \exp\left(\frac{1}{2}(-U(\boldsymbol{\theta}^{(t-1)}) + U(\boldsymbol{\theta}^{(t)})) - (\boldsymbol{\theta}^{(t-1)} - \boldsymbol{\theta}^{(t)})^\top\left(\frac{1}{2\alpha}I + \frac{1}{4}\int_0^1 \nabla^2 U((1-s)\boldsymbol{\theta}^{(t-1)} + s\boldsymbol{\theta}^{(t)})ds\right)\cdot\right. \\
&\left.(\boldsymbol{\theta}^{(t-1)} - \boldsymbol{\theta}^{(t)}) + \frac{1}{2\eta}\tanh\left(\frac{\boldsymbol{\theta}_a^{(t-1)} - \boldsymbol{\theta}^{(t-1)}}{2\eta}\right)^\top(\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)})\right)
\end{aligned}
$$

The third line is true because we replace the linear gradient term $\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})^\top(\boldsymbol{\theta}' - \boldsymbol{\theta})$ with Taylor expansion with integral remainder.

Consequently, the modified normalizing constant becomes

$$
\begin{aligned}
Z(\widetilde{\boldsymbol{\theta}}^{(t-1)}) = \sum_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \exp\Bigg(&\frac{1}{2}(-U(\boldsymbol{\theta}^{(t-1)}) + U(\boldsymbol{\theta}')) - (\boldsymbol{\theta}^{(t-1)} - \boldsymbol{\theta}')^\top\left(\frac{1}{2\alpha}I + \frac{1}{4}\int_0^1 \nabla^2 U((1-s)\boldsymbol{\theta}^{(t-1)} + s\boldsymbol{\theta}')ds\right)(\boldsymbol{\theta}^{(t-1)} - \boldsymbol{\theta}') \\
&+ \frac{1}{2\eta}\tanh\left(\frac{\boldsymbol{\theta}_a^{(t-1)} - \boldsymbol{\theta}^{(t-1)}}{2\eta}\right)^\top(\boldsymbol{\theta}' - \boldsymbol{\theta}^{(t-1)})\Bigg).
\end{aligned}
$$

Recall, from Assumption 5.1,U is M-gradient Lipschitz, we have

$$
\frac{1}{\alpha}I + \frac{1}{2}\int_0^1 \nabla^2 U((1-s)\boldsymbol{\theta} + s\boldsymbol{\theta}')\,ds) \geq \left(\frac{1}{\alpha} - \frac{M}{2}\right)I
$$

Since $\alpha < \frac{2}{M}$, the matrix $\left(\frac{1}{2\alpha} - \frac{M}{2}\right)I$is positive definite.

This implies,

$$
\begin{aligned}
Z(\widetilde{\boldsymbol{\theta}}^{(t-1)}) &\leq \sum_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \exp\left(\frac{1}{2}(-U(\boldsymbol{\theta}^{(t-1)}) + U(\boldsymbol{\theta}')) + \frac{1}{2\eta}\tanh\left(\frac{\boldsymbol{\theta}_a^{(t-1)} - \boldsymbol{\theta}^{(t-1)}}{2\eta}\right)^\top(\boldsymbol{\theta}' - \boldsymbol{\theta}^{(t-1)})\right) \\
&\leq \exp\left(\frac{-U(\boldsymbol{\theta}^{(t-1)})}{2}\right)\sum_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \exp\left(\frac{1}{2}(U(\boldsymbol{\theta}')) + \frac{1}{2\eta}||\tanh\left(\frac{\boldsymbol{\theta}_a^{(t-1)} - \boldsymbol{\theta}^{(t-1)}}{2\eta}\right)||\cdot||(\boldsymbol{\theta}' - \boldsymbol{\theta}^{(t-1)})||\right) \\
&\leq \exp\left(\frac{-U(\boldsymbol{\theta}^{(t-1)})}{2} + \frac{\sqrt{d}}{2\eta}\text{diam}(\boldsymbol{\Theta})\right)\sum_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \exp\left(\frac{1}{2}(U(\boldsymbol{\theta}'))\right)
\end{aligned}
$$

Since Assumption 5.2 holds true in this setting, we have an $m > 0$ such that for any $\boldsymbol{\theta} \in conv(\boldsymbol{\Theta})$

$$
-\nabla^2 U(\boldsymbol{\theta}) \geq m\,I.
$$

From this, one notes that,

$$\begin{aligned}
Z(\widetilde{\boldsymbol{\theta}}^{(t-1)}) &\geq \sum_{\boldsymbol{\theta'}\in\boldsymbol{\Theta}} \exp\left(\frac{1}{2}(-U(\boldsymbol{\theta}^{(t-1)}) + U(\boldsymbol{\theta'})) + \frac{1}{2\eta}\tanh\left(\frac{\boldsymbol{\theta}_a^{(t-1)} - \boldsymbol{\theta}^{(t-1)}}{2\eta}\right)^\top (\boldsymbol{\theta'} - \boldsymbol{\theta}^{(t-1)})\right)\cdot \\
&\quad \exp\left\{-\frac{1}{2}\left(\frac{1}{\alpha} - \frac{m}{2}\right)\,\mathrm{diam}(\boldsymbol{\Theta})^2\right\} \\
&\geq \exp\left\{-\frac{1}{2}\left(\frac{1}{\alpha} - \frac{m}{2}\right)\,\mathrm{diam}^2(\boldsymbol{\Theta}) + \frac{-U(\boldsymbol{\theta}^{(t-1)})}{2}\right\}\cdot \\
&\quad \sum_{\boldsymbol{\theta'}\in\boldsymbol{\Theta}} \exp\left(\frac{1}{2}(U(\boldsymbol{\theta'})) - \frac{1}{2\eta}||\tanh\left(\frac{\boldsymbol{\theta}_a^{(t-1)} - \boldsymbol{\theta}^{(t-1)}}{2\eta}\right)||\cdot||(\boldsymbol{\theta'} - \boldsymbol{\theta}^{(t-1)})||\right) \\
&\geq \exp\left(\frac{-U(\boldsymbol{\theta}^{(t-1)})}{2} - \frac{\sqrt{d}}{2\eta}\mathrm{diam}(\boldsymbol{\Theta}) - \frac{1}{2}\left(\frac{1}{\alpha} - \frac{m}{2}\right)\,\mathrm{diam}^2(\boldsymbol{\Theta})\right)\sum_{\boldsymbol{\theta'}\in\boldsymbol{\Theta}}\exp\left(\frac{U(\boldsymbol{\theta'})}{2}\right)
\end{aligned}$$

In other words,

$$\exp\left(-\frac{\sqrt{d}\mathrm{diam}(\boldsymbol{\Theta})}{2\eta} - \frac{1}{2}\left(\frac{1}{\alpha} - \frac{m}{2}\right)\,\mathrm{diam}^2(\boldsymbol{\Theta})\right) \leq \frac{Z(\widetilde{\boldsymbol{\theta}}^{(t-1)})}{\exp(\frac{-U(\boldsymbol{\theta}^{(t-1)})}{2})\sum_{\boldsymbol{\theta'}\in\boldsymbol{\Theta}}\exp\left(\frac{U(\boldsymbol{\theta'})}{2}\right)} \leq \exp\left(\frac{\sqrt{d}\mathrm{diam}(\boldsymbol{\Theta})}{2\eta}\right)$$

Consequently,

$$\frac{\frac{Z(\widetilde{\boldsymbol{\theta}}^{(t-1)})}{\sum_{x\in\boldsymbol{\Theta}}\exp\left(\frac{U(x)}{2}\right)\exp\left(-\frac{U(\boldsymbol{\theta}^{(t-1)})}{2}\right)}}{\frac{Z(\widetilde{\boldsymbol{\theta}}^{(t)})}{\sum_{x\in\boldsymbol{\Theta}}\exp\left(\frac{U(x)}{2}\right)\exp\left(-\frac{U(\boldsymbol{\theta})^{(t)}}{2}\right)}} \geq \frac{\exp\left(-\frac{\sqrt{d}\mathrm{diam}(\boldsymbol{\Theta})}{2\eta} - \frac{(2-m\alpha)\mathrm{diam}^2(\boldsymbol{\Theta})}{4\alpha}\right)}{\exp\left(\frac{\sqrt{d}\mathrm{diam}(\boldsymbol{\Theta})}{2\eta}\right)}$$

This implies, for any two states $\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}'$,

$$\frac{Z(\widetilde{\boldsymbol{\theta}})}{Z(\widetilde{\boldsymbol{\theta}}')} \geq \exp\left(\frac{1}{2}(-U(\boldsymbol{\theta}) + U(\boldsymbol{\theta}'))\right)\exp\left(-\frac{\sqrt{d}\mathrm{diam}(\boldsymbol{\Theta})}{\eta} - \frac{(2-m\alpha)\mathrm{diam}^2(\boldsymbol{\Theta})}{4\alpha}\right)$$

Thus,

$$\begin{aligned}
q_{\mathrm{DMALA}}(\widetilde{\boldsymbol{\theta}}^{(t)} \mid \widetilde{\boldsymbol{\theta}}^{(t-1)}) &\geq \frac{1}{Z(\widetilde{\boldsymbol{\theta}}^{(t-1)})}\exp\left(\frac{1}{2}\left\langle\nabla U(\boldsymbol{\theta}^{(t-1)}), \boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}\right\rangle + \frac{1}{2\eta}\tanh\left(\frac{\boldsymbol{\theta}_a^{(t-1)} - \boldsymbol{\theta}^{(t-1)}}{2\eta}\right)^\top(\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)})\right) \\
&\geq \frac{1}{Z(\widetilde{\boldsymbol{\theta}}^{(t-1)})}\exp\left(\frac{1}{2}\left\langle\nabla U(\boldsymbol{\theta}^{(t-1)}), \boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}\right\rangle - \frac{1}{2\eta}||\tanh\left(\frac{\boldsymbol{\theta}_a^{(t-1)} - \boldsymbol{\theta}^{(t-1)}}{2\eta}\right)||\cdot||\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}||\right) \\
&\geq \frac{1}{Z(\widetilde{\boldsymbol{\theta}}^{(t-1)})}\exp\left(\frac{1}{2}\left\langle\nabla U(\boldsymbol{\theta}^{(t-1)}), \boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}\right\rangle - \frac{\sqrt{d}\mathrm{diam}(\boldsymbol{\Theta})}{2\eta}\right)
\end{aligned}$$

We also note that

$$\begin{aligned}
-\frac{1}{2}\left\langle\nabla U(\boldsymbol{\theta}), \boldsymbol{\theta}' - \boldsymbol{\theta}\right\rangle + \frac{1}{2\alpha}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2 &= \frac{1}{2}\left\langle -\nabla U(\boldsymbol{\theta}) + \nabla U(a), \boldsymbol{\theta}' - \boldsymbol{\theta}\right\rangle + \frac{1}{2}\left\langle -\nabla U(a), \boldsymbol{\theta}' - \boldsymbol{\theta}\right\rangle + \frac{1}{2\alpha}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2 \\
&\leq \frac{1}{2}\left\langle -\nabla U(\boldsymbol{\theta}) + \nabla U(a), \boldsymbol{\theta}' - \boldsymbol{\theta}\right\rangle + \frac{1}{2}\left\langle -\nabla U(a), \boldsymbol{\theta}' - \boldsymbol{\theta}\right\rangle + \frac{1}{2\alpha}diam(\boldsymbol{\Theta})^2 \\
&\leq \frac{1}{2}\left\|-\nabla U(\boldsymbol{\theta}) + \nabla U(a)\right\|\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| + \frac{1}{2}\left\|\nabla U(a)\right\|\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| + \frac{1}{2\alpha}diam(\boldsymbol{\Theta})^2 \\
&\leq \frac{1}{2}\|-\nabla U(\boldsymbol{\theta}) + \nabla U(a)\|diam(\boldsymbol{\Theta}) + \frac{1}{2}\|\nabla U(a)\|diam(\boldsymbol{\Theta}) + \frac{1}{2\alpha}diam(\boldsymbol{\Theta})^2 \\
&\leq \left(\frac{1}{2}M + \frac{1}{2\alpha}\right)\,diam(\boldsymbol{\Theta})^2 + \frac{1}{2}\|\nabla U(a)\|\,diam(\boldsymbol{\Theta}).
\end{aligned}$$

Therefore,

$$
\begin{aligned}
q_{\mathrm{DMALA}}(\boldsymbol{\theta}^{(t)} \mid \widetilde{\boldsymbol{\theta}}^{(t-1)}) &\geq \frac{1}{Z(\widetilde{\boldsymbol{\theta}}^{(t-1)})} \exp\left((-\frac{M}{2} - \frac{1}{2\alpha})\mathrm{diam}^2(\boldsymbol{\Theta}) - \frac{1}{2}\|\nabla U(a)\|\mathrm{diam}(\boldsymbol{\Theta}) - \frac{\sqrt{d}\mathrm{diam}(\boldsymbol{\Theta})}{2\eta}\right) \\
&\geq \frac{\exp\left((-\frac{M}{2} - \frac{1}{2\alpha})\mathrm{diam}^2(\boldsymbol{\Theta}) - \frac{1}{2}\|\nabla U(a)\|\mathrm{diam}(\boldsymbol{\Theta}) - \frac{\sqrt{d}\mathrm{diam}(\boldsymbol{\Theta})}{2\eta}\right)}{\exp\left(\frac{-U(\boldsymbol{\theta}^{(t-1)})}{2} + \frac{\sqrt{d}\mathrm{diam}(\boldsymbol{\Theta})}{2\eta}\right) \sum_{\boldsymbol{\theta'} \in \boldsymbol{\Theta}} \exp\left(\frac{1}{2}(U(\boldsymbol{\theta'}))\right)}
\end{aligned}
$$

The acceptance ratio is given as:

$$
\begin{aligned}
\rho_{\mathrm{DMALA}}(\boldsymbol{\theta}^{(t)} \mid \widetilde{\boldsymbol{\theta}}^{(t-1)}) &= \frac{\pi(\widetilde{\boldsymbol{\theta}}^{(t)}) q_{\mathrm{DMALA}}(\boldsymbol{\theta}^{(t-1)} \mid \widetilde{\boldsymbol{\theta}}^{(t)})}{\pi(\widetilde{\boldsymbol{\theta}}^{(t-1)}) q_{\mathrm{DMALA}}(\boldsymbol{\theta}^{(t)} \mid \widetilde{\boldsymbol{\theta}}^{(t-1)})} \\
&= \exp\left\{ U(\boldsymbol{\theta}^{(t)}) - U(\boldsymbol{\theta}^{(t-1)}) - 2\ln\left(\cosh\left(\frac{\boldsymbol{\theta}_a^{(t-1)} - \boldsymbol{\theta}^{(t)}}{2\eta}\right)\right) \right. \\
&\quad \left. + 2\ln\left(\cosh\left(\frac{\boldsymbol{\theta}_a^{(t-1)} - \boldsymbol{\theta}^{(t-1)}}{2\eta}\right)\right)\right\} \cdot \frac{\tilde{Z}}{\tilde{Z}} \cdot \\
&\quad \exp\left\{ U(\boldsymbol{\theta}^{(t-1)}) - U(\boldsymbol{\theta}^{(t)}) + \frac{1}{2\eta}\tanh\left(\frac{\boldsymbol{\theta}_a^{(t-1)} - \boldsymbol{\theta}^{(t)}}{2\eta}\right)^\top (\boldsymbol{\theta}^{(t-1)} - \boldsymbol{\theta}^{(t)}) \right. \\
&\quad \left. - \frac{1}{2\eta}\tanh\left(\frac{\boldsymbol{\theta}_a^{(t-1)} - \boldsymbol{\theta}^{(t-1)}}{2\eta}\right)^\top (\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}) \right\} \cdot \frac{Z(\widetilde{\boldsymbol{\theta}}^{(t-1)})}{Z(\widetilde{\boldsymbol{\theta}}^{(t)})} \\
&= \exp\left\{ -2\ln\left(\frac{\cosh\left(\frac{\boldsymbol{\theta}_a^{(t-1)} - \boldsymbol{\theta}^{(t)}}{2\eta}\right)}{\cosh\left(\frac{\boldsymbol{\theta}_a^{(t-1)} - \boldsymbol{\theta}^{(t-1)}}{2\eta}\right)}\right) - \frac{1}{2\eta}(\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)})^\top \right. \\
&\quad \left. (\tanh\left(\frac{\boldsymbol{\theta}_a^{(t-1)} - \boldsymbol{\theta}^{(t)}}{2\eta}\right) + \tanh\left(\frac{\boldsymbol{\theta}_a^{(t-1)} - \boldsymbol{\theta}^{(t-1)}}{2\eta}\right)) \right\} \cdot \frac{Z(\widetilde{\boldsymbol{\theta}}^{(t-1)})}{Z(\widetilde{\boldsymbol{\theta}}^{(t)})} \\
&\geq \exp\left\{ -\frac{1}{\eta}||\boldsymbol{\theta}^{(t-1)} - \boldsymbol{\theta}^{(t)}|| - \frac{1}{2\eta}(\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)})^\top \right. \\
&\quad \left. (\tanh\left(\frac{\boldsymbol{\theta}_a^{(t-1)} - \boldsymbol{\theta}^{(t)}}{2\eta}\right) + \tanh\left(\frac{\boldsymbol{\theta}_a^{(t-1)} - \boldsymbol{\theta}^{(t-1)}}{2\eta}\right)) \right\} \cdot \frac{Z(\widetilde{\boldsymbol{\theta}}^{(t-1)})}{Z(\widetilde{\boldsymbol{\theta}}^{(t)})} \\
&\geq \exp\left\{ -\frac{1}{\eta}||\boldsymbol{\theta}^{(t-1)} - \boldsymbol{\theta}^{(t)}|| - \frac{\sqrt{d}}{\eta}||\boldsymbol{\theta}^{(t-1)} - \boldsymbol{\theta}^{(t)}|| \right\} \cdot \frac{Z(\widetilde{\boldsymbol{\theta}}^{(t-1)})}{Z(\widetilde{\boldsymbol{\theta}}^{(t)})} \\
&\geq \exp\left\{ -\frac{(\sqrt{d}+1)}{\eta}\mathrm{diam}(\boldsymbol{\Theta}) \right\} \cdot \frac{Z(\widetilde{\boldsymbol{\theta}}^{(t-1)})}{Z(\widetilde{\boldsymbol{\theta}}^{(t)})}
\end{aligned}
$$

where $\tilde{Z}$ is the normalizing constant for $\pi(\widetilde{\boldsymbol{\theta}})$.

with Acceptance Probability

$$
\mathcal{A}_{\mathrm{DMALA}}(\boldsymbol{\theta}^{(t)} \mid \widetilde{\boldsymbol{\theta}}^{(t-1)}) = \left(\rho_{\mathrm{DMALA}}(\boldsymbol{\theta}^{(t)} \mid \widetilde{\boldsymbol{\theta}}^{(t-1)}) \wedge 1\right)
$$

The overall L-stepped kernel is then given by: $\boldsymbol{\theta}^{(t=0)} = \boldsymbol{\theta}^{(i)}_{\text{init}}, \boldsymbol{\theta}^{(t=L)} = \boldsymbol{\theta}^{(i)}$

$$
\begin{aligned}
p_{\text{DMALA}}(\boldsymbol{\theta}^{(i)} \mid \widetilde{\boldsymbol{\theta}}^{(i)}_{\text{init}}) &= \prod_{t=1}^{L} \left[q_{\text{DMALA}}(\boldsymbol{\theta}^{(t)} \mid \widetilde{\boldsymbol{\theta}}^{(t-1)}) \cdot \mathcal{A}_{\text{DMALA}}(\boldsymbol{\theta}^{(t)} | \widetilde{\boldsymbol{\theta}}^{(t-1)}) + \left(1 - L(\widetilde{\boldsymbol{\theta}}^{(t-1)})\right) \delta_{\boldsymbol{\theta}}(\boldsymbol{\theta}^{(t)}) \right] \\
&\geq \prod_{t=1}^{L} \left[q_{\text{DMALA}}(\boldsymbol{\theta}^{(t)} \mid \widetilde{\boldsymbol{\theta}}^{(t-1)}) \cdot \mathcal{A}_{\text{DMALA}}(\boldsymbol{\theta}^{(t)} | \widetilde{\boldsymbol{\theta}}^{(t-1)}) \right] \\
&\geq \prod_{t=1}^{L} \left[q_{\text{DMALA}}(\boldsymbol{\theta}^{(t)} \mid \widetilde{\boldsymbol{\theta}}^{(t-1)}) \cdot \exp\left\{ -\frac{(\sqrt{d}+1)}{\eta} \text{diam}(\boldsymbol{\Theta}) \right\} \cdot \frac{Z(\widetilde{\boldsymbol{\theta}}^{(t-1)})}{Z(\widetilde{\boldsymbol{\theta}}^{(t)})} \right] \\
&\geq \prod_{t=1}^{L} \left[q_{\text{DMALA}}(\boldsymbol{\theta}^{(t)} \mid \widetilde{\boldsymbol{\theta}}^{(t-1)}) \cdot \exp\left(-\frac{(\sqrt{d}+1)}{\eta} \text{diam}(\boldsymbol{\Theta}) \right) \cdot \exp\left(\frac{1}{2} \left[-U(\boldsymbol{\theta}^{(t-1)}) + U(\boldsymbol{\theta}^{(t)}) \right] \right) \cdot \right. \\
&\quad \left. \exp\left(-\frac{\sqrt{d}\,\text{diam}(\boldsymbol{\Theta})}{\eta} - \frac{(2-m\alpha)\,\text{diam}^2(\boldsymbol{\Theta})}{4\alpha} \right) \right] \\
&= \prod_{t=1}^{L} \left[q_{\text{DMALA}}(\boldsymbol{\theta}^{(t)} \mid \widetilde{\boldsymbol{\theta}}^{(t-1)}) \cdot \exp\left(-\frac{(2\sqrt{d}+1)}{\eta} \text{diam}(\boldsymbol{\Theta}) + \frac{1}{2} \left(-U(\boldsymbol{\theta}^{(t-1)}) + U(\boldsymbol{\theta}^{(t)}) \right) \right. \right. \\
&\quad \left. \left. - \frac{(2-m\alpha)\,\text{diam}^2(\boldsymbol{\Theta})}{4\alpha} \right) \right] \\
&\geq \prod_{t=1}^{L} \left[\frac{\exp\left(-\left(\frac{M}{2} + \frac{1}{2\alpha}\right) \text{diam}^2(\boldsymbol{\Theta}) - \frac{1}{2}\|\nabla U(a)\| \cdot \text{diam}(\boldsymbol{\Theta}) - \frac{\sqrt{d} \cdot \text{diam}(\boldsymbol{\Theta})}{2\eta} \right)}{\exp\left(-\frac{1}{2} U(\boldsymbol{\theta}^{(t-1)}) + \frac{\sqrt{d} \cdot \text{diam}(\boldsymbol{\Theta})}{2\eta} \right) \sum_{\boldsymbol{\theta'} \in \boldsymbol{\Theta}} \exp\left(\frac{1}{2} U(\boldsymbol{\theta'}) \right)} \right. \\
&\quad \left. \cdot \exp\left(-\frac{(2\sqrt{d}+1)}{\eta} \cdot \text{diam}(\boldsymbol{\Theta}) + \frac{1}{2} \left(U(\boldsymbol{\theta}^{(t)}) - U(\boldsymbol{\theta}^{(t-1)}) \right) - \frac{(2-m\alpha)}{4\alpha} \cdot \text{diam}^2(\boldsymbol{\Theta}) \right) \right] \\
&= \prod_{t=1}^{L} \left[\frac{\exp\left(-\left(\frac{M}{2} + \frac{1}{2\alpha}\right) \text{diam}^2(\boldsymbol{\Theta}) - \frac{1}{2}\|\nabla U(a)\| \cdot \text{diam}(\boldsymbol{\Theta}) - \frac{\sqrt{d} \cdot \text{diam}(\boldsymbol{\Theta})}{\eta} \right)}{\sum_{\boldsymbol{\theta'} \in \boldsymbol{\Theta}} \exp\left(\frac{1}{2} U(\boldsymbol{\theta'}) \right)} \right. \\
&\quad \left. \cdot \exp\left(-\frac{(2\sqrt{d}+1)}{\eta} \cdot \text{diam}(\boldsymbol{\Theta}) + \frac{1}{2} U(\boldsymbol{\theta}^{(t)}) - \frac{(2-m\alpha)}{4\alpha} \cdot \text{diam}^2(\boldsymbol{\Theta}) \right) \right] \\
&= \frac{\exp(\sum_{t=1}^{L} \frac{U(\boldsymbol{\theta}^{(t-1)})}{2})}{\left(\sum_{\boldsymbol{\theta'} \in \boldsymbol{\Theta}} \exp\left(\frac{U(\boldsymbol{\theta'})}{2} \right) \right)^L} \cdot \\
&\exp\left\{ L \left((-\frac{M}{2} - \frac{1}{\alpha} + \frac{m}{4}) \text{diam}^2(\boldsymbol{\Theta}) - (\frac{1}{2}\|\nabla U(a)\| + \frac{3\sqrt{d}+1}{\eta}) \text{diam}(\boldsymbol{\Theta}) \right) \right\} \\
&= \nu_i(A) \exp\left\{ L \left((-\frac{M}{2} - \frac{1}{\alpha} + \frac{m}{4}) \text{diam}^2(\boldsymbol{\Theta}) - (\frac{1}{2}\|\nabla U(a)\| + \frac{3\sqrt{d}+1}{\eta}) \text{diam}(\boldsymbol{\Theta}) \right) \right\}
\end{aligned}
$$

Thus, we were able to lower bound the L-stepped score-based-denoising kernel without any dependence on $\boldsymbol{\theta}_a$.

□

C.5 Proof of Theorem 5.5

Proof. Our kernel of interest is the global marginal transition kernel after G gibbs sweeps i.e.

$$
\begin{aligned}
\kappa_{(\text{HiSS})}(\boldsymbol{\theta}^{(G)} \mid \boldsymbol{\theta}^{(0)}) &= \prod_{i=1}^{G} \kappa(\boldsymbol{\theta}^{(i)} \mid \boldsymbol{\theta}^{(i-1)}) \\
&= \prod_{i=1}^{G} \int p_{\text{MwG}}(\boldsymbol{\theta}^{(i)}_{\text{init}}, \boldsymbol{\theta}^{(i-1)}_{a} \mid \boldsymbol{\theta}^{(i-1)}) \cdot p_{\text{DMALA}}(\boldsymbol{\theta}^{(i)} \mid [{\boldsymbol{\theta}^{(i)}_{\text{init}}}^{T}, {\boldsymbol{\theta}^{(i-1)}_{a}}^{T}]^T)\, d\boldsymbol{\theta}^{(i-1)}_{a}
\end{aligned}
$$

In order to lower bound $\kappa_{(\text{HiSS})}(\boldsymbol{\theta}^{(G)} \mid \boldsymbol{\theta}^{(0)})$, we derive a lower bound for $\int p_{\text{MwG}}(\boldsymbol{\theta}^{(i)}_{\text{init}}, \boldsymbol{\theta}^{(i-1)}_{a} \mid \boldsymbol{\theta}^{(i-1)}) d\boldsymbol{\theta}^{(i-1)}_{a}$ and use Lemma (C.1) to lower bound $p_{\text{DMALA}}(\boldsymbol{\theta}^{(i)} \mid [{\boldsymbol{\theta}^{(i)}_{\text{init}}}^{T}, {\boldsymbol{\theta}^{(i-1)}_{a}}^{T}]^T)$ independent of the auxiliary variable.
By virtue of coordinatewise noising, we can say,

$$
q_{\text{noise}}(\boldsymbol{\theta}_a \mid \boldsymbol{\theta}) \propto \frac{1}{(4\eta)^d} \prod_{i=1}^{d} \text{sech}^2\left(\frac{(\theta_a)_i - \theta_i}{2\eta} \right) \propto \frac{1}{(4\eta)^d} \text{sech}^2\left(\frac{\boldsymbol{\theta}_a - \boldsymbol{\theta}}{2\eta} \right)
$$

Similarly, for denoising,

$$q_{\text{denoise}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_a) \propto \prod_{i=1}^{d} \exp\left(-2\ln\left(\frac{(\theta_a)_i - \theta_i}{2\eta}\right)\right) \propto \operatorname{sech}^2\left(\frac{\boldsymbol{\theta}_a - \boldsymbol{\theta}}{2\eta}\right)$$

We know by definition,

$$\begin{aligned}
p_{\text{noise-denoise}}(\boldsymbol{\theta}'|\boldsymbol{\theta}) &\propto \int_{\boldsymbol{\theta}_a} q_{\text{noise}}(\boldsymbol{\theta}_a|\boldsymbol{\theta}) q_{\text{denoise}}(\boldsymbol{\theta}'|\boldsymbol{\theta}_a) d\boldsymbol{\theta}_a \\
&= \int_{\boldsymbol{\theta}_a} \frac{1}{(4\eta)^d} \operatorname{sech}^2(\frac{\boldsymbol{\theta} - \boldsymbol{\theta}_a}{2\eta}) \operatorname{sech}^2(\frac{\boldsymbol{\theta}_a - \boldsymbol{\theta}'}{2\eta}) d\boldsymbol{\theta}_a \\
&\geq \frac{1}{(4\eta)^d} \int_{\boldsymbol{\theta}_a} e^{-\frac{|\boldsymbol{\theta} - \boldsymbol{\theta}_a|}{\eta}} \cdot e^{-\frac{|\boldsymbol{\theta}_a - \boldsymbol{\theta}'|}{\eta}} d\boldsymbol{\theta}_a \\
&= \frac{1}{(4\eta)^d} \int_{\boldsymbol{\theta}_a} e^{-\frac{|\boldsymbol{\theta} - \boldsymbol{\theta}'|}{\eta}} \cdot e^{-2\frac{|\boldsymbol{\theta}_a - \frac{\boldsymbol{\theta}' + \boldsymbol{\theta}}{2}|}{\eta}} d\boldsymbol{\theta}_a \\
&= \frac{1}{(4\eta)^d} e^{-\frac{|\boldsymbol{\theta} - \boldsymbol{\theta}'|}{\eta}} \int_{\boldsymbol{\theta}_a \in R^d} e^{-2\frac{|\boldsymbol{\theta}_a - \frac{\boldsymbol{\theta}' + \boldsymbol{\theta}}{2}|}{\eta}} d\boldsymbol{\theta}_a \\
&= \frac{1}{(4\eta)^d} e^{-\frac{|\boldsymbol{\theta} - \boldsymbol{\theta}'|}{\eta}} (\eta)^d \\
&= \frac{1}{4^d} e^{-\frac{|\boldsymbol{\theta} - \boldsymbol{\theta}'|}{\eta}}
\end{aligned}$$

$$\begin{aligned}
Z_{\text{noise-denoise}}(\boldsymbol{\theta}) &= \sum_{x \in \boldsymbol{\Theta}} p_{\text{noise-denoise}}(x|\boldsymbol{\theta}) \\
&\geq \sum_{x \in \boldsymbol{\Theta}} \frac{1}{4^d} e^{-\frac{|\boldsymbol{\theta} - x|}{\eta}} \\
&\geq \frac{|\boldsymbol{\Theta}|}{4^d} e^{-\frac{\Delta(\boldsymbol{\Theta})}{\eta}}
\end{aligned}$$

Similarly,

$$\begin{aligned}
p_{\text{noise-denoise}}(\boldsymbol{\theta}'|\boldsymbol{\theta}) &\propto \int_{\boldsymbol{\theta}_a} q_{\text{noise}}(\boldsymbol{\theta}_a|\boldsymbol{\theta}) q_{\text{denoise}}(\boldsymbol{\theta}'|\boldsymbol{\theta}_a) d\boldsymbol{\theta}_a \\
&= \int_{\boldsymbol{\theta}_a} \frac{1}{(4\eta)^d} \operatorname{sech}^2(\frac{\boldsymbol{\theta} - \boldsymbol{\theta}_a}{2\eta}) \operatorname{sech}^2(\frac{\boldsymbol{\theta}_a - \boldsymbol{\theta}'}{2\eta}) d\boldsymbol{\theta}_a \\
&\leq \frac{1}{(4\eta)^d} \int_{\boldsymbol{\theta}_a} 4e^{-\frac{|\boldsymbol{\theta} - \boldsymbol{\theta}_a|}{\eta}} \cdot 4e^{-\frac{|\boldsymbol{\theta}_a - \boldsymbol{\theta}'|}{\eta}} d\boldsymbol{\theta}_a \\
&= \frac{16}{(4\eta)^d} \int_{\boldsymbol{\theta}_a} e^{-\frac{|\boldsymbol{\theta} - \boldsymbol{\theta}'|}{\eta}} \cdot e^{-2\frac{|\boldsymbol{\theta}_a - \frac{\boldsymbol{\theta}' + \boldsymbol{\theta}}{2}|}{\eta}} d\boldsymbol{\theta}_a \\
&= \frac{16}{(4\eta)^d} e^{-\frac{|\boldsymbol{\theta} - \boldsymbol{\theta}'|}{\eta}} \int_{\boldsymbol{\theta}_a \in R^d} e^{-2\frac{|\boldsymbol{\theta}_a - \frac{\boldsymbol{\theta}' + \boldsymbol{\theta}}{2}|}{\eta}} d\boldsymbol{\theta}_a \\
&= \frac{16}{(4\eta)^d} e^{-\frac{|\boldsymbol{\theta} - \boldsymbol{\theta}'|}{\eta}} (\eta)^d \\
&= \frac{16}{4^d} e^{-\frac{|\boldsymbol{\theta} - \boldsymbol{\theta}'|}{\eta}}
\end{aligned}$$

$$
\begin{aligned}
Z_{\text{noise-denoise}}(\boldsymbol{\theta}) &= \sum_{x\in\boldsymbol{\Theta}} p_{\text{noise-denoise}}(x|\boldsymbol{\theta}) \\
&\leq \sum_{x\in\boldsymbol{\Theta}} \frac{16}{4^d} e^{-\frac{|\boldsymbol{\theta}-x|}{\eta}} \\
&\leq \frac{16}{4^d}|\boldsymbol{\Theta}|
\end{aligned}
$$

We also note that for any arbitrary $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$, the following holds true because of Assumption 5.1.

$$
\begin{aligned}
-\langle \nabla U(\boldsymbol{\theta}), \boldsymbol{\theta}' - \boldsymbol{\theta}\rangle &= \langle -\nabla U(\boldsymbol{\theta}) + \nabla U(a), \boldsymbol{\theta}' - \boldsymbol{\theta}\rangle + \langle -\nabla U(a), \boldsymbol{\theta}' - \boldsymbol{\theta}\rangle \\
&\leq \langle -\nabla U(\boldsymbol{\theta}) + \nabla U(a), \boldsymbol{\theta}' - \boldsymbol{\theta}\rangle + \langle -\nabla U(a), \boldsymbol{\theta}' - \boldsymbol{\theta}\rangle \\
&\leq \|-\nabla U(\boldsymbol{\theta}) + \nabla U(a)\|\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| + \|\nabla U(a)\|\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \\
&\leq \| - \nabla U(\boldsymbol{\theta}) + \nabla U(a)\| diam(\boldsymbol{\Theta}) + \|\nabla U(a)\| diam(\boldsymbol{\Theta}) \\
&\leq (M)\ diam(\boldsymbol{\Theta})^2 + \|\nabla U(a)\|\, diam(\boldsymbol{\Theta}).
\end{aligned}
$$

From (8) in Section 4, we can see

$$
\begin{aligned}
a_{\text{init}}(\boldsymbol{\theta}'_{\text{init}}|\boldsymbol{\theta}^{(i-1)}) &= \min\left(1, \frac{\pi(\boldsymbol{\theta}'_{\text{init}}) q_{\text{noise}}(\boldsymbol{\theta}_a^{(i-1)}|\boldsymbol{\theta}'_{\text{init}}) q_{\text{denoise}}(\boldsymbol{\theta}^{(i-1)} \mid \boldsymbol{\theta}_a^{(i-1)})}{\pi(\boldsymbol{\theta}^{(i-1)}) q_{\text{noise}}(\boldsymbol{\theta}_a^{(i-1)}|\boldsymbol{\theta}^{(i-1)}) q_{\text{denoise}}(\boldsymbol{\theta}'_{\text{init}} \mid \boldsymbol{\theta}_a^{(i-1)})}\right) \\
&= \min\left(1, \frac{\pi(\boldsymbol{\theta}'_{\text{init}}) p_{\text{noise-denoise}}(\boldsymbol{\theta}^{(i-1)}|\boldsymbol{\theta}'_{\text{init}})}{\pi(\boldsymbol{\theta}^{(i-1)}) p_{\text{noise-denoise}}(\boldsymbol{\theta}'_{\text{init}}|\boldsymbol{\theta}^{(i-1)})}\right) \\
&\geq \frac{\pi(\boldsymbol{\theta}'_{\text{init}})}{\pi(\boldsymbol{\theta}^{(i-1)})} \cdot \frac{p_{\text{noise-denoise}}(\boldsymbol{\theta}^{(i-1)}|\boldsymbol{\theta}'_{\text{init}})}{p_{\text{noise-denoise}}(\boldsymbol{\theta}'_{\text{init}}|\boldsymbol{\theta}^{(i-1)})} \cdot \frac{Z}{Z} \cdot \frac{Z_{\text{noise-denoise}}(\boldsymbol{\theta}^{(i-1)})}{Z_{\text{noise-denoise}}(\boldsymbol{\theta}'_{\text{init}})} \\
&\geq \exp\left(U(\boldsymbol{\theta}'_{\text{init}}) - U(\boldsymbol{\theta}^{(i-1)})\right) \frac{\frac{1}{4^d} e^{-\frac{|\boldsymbol{\theta}'_{\text{init}} - \boldsymbol{\theta}^{(i-1)}|}{\eta}}}{\frac{16}{4^d} e^{-\frac{|\boldsymbol{\theta}^{(i-1)} - \boldsymbol{\theta}'_{\text{init}}|}{\eta}}} \cdot \frac{\frac{|\boldsymbol{\Theta}|}{4^d} e^{-\frac{\Delta(\boldsymbol{\Theta})}{\eta}}}{\frac{16|\boldsymbol{\Theta}|}{4^d}} \\
&\geq \frac{1}{2^8} \exp\left(\left\langle \nabla U(\boldsymbol{\theta}^{(i-1)}), \boldsymbol{\theta}'_{\text{init}} - \boldsymbol{\theta}^{(i-1)} \right\rangle\right) \cdot \exp\left(-\frac{\Delta(\boldsymbol{\Theta})}{\eta}\right) \\
&\geq \frac{1}{2^8} \exp\left(-M diam(\boldsymbol{\Theta})^2 - \|\nabla U(a)\| diam(\boldsymbol{\Theta}) - \frac{\Delta(\boldsymbol{\Theta})}{\eta}\right)
\end{aligned}
$$

Putting everything together,

$$
\begin{aligned}
\int p_{\text{MwG}}(\boldsymbol{\theta}_{\text{init}}^{(i)}, \boldsymbol{\theta}_a^{(i-1)} \mid \boldsymbol{\theta}^{(i-1)})\, d\boldsymbol{\theta}_a^{(i-1)} &= \int q_{\text{noise}}(\boldsymbol{\theta}_a^{(i-1)} \mid \boldsymbol{\theta}^{(i-1)}) q_{\text{denoise}}(\boldsymbol{\theta}'_{\text{init}} \mid \boldsymbol{\theta}_a^{(i-1)}) a_{\text{init}}(\boldsymbol{\theta}'_{\text{init}}|\boldsymbol{\theta}^{(i-1)}) \\
&\quad + \left(1 - L(\widetilde{\boldsymbol{\theta}}^{(i-1)})\right) \delta(\boldsymbol{\theta}_{\text{init}}^{(i)})\, d\boldsymbol{\theta}_a^{(i-1)} \\
&\geq \int q_{\text{noise}}(\boldsymbol{\theta}_a^{(i-1)} \mid \boldsymbol{\theta}^{(i-1)}) q_{\text{denoise}}(\boldsymbol{\theta}'_{\text{init}} \mid \boldsymbol{\theta}_a^{(i-1)}) a_{\text{init}}(\boldsymbol{\theta}'_{\text{init}}|\boldsymbol{\theta}^{(i-1)}) d\boldsymbol{\theta}_a^{(i-1)} \\
&= a_{\text{init}}(\boldsymbol{\theta}'_{\text{init}}|\boldsymbol{\theta}^{(i-1)}) \int q_{\text{noise}}(\boldsymbol{\theta}_a^{(i-1)} \mid \boldsymbol{\theta}^{(i-1)}) q_{\text{denoise}}(\boldsymbol{\theta}'_{\text{init}} \mid \boldsymbol{\theta}_a^{(i-1)}) d\boldsymbol{\theta}_a^{(i-1)} \\
&= a_{\text{init}}(\boldsymbol{\theta}'_{\text{init}}|\boldsymbol{\theta}^{(i-1)}) \frac{p_{\text{noise-denoise}}(\boldsymbol{\theta}'_{\text{init}}|\boldsymbol{\theta}^{(i-1)})}{Z_{\text{noise-denoise}}(\boldsymbol{\theta}^{(i-1)})} \\
&\geq \frac{1}{2^8} \cdot \frac{\frac{1}{4^d} e^{-\frac{|\boldsymbol{\theta}^{(i-1)} - \boldsymbol{\theta}'_{\text{init}}|}{\eta}}}{\frac{16}{4^d}|\boldsymbol{\Theta}|} \exp\left(-M \text{diam}^2(\boldsymbol{\Theta}) - \|\nabla U(a)\| \text{diam}(\boldsymbol{\Theta}) - \frac{\Delta(\boldsymbol{\Theta})}{\eta}\right) \\
&\geq \frac{1}{2^{12}|\boldsymbol{\Theta}|} \exp\left(-M \text{diam}^2(\boldsymbol{\Theta}) - \|\nabla U(a)\| \text{diam}(\boldsymbol{\Theta}) - 2\frac{\Delta(\boldsymbol{\Theta})}{\eta}\right)
\end{aligned}
$$

Thus,

$$\int p_{\mathrm{MwG}}(\boldsymbol{\theta}_{\mathrm{init}}^{(i)}, \boldsymbol{\theta}_a^{(i-1)} \mid \boldsymbol{\theta}^{(i-1)})\, d\boldsymbol{\theta}_a^{(i-1)} \geq \frac{1}{2^{12}|\boldsymbol{\Theta}|} \exp\left(-M\mathrm{diam}^2(\boldsymbol{\Theta}) - \|\nabla U(a)\|\mathrm{diam}(\boldsymbol{\Theta}) - 2\frac{\Delta(\boldsymbol{\Theta})}{\eta}\right) \tag{18}$$

Combining Equation (18) and Lemma C.1, we see,

$$\begin{aligned}
\kappa_{(\mathrm{HiSS})}(\boldsymbol{\theta}^{(G)} \mid \boldsymbol{\theta}^{(0)}) &= \prod_{i=1}^{G} \kappa(\boldsymbol{\theta}^{(i)} \mid \boldsymbol{\theta}^{(i-1)}) \\
&= \prod_{i=1}^{G} \int p_{\mathrm{MwG}}(\boldsymbol{\theta}_{\mathrm{init}}^{(i)}, \boldsymbol{\theta}_a^{(i-1)} \mid \boldsymbol{\theta}^{(i-1)}) \cdot p_{\mathrm{DMALA}}(\boldsymbol{\theta}^{(i)} \mid [{\boldsymbol{\theta}_{\mathrm{init}}^{(i)}}^T, {\boldsymbol{\theta}_a^{(i-1)}}^T]^T)\, d\boldsymbol{\theta}_a^{(i-1)} \\
&\geq \prod_{i=1}^{G} \exp\left\{(-\frac{ML}{2} - \frac{L}{\alpha} + \frac{mL}{4})\mathrm{diam}^2(\boldsymbol{\Theta}) - (\frac{L}{2}\|\nabla U(a)\| + \frac{3L\sqrt{d}+L}{\eta})\mathrm{diam}(\boldsymbol{\Theta})\right\} \cdot \\
&\int p_{\mathrm{MwG}}(\boldsymbol{\theta}_{\mathrm{init}}^{(i)}, \boldsymbol{\theta}_a^{(i-1)} \mid \boldsymbol{\theta}^{(i-1)}) d\boldsymbol{\theta}_a^{(i-1)} \\
&\geq \prod_{i=1}^{G} \left[\nu_i(A) \exp\left\{(-\frac{ML}{2} - \frac{L}{\alpha} + \frac{mL}{4})\mathrm{diam}^2(\boldsymbol{\Theta}) - (\frac{L}{2}\|\nabla U(a)\| + \frac{3L\sqrt{d}+L}{\eta})\mathrm{diam}(\boldsymbol{\Theta})\right\} \cdot \right. \\
&\left. \frac{\exp\left(-M\mathrm{diam}^2(\boldsymbol{\Theta}) - \|\nabla U(a)\|\mathrm{diam}(\boldsymbol{\Theta}) - 2\frac{\Delta(\boldsymbol{\Theta})}{\eta}\right)}{2^{12}|\boldsymbol{\Theta}|}\right] \\
&\geq \frac{1}{2^{12G}|\boldsymbol{\Theta}|^G} \cdot \prod_{i=1}^{G} \left[\nu_i(A) \exp\left\{(-\frac{ML}{2} - \frac{L}{\alpha} + \frac{mL}{4})\mathrm{diam}^2(\boldsymbol{\Theta}) - (\frac{L}{2}\|\nabla U(a)\| \right.\right. \\
&\left.\left. + \frac{3L\sqrt{d}+L}{\eta})\mathrm{diam}(\boldsymbol{\Theta})\right\} \cdot \exp\left(-M\mathrm{diam}^2(\boldsymbol{\Theta}) - \|\nabla U(a)\|\mathrm{diam}(\boldsymbol{\Theta}) - \frac{2\sqrt{d}}{\eta}\mathrm{diam}(\boldsymbol{\Theta})\right)\right] \\
&= \nu(A') \cdot \exp\left\{(-M(\frac{LG}{2}+G) - \frac{LG}{\alpha} + \frac{mLG}{4})diam(\boldsymbol{\Theta})^2 + (\frac{3LG\sqrt{d}+LG-2G\sqrt{d}}{\eta} - \right. \\
&\left.(\frac{LG}{2}+G)\|\nabla U(a)\|)diam(\boldsymbol{\Theta})\right\} \\
&= \nu(A') \cdot \exp\left\{(-M(\frac{LG}{2}+G) - \frac{LG}{\alpha} + \frac{mLG}{4})diam(\boldsymbol{\Theta})^2 + (\frac{G\sqrt{d}(3L-2)+LG}{\eta} - \right. \\
&\left.(\frac{LG}{2}+G)\|\nabla U(a)\|)diam(\boldsymbol{\Theta})\right\}
\end{aligned}$$

where $\nu(A') \in [0,1]$ is a valid probability measure over L trajectories across G Gibbs sweeps i.e. $A' \subseteq \boldsymbol{\Theta}^{L\times G}$.

□

Corollary C.2 (Analytical Condition for Fast Mixing)**.** *Under the assumptions of Theorem 5.6, suppose the proposal scale $\eta > 0$, and the number of Gibbs sweeps $G \geq 1$, refinements per sweeps $L \geq 1$, and parameter space diameter $\mathrm{diam}(\boldsymbol{\Theta}) > 0$. Also, let $m < 4M$ and $\alpha < \frac{2}{M}$. In particular, fast mixing (i.e., large ϵ_α) occurs whenever*

$$\|\nabla U(a)\| \leq \frac{\sqrt{d}(3L-2)+L}{\eta\left(\frac{L}{2}+1\right)}.$$

Proof. Let convergence factor be expressed as

$$\epsilon_\alpha = \exp\left\{A \cdot \mathrm{diam}(\boldsymbol{\Theta})^2 + B \cdot \mathrm{diam}(\boldsymbol{\Theta})\right\}$$

where

$$A = -M\left(\frac{LG}{2} + G\right) - \frac{LG}{\alpha} + \frac{mLG}{4},$$
$$B = \frac{G\sqrt{d}(3L-2) + LG}{\eta} - \left(\frac{LG}{2} + G\right) \cdot \|\nabla U(a)\|.$$

For faster mixing, the total variation distance must decay quickly. Thus,$\epsilon_\alpha \to 1$, making $\left\{A \cdot \text{diam}(\boldsymbol{\Theta})^2 + B \cdot \text{diam}(\boldsymbol{\Theta})\right\} \to 0$.

We know, $\text{diam}(\boldsymbol{\Theta}) > 0$. Therefore, $\text{diam}(\boldsymbol{\Theta}) = -\frac{B}{A} > 0$.

Trivially, if $m < 4M$ and $\alpha < \frac{2}{M}$,then $A < 0$. B becomes non-negative when,

$$\|\nabla U(a)\| < \frac{\sqrt{d}(3L-2) + L}{\eta(\frac{L}{2} + 1)}$$

For fixed values of η, L, and d, favorable local mixing is possible only if the flattest region of the energy landscape is not too steep i.e. ($\|\nabla U(a)\| \to 0$). In other words, if the pull-back from the gradient is less than exploration strength of the sampler, chain mixes well. This inherently shows HiSS's success in disconnected energy regimes.

In contrast, a symmetric analysis of DMALA (Pynadath et al. (2024)) reveals that local mixing improves only when $\|\nabla U(a)\| > 0$, which relies on the assumption that the energy landscape is very well connected. □

D Additional Experimental Results

Coverage

Let $\boldsymbol{\Theta}$ denote the set of all possible discrete states with cardinality $|\boldsymbol{\Theta}|$, and let:

- $S = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N\}$ be the set of samples generated by the MCMC sampler after N iterations.
- $\mathcal{V}(S) \subseteq \boldsymbol{\Theta}$ represent the subset of states in $\boldsymbol{\Theta}$ visited at least once by the sampler.

The **coverage** C of the MCMC sampler is defined as:

$$C = \frac{|\mathcal{V}(S)|}{|\boldsymbol{\Theta}|},$$

where:

- $|\mathcal{V}(S)|$ is the number of unique states visited by the sampler.
- $|\boldsymbol{\Theta}|$ is the total number of discrete states.

Properties of Coverage

- **Range**: $C \in [0, 1]$:
  - $C = 0$: No states were visited, i.e., $|\mathcal{V}(S)| = 0$.
  - $C = 1$: All possible states were visited, i.e., $|\mathcal{V}(S)| = |\boldsymbol{\Theta}|$.
- **Interpretation**:
  - C represents the fraction of the state space explored by the sampler.
  - A higher C indicates better exploration and diversity of the sampled states.

D.1 4D Joint Bernoulli

For Section 6, we manually tuning the step size α to 0.2. For HiSS, we set $G = 5, L = 2$, and $\eta = 4$. To provide additional insights into the functionality of HiSS, we explore their behavior on the 4D Joint Bernoulli Distribution, which serves as the simplest low-dimensional case among our experiments. This aids in visualizing and understanding the sampling process.

Target Distribution

The following represents the probability mass function (PMF) for the 4D Joint Bernoulli Distribution used in our test case. The distribution has 16 states with the corresponding probabilities:

$$P_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = \begin{cases} 0.588204 & \text{if } \boldsymbol{\theta} = 0000, \\ 5.882e-6 & \text{if } \boldsymbol{\theta} = 0001, \\ 5.882e-6 & \text{if } \boldsymbol{\theta} = 0010, \\ 5.882e-6 & \text{if } \boldsymbol{\theta} = 0011, \\ 5.882e-6 & \text{if } \boldsymbol{\theta} = 0100, \\ 5.882e-6 & \text{if } \boldsymbol{\theta} = 0101, \\ 5.882e-6 & \text{if } \boldsymbol{\theta} = 0110, \\ 5.882e-6 & \text{if } \boldsymbol{\theta} = 0111, \\ 5.882e-6 & \text{if } \boldsymbol{\theta} = 1000, \\ 5.882e-6 & \text{if } \boldsymbol{\theta} = 1001, \\ 5.882e-6 & \text{if } \boldsymbol{\theta} = 1010, \\ 5.882e-6 & \text{if } \boldsymbol{\theta} = 1011, \\ 5.882e-6 & \text{if } \boldsymbol{\theta} = 1100, \\ 5.882e-6 & \text{if } \boldsymbol{\theta} = 1101, \\ 0.294102 & \text{if } \boldsymbol{\theta} = 1110, \\ 0.117641 & \text{if } \boldsymbol{\theta} = 1111. \end{cases}$$

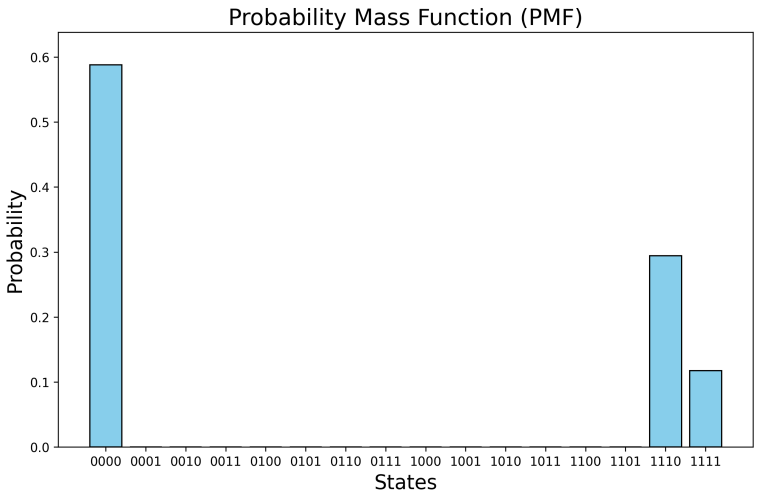

Figure 11: Target Distribution for 4D Joint Bernoulli

The state $\boldsymbol{\theta} = 0000$ has the highest probability ($P_{\boldsymbol{\Theta}}(0000) = 0.588204$), indicating it is the dominant mode. States $\boldsymbol{\theta} = 1110$ and $\boldsymbol{\theta} = 1111$ have moderate probabilities (0.294102 and 0.117641 , respectively), while all other states have extremely low probabilities (5.882×10^{-6}), highlighting a multimodal distribution with sharp peaks and a long tail of negligible probabilities.

Coverage Analysis

As shown in Figure 12, HiSS exhibits a clear upward trend in average coverage throughout the iterations, showcasing its superior ability to explore diverse modes effectively. This improvement can be attributed to the MwG sweep mechanism, which enhances its exploratory capacity. In contrast, other samplers, such as GWG, DMALA, ACS, and PT+DMALA, appear to stagnate early in the iterations, failing to escape initial regions and achieve broader coverage.

D.2 Ising Sampling

Target Distribution

For Section 6, for HiSS, we set $G = 10, L = 2, \alpha = 0.2$, and $\eta = 4$. For our setup, the interaction matrix $\mathbf{W}$ is essentially a cross-diagonal matrix. This is intentional, causing the Ising Model to be sparse, breaking the symmetry-like effect. $\mathbf{W}$'s construction is motivated by the study of frustrated and anisotropic systems(Edwards and Anderson, 1975; Chaikin and Lubensky, 1995), where competing interactions and directional dependencies govern dynamics. For $d = 9$, the Ising model contains $2^9 = 512$ discrete states. The probability distribution, visualized in Figure 13, reveals that 32 prominent states dominate the landscape, accounting for $\frac{32}{512} = 6.25\%$ of the total state space. These high-probability states illustrate the multimodal nature of the model, where efficient sampling requires the ability to transition effectively between modes.

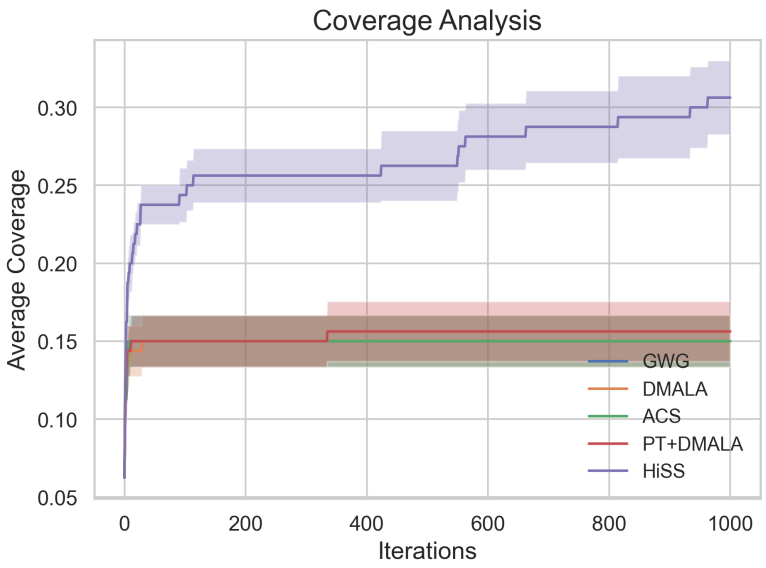

Figure 12: Coverage Analysis for 4D Bernoulli

Coverage Analysis

In Figure 14, HiSS demonstrates excellent performance, rapidly converging to cover these prominent states. Its upward trajectory showcases its ability to explore the multimodal landscape efficiently, achieving the theoretical limit of 0.0625 well before other samplers. This success can be attributed to mechanisms like the MwG Gibbs sweep, which ensures good mixing and fast convergence. While PT+DMALA achieves comparable coverage in the long run, its inefficiency at earlier iterations highlights its limitations for tasks requiring faster convergence. In contrast, samplers like GWG, DMALA, and ACS struggle to escape initial regions of the probability landscape and fail to achieve sufficient mixing. These methods stall early, underscoring their inability to effectively explore the multimodal nature of the Ising model.

Tuning G and L

Increasing G enhances global mixing by allowing more opportunities to transition between modes, ensuring broader coverage of the state space. However, excessive G without sufficient refinement can lead to a *jumpy* process. Conversely, increasing L enables smoother and more informed transitions, improving local convergence. Yet, overly large L risks making transitions too deterministic, potentially trapping the chain in local modes. Thus, the optimal configuration depends on the energy landscape: smoother distributions benefit from higher L and moderate G, whereas highly disconnected, multimodal landscapes require a larger G with appropriately scaled η. Parameter selection should therefore be guided by the structure of the target distribution and the desired tradeoff between exploration and exploitation.

To study this phenomena, we fixed the product $G \times L$ and evaluated all integer factor combinations (e.g., $G = 1, L = 50; G = 2, L = 25; \cdots, G = 50, L = 1$) for a fixed α and η. We then plotted the Total Variational Distance versus runtime. The observed trend aligns with our theoretical intuition: higher G and lower L enhance convergence speed (due to improved mode mixing) but come with increased runtime overhead (due to more Metropolis-within-Gibbs steps). Conversely, lower G and higher L slow down convergence but reduce computational cost (as illustrated in Figure 15).

Criticality

To investigate the limits of gradient-based sampling, we analyze 2D Ising model under Criticality($\beta_c = 0.4407$) for $d = 24 \times 24$, with $\alpha = \frac{\beta_c}{2}$ accounting for the double-counting of bonds in the quadratic form and $b = 0$. Interestingly, we observed that HiSS and the baseline DMALA exhibit almost identical performance, with both converging to the theoretical internal energy (≈ -1.4402). This suggests that while critical systems suffer from

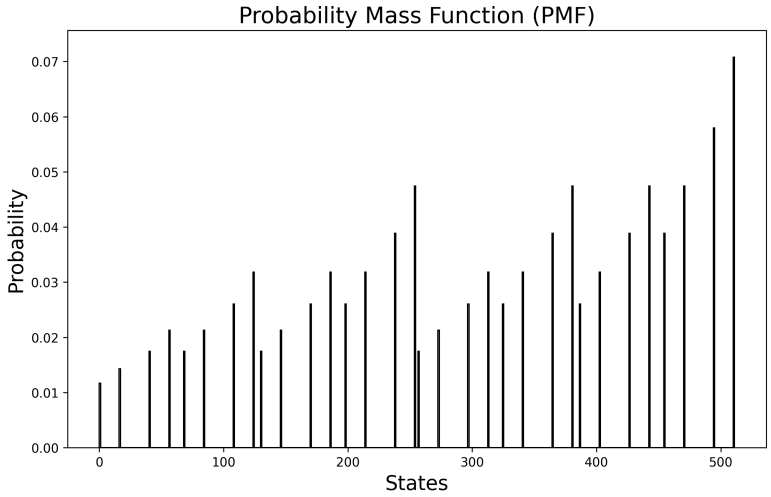

Figure 13: Ising Model Distribution

'critical slowing down', they do not necessarily exhibit the disconnected energy landscape that traps gradient samplers. The gradients in the critical Ising model still provide a valid path for global exploration(Figure 16).

D.3 Computational Complexity Analysis (Energy Evaluations)

While Section 6 provides a practical convergence analysis based on wall-clock runtime (which accounts for implementation overheads, memory access, and communication latency), it is also valuable to analyze the theoretical computational strain of each sampler in terms of Total Number of Function Evaluations (TNFE). We define the Total NFE as the aggregate number of energy function calls required to generate the full set of samples used in our results. In many scientific applications, the energy function $U(\boldsymbol{\theta})$ is the computational bottleneck, making NFE a critical metric independent of hardware optimizations (Noé et al., 2019; Song et al., 2021).

We define the cost of a single gradient-based transition (e.g., one step of DMALA or GWG) as $\mathcal{C}_{grad} = 4$ energy calls:

Gradient Calculation: Requires 2 evaluations (forward and backward passes, or function evaluations at $\boldsymbol{\theta}$ and new $\boldsymbol{\theta}$) to approximate or compute $\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$. Metropolis-Hastings (MH) Correction: Requires 2 evaluations to compute the energy of the proposed state $U(\boldsymbol{\theta}')$ and the current state $U(\boldsymbol{\theta})$ for the acceptance ratio.

For all experiments, we collect N_S samples. Let N be number of independent chains (batch size. We assume a fixed budget of refinement steps per sample across baselines to ensure parity.

- Baseline Gradient Samplers (DMALA, GWG, ACS)For standard gradient-based samplers, the process is a sequential application of the kernel. With S/GL refinement steps per sample:

$$\text{NFE}_{base} = N \times N_S \times S \times \mathcal{C}_{grad}$$

- HiSS (Proposed Method) HiSS introduces a hierarchical structure with an outer loop (G sweeps) and an inner refinement loop (L steps). The cost per sweep includes the MH step for the denoised proposal (2 evaluations) and the gradient refinement ($L \times \mathcal{C}_{grad}$).

$$\text{NFE}_{HiSS} = N \times N_S \times G \times [2 + (L \times \mathcal{C}_{grad})]$$

- Parallel Tempering (PT+DMALA) Parallel Tempering incurs significantly higher computational costs due to the maintenance of multiple auxiliary chains at higher temperatures. For a batch of T chains, each utilizing K temperature levels, with swap attempts every I steps, the cost is twofold:

  Refinement Cost: All K replicas must undergo Langevin dynamics. Swap Cost: Every I steps, energy differences between adjacent chains must be computed to satisfy the exchange criterion.

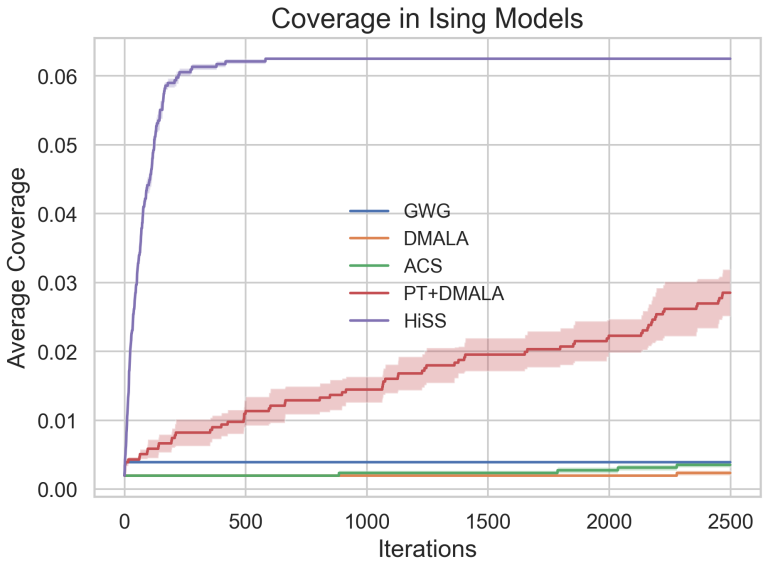

Figure 14: Coverage Analysis for Ising Model

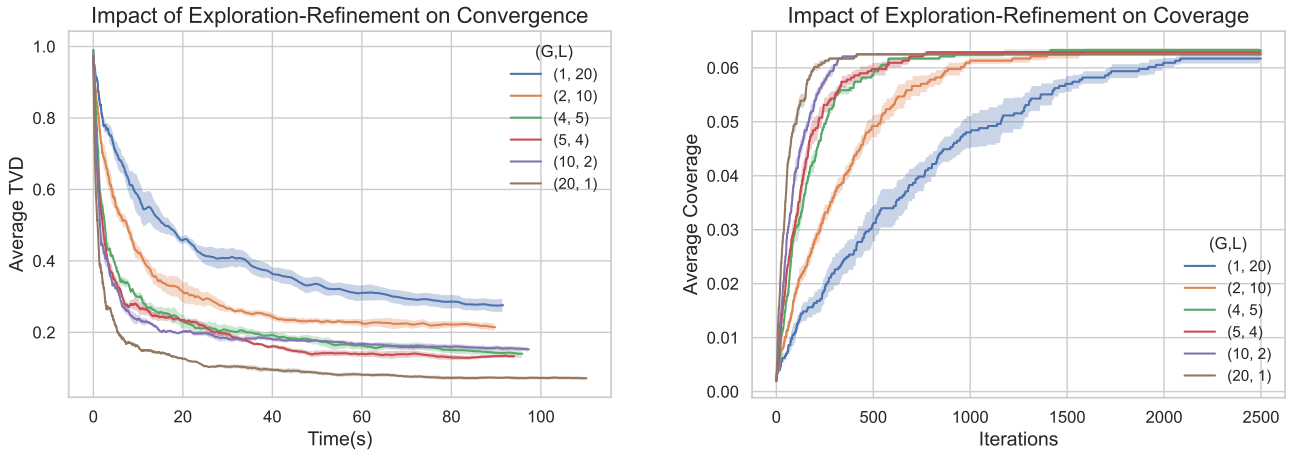

Figure 15: Impact of Gibbs Sweeps and Refinement Iterations in Ising models.

$$\text{NFE}_{PT} = \underbrace{N_S \times N \times K \times S \times \mathcal{C}_{grad}}_{\text{Thermodynamic Refinement}} + \underbrace{\left\lfloor \frac{N_S \cdot S}{I} \right\rfloor \times N \times (K-1) \times 2}_{\text{Swap Communication}}$$

For 4D Joint Bernoulli Sample, $N = 10, N_S = 10^3$, $S = 10 \quad (G = 5, L = 2)$, $K = 5$ temperatures, $I = 4$.

From Table 3, Conclusion: While HiSS incurs a marginal increase in NFE compared to pure DMALA (5×10^5 vs 4×10^5) to support the auxiliary variable mechanics; PT requires 2.2×10^6 NFEs(a 4.4× with respect to HiSS increase in raw energy computations). This confirms that PT's "slower" wall-clock convergence (Figure 2) is not just due to communication overhead, but due to the sheer volume of wasted computation on high-temperature auxiliary chains that are discarded during inference. HiSS achieves mode-hopping more efficiently by using a single continuous auxiliary variable rather than K discrete replicas. TNFE values for Ising Models are reported in Table 4.

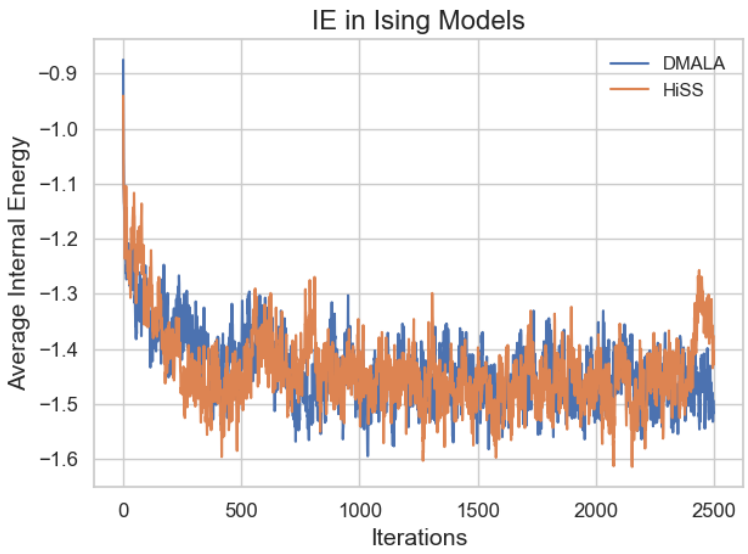

Figure 16: Criticality Ising Model

Table 3: Number of Energy Evaluations in 4D Joint Bernoulli

Sampler	**# Energy Evaluations**
GWG	4.0×10^5
DMALA	4.0×10^5
ACS	4.0×10^5
PT+DMALA	2.2×10^6
HiSS	5.0×10^5

D.4 Traveling Salesman Problem

For results presented under Section 6, for HiSS, we set $G = 10$, $L = 4$, $\alpha = 0.02$, and $\eta = 2$.

In this section, we gauge to assess the impact of η on the quality of the solutions sampled for HiSS. By employing $\alpha = 10^{-4}$, $G = 10$, and $L = 4$, we plot the average PMC and Jaccard Similarity metrics, along with their respective standard deviations, as η is progressively increased.

As evident from Figure 17, the sample diversity improves as η increases. This observation aligns with logical intuition, as larger η enables the sampler to explore the state space more effectively.

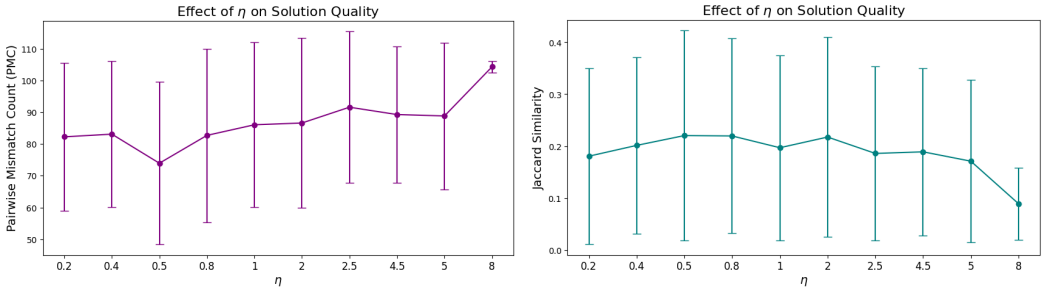

Figure 17: Impact of scale of logistic noise on solution quality.

Table 4: Number of Energy Evaluations in Ising Models

Sampler	**# Energy Evaluations**
GWG	1.00×10^6
DMALA	1.00×10^6
ACS	1.00×10^6
PT+DMALA	6.00×10^6
HiSS	1.25×10^6

D.5 Binary Bayesian Neural Networks

For Section 6, for HiSS, we set the parameters $G = 10$, $L = 5$, and $\alpha = 0.1$. As shown in Table 5, for the smaller Breast Cancer dataset (with $\eta = 0.005$), DMALA outperforms HiSS in terms of average training log-likelihood and RMSE. However, on the larger COMPAS dataset (with $\eta = 2$),HIV dataset($\eta = 4$), and Blog($\eta = 5$) HiSS achieves the lowest training RMSE. For each dataset, η values are chosen based on validation RMSE.

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

Dataset	Average Training Log-likelihood (↑)					Average Training Root-Mean Square Error (↓)				
	GWG	DMALA	ACS	PT+DMALA	HiSS	GWG	DMALA	ACS	PT+DMALA	HiSS
Breast Cancer	-0.0708 ±0.0089	**-0.0693** ±0.0025	-0.0938 ±0.0060	-0.0721 ±0.0049	-0.0728 ±0.0028	0.0283 ±0.0032	**0.0276** ±0.0011	0.0315 ±0.0018	0.0287 ±0.002	0.0291 ±0.0013
COMPAS	-0.3130 ± 0.0068	**-0.3121** ± 0.0027	-0.3139 ± 0.0037	-0.3149 ± 0.0057	-0.3697 ± 0.0073	0.2213 ±0.0019	0.2219 ±0.0020	0.2230 ±0.0020	0.2215 ± 0.0020	**0.2172** ± 0.0024
HIV	-0.7453 ± 0.0750	-0.7746 ± 0.0000	-0.7746 ± 0.0000	-0.7746 ± 0.0000	**-0.4578** ± 0.0101	0.7299 ± 0.1205	0.7746 ± 0.0000	0.7746 ± 0.0000	0.7746 ± 0.0000	**0.2554** ± 0.0100
Blog	**-0.3458** ± 0.0127	-0.3476 ± 0.0000	-0.3476 ± 0.0000	-0.3476 ± 0.0000	-0.4094 ± 0.0092	0.3320± 0.0453	0.3476 ± 0.0000	0.3476 ± 0.0000	0.3476 ± 0.0000	**0.2043** ± 0.0088

Dataset Details

- **Breast Cancer** (Wolberg et al., 1993): This dataset contains 569 instances of digitized fine needle aspirates (FNAs) of breast masses. The task involves predicting whether the instance is benign or malignant. For prediction, we use 30 real-valued attributes, and the dimensionality of sampling vector is 3,201.
- **COMPAS** (J. Angwin and Kirchner, 2016): This dataset includes criminal records of 6,172 individuals arrested in Florida. The task is to predict whether an individual will re-offend within two years. We utilize 13 attributes for prediction. The dimensionality of sampling vector is 1,501.
- **HIV** (Rgnvaldsson, 2015) : This dataset contains 1,625 instances of octamers (8-amino-acid sequences). The binary classification task is to predict whether a sequence is a cleavage site for the HIV-1 protease enzyme. The input sequences are transformed into features via one-hot encoding across 20 standard amino acids, resulting in a 160-dimensional binary feature vector for each instance. This makes the dimensionality of the sampling vector 16,201.
- **Blog** (Buza, 2014) : This 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. This makes the dimensionality of the sampling vector 28,201.

To create a challenging, disconnected posterior landscape characterized by isolated modes, we introduce a sparsity inducing prior on the network weights for HIV and Blog Datasets. This approach is conceptually motivated by the classic Spike-and-Slab framework for Bayesian variable selection (George and McCulloch, 1993; Koyejo et al., 2014). In practice, we implement this as a Laplace prior, which is the Bayesian equivalent of the well-known L1/Lasso penalty (Tibshirani, 1996; Park and Casella, 2008). Applying such priors to encourage sparsity and prune connections is a highly active area of research, with recent applications to both continuous and binarized neural networks (Louizos et al., 2018). This prior forces the BNN to find solutions where most weights are in a default *off* state, creating deep energy wells at sparse configurations and high energy barriers between them.