Slithering through Gaps: Capturing Discrete Isolated Modes via Logistic Bridging

Anonymous Author(s)
Affiliation
Address
email

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) [52, 44], 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[5, 37, 46]. Techniques like cyclical step size[69], PT(parallel tempering)[59], and flat-histogram approaches[4] have been developed to address this issue, improving mixing efficiency in continuous spaces by facilitating mode transitions.

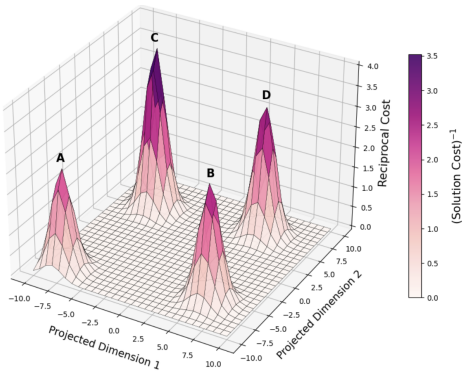

Figure 1: A visualization of the multimodal cost landscape in the Traveling Salesman Problem, showcasing solutions separated by low energy barriers.

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)[23, 14, 33], protein coupling prediction[32], low-precision neural networks[11], and combinatorial optimization

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

[2]. 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)[71] and Gibbs-With-Gradient (GWG)[22], 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[47] 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**yperbol**i**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 implementation at `https://anonymous.4open.science/r/HiSS-8004/README.md`.

2 Related Works

Gradient-Based Discrete Sampling. Gradient-based discrete sampling methods have significantly advanced, enhancing efficiency and applicability. Locally informed proposals[68] proposals using probability ratios improved convergence, while later gradient-based approaches[21] to approximate the probability ratio, boosted efficiency . Subsequent works proposed various gradient-based techniques [50, 57, 58, 56, 67]. DLP 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 [40], cyclical step sizes [69], parallel tempering [59], Swendsen-Wang algorithm [60], flat histograms [4, 13], and Wolff algorithm [66] have been widely employed to facilitate mode transitions. More recently, Diffusive Gibbs (DiGs) [10] used Gaussian Convolution on continuous multimodal distributions.

Multimodal Sampling in Discrete Spaces. Most prior works on multimodal discrete sampling rely on combinatorial or swap-based proposals like parallel tempering [59], Wang-Landau Sampling[63], and early Mode-Jumping MCMC [39]. Based on gradient-based discrete samplers, [47] 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.

Unlike prior works, HiSS efficiently samples from disconnected modes in discrete spaces by running a single chain. A detailed comparison between these methods is provided in Appendix B

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 [21, 57, 71]: 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.

Discrete Langevin Proposal. The DLP is an extension of the Langevin algorithm tailored for discrete spaces [71]. At a given position $\boldsymbol{\theta}$, the proposal distribution $q(\cdot|\boldsymbol{\theta})$ determines the next position. The proposal distribution is formulated as:

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

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 [1]. 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:

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

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 [30, 38]. We empirically demonstrate the advantages of logistic kernel in Section 6.5 and Appendix A.

Inspired by the coupling method used in [34, 42], 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:*

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

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, the gradient of it is also easy to compute, $\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\boldsymbol{\theta}_a) = \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) + \frac{1}{\eta}\tanh\left(\frac{\boldsymbol{\theta}_a - \boldsymbol{\theta}}{2\eta}\right)$. This allows us to use gradient-based discrete samplers like DMALA [71]. Similarly, the conditional distribution for the auxiliary variable $\boldsymbol{\theta}_a$ is, $p(\boldsymbol{\theta}_a|\boldsymbol{\theta}) \sim Logistic(\boldsymbol{\theta}, \eta)$, 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 (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 $\boldsymbol{\theta}$.

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: $\mathcal{S}$

4.3.3 Metropolis-Hastings Acceptance for Denoised Proposal

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

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

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

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 C. 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 [10], 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.

5 Theoretical Analysis

In this section, we provide asymptotic and non-asymptotic convergence guarantees for HiSS. We make similar assumptions as in [47]. 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 [8, 12, 16]. Under Assumption 5.2, $U(\cdot)$ is m-strongly concave on conv($\boldsymbol{\Theta}$), following Lemma C.3 from [47].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 $\pi(\boldsymbol{\theta})$ and 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 $\epsilon_\alpha = \exp\left\{(-M(\frac{LG}{2}+G) - \frac{LG}{\alpha} + \frac{mLG}{4})diam(\boldsymbol{\Theta})^2 + (\frac{G\sqrt{d}(3L-2)+LG}{\eta} - (\frac{LG}{2}+G)\|\nabla U(a)\|)diam(\boldsymbol{\Theta})\right\}$

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 D for detailed proofs.

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 [71, 42]. We compare HiSS against some popular pure gradientd-based baselines like Gibbs with Gradient (GWG) [21], Discrete Metropolis-Adjusted Langevin Algorithm (DMALA) [71]. Other baselines targeting discrete multimodal distributions such as , Automatic Cyclical Sampler (ACS) [47], and Parallel Tempering(PT) [59] are also included. Being consistent with [10], 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 [17, 31]. We use the same stepsize α for DMALA, ACS, PT+DMALA, and HiSS to ensure fair comparison. We also ensure the number of iterations per sample for all samplers across all tasks is the same for fair runtime comparison.

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 10 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

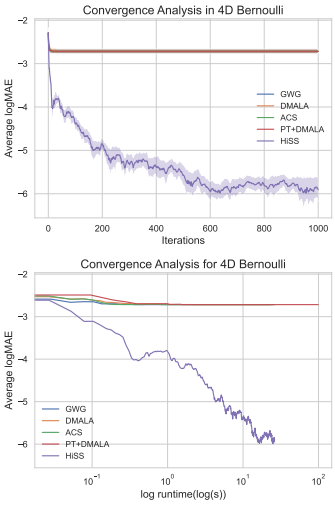

Figure 2: 4D Joint Bernoulli

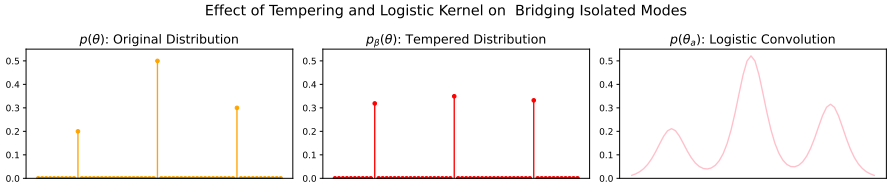

Figure 3: Comparison of the target distribution, tempered distribution, and $p(\boldsymbol{\theta}_a)$ in HiSS.

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

Why HiSS Outperforms PT? HiSS significantly outperforms PT because PT adjusts the inverse temperature $\beta = \frac{1}{T}$ to enhance exploration, but disconnected modes 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 log root-mean-square error (RMSE) between the estimated mean and the true mean with standard error against iterations and runtime in Figure 4. The convergence analysis reveals that HiSS significantly outperforms other samplers, achieving the lowest log RMSE (-3.0) 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 RMSE levels. HiSS achieves convergence faster than other methods, while PT+DMALA has slower runtime performance due to high communication overhead. GWG, DMALA, and ACS show limited improvement. We provide hyperparameter settings, additional results, and diagnostics in the Appendix E.

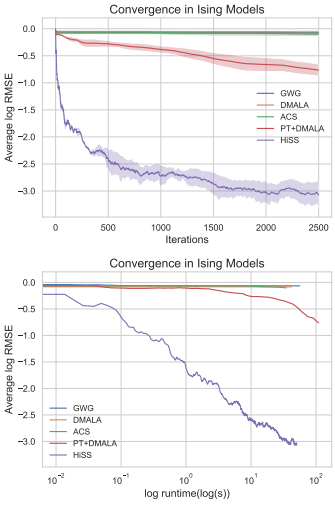

Figure 4: Ising Model

6.3 Binary Bayesian Neural Networks

The posterior distribution of binary neural networks (BNNs) [11, 48, 36] is highly multimodal, characterized by disconnected or isolated modes[70, 27]. To investigate this, we perform regression tasks on two UCI datasets [15], 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 500 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 1). Notably, HiSS outperforms other baseline methods in terms of generalization to unseen data. Consistent with the findings of [10] 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 E.

Table 1: Experiment results with binary Bayesian neural networks on two 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	**-0.0237** ± 0.0014	0.1553 ± 0.0082	0.1550 ± 0.0047	0.1673 ± 0.0065	0.1568 ± 0.0060	**0.1541** ± 0.0047
COMPAS	-0.2265 ± 0.0025	-0.2271 ± 0.0025	-0.2284 ± 0.0025	-0.2265 ± 0.0026	**-0.2237** ± 0.0039	0.4759 ± 0.0027	0.4766 ± 0.0026	0.4779 ± 0.0026	0.4759 ± 0.0027	**0.4729** ± 0.0042

6.4 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:

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

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 [49]. 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 2: Performance comparison of different samplers on the eil14 dataset.

Sampler	**Cost**(↓)	**PMC** (↑)	**Jaccard Similarity** (↓)	**No. of Unique Solutions**(↑)
GWG	370.7086 ± 19.7404	027.0000 ± 00.0000	0.5555 ± 0.0000	02
DMALA	339.0111 ± 43.5156	084.2444 ± 23.4995	0.1483 ± 0.1481	10
ACS	382.8639 ± 02.5501	**105.0000** ± 00.0000	**0.0370** ± 0.0000	02
PT+DMALA	337.1584 ± 44.9881	080.4167 ± 19.7419	0.1402 ± 0.1357	09
HiSS	**277.9008** ± 19.8467	103.0727 ± 09.1807	0.0990 ± 0.0964	**11**

Table 2 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 [7, 24, 62] and Kendall's Tau [19], and Jaccard similarity [35] while ensuring samples discovered are low-cost solutions. We provide additional insights and hyperparameter settings in Appendix E.

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 (log-MAE) 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.

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

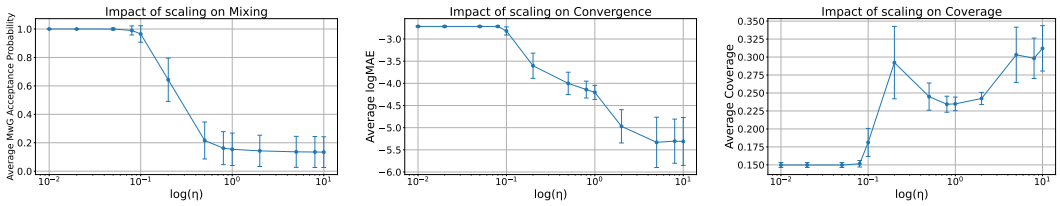

Figure 7: Sensitivity analysis of HiSS under different configurations.

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.

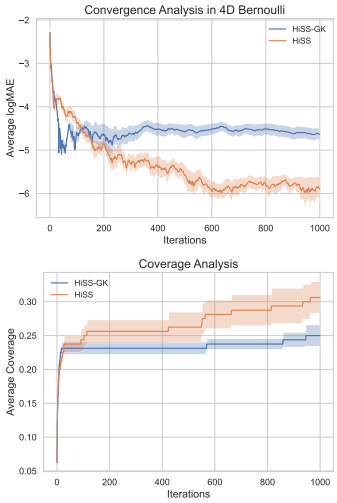

Figure 5: Gaussian Kernel vs Logistic Kernel

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 E) 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 convergence (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.

7 Conclusion

7.1 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.

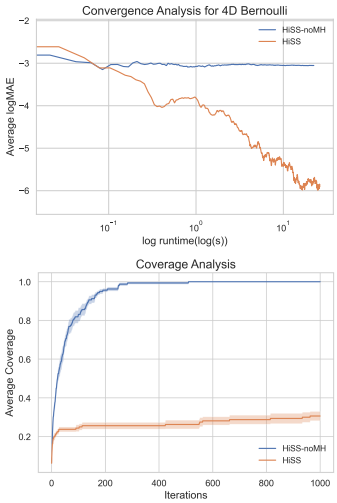

Figure 6: MH vs no-MH

7.2 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.

References

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

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

[3] M. Banterle, C. Grazian, A. Lee, and C. P. Robert. Accelerating metropolis-hastings algorithms by delayed acceptance, 2015.

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

[5] M. Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017.

[6] C. M. Bishop. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[7] J. Bossek. A comprehensive analysis of permutation distance measures in combinatorial optimization. *Evolutionary Computation*, 28(4):589–618, 2020.

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

[9] P. M. Chaikin and T. C. Lubensky. *Principles of Condensed Matter Physics*. Cambridge University Press, 1995.

[10] W. Chen, M. Zhang, B. Paige, J. M. Hernández-Lobato, and D. Barber. Diffusive gibbs sampling, 2024.

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

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

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

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

[15] D. Dua and C. Graff. UCI machine learning repository, 2017.

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

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

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

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

[20] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. Wiley, 1971.

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

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

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

[24] D. Helgemo, N. Wagner, E. Fink, J. Pollack, and S. E. Smith. Evaluating structural diversity in evolutionary algorithms for tsp. In *Genetic and Evolutionary Computation Conference (GECCO)*, pages 612–620. ACM, 2021.

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

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

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

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

[29] G. L. Jones. On the markov chain central limit theorem. 2004.

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

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

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

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

[34] B. Li and R. Zhang. Entropy-MCMC: Sampling from flat basins with ease. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.

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

[36] X. Liu, M. Ye, D. Zhou, and Q. Liu. Post-training quantization with multiple points: Mixed precision without mixed precision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8697–8705, 2021.

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

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

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

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

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

[42] P. Mohanty, R. Bhattacharya, and R. Zhang. Entropy-guided sampling of flat modes in discrete spaces, 2025.

[43] M. M. Morshed and V. Boddeti. Diverseflow: Sample-efficient diverse mode coverage in flows. *arXiv preprint arXiv:2504.07894*, 2025.

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

[45] E. Nummelin. *General Irreducible Markov Chains and Non-negative Operators*. Cambridge University Press, 1984.

[46] E. Pompe, C. Holmes, and K. Łatuszyński. A framework for adaptive mcmc targeting multimodal distributions. 2020.

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

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

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

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

[51] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, New York, 1999.

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

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

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

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

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

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

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

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

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

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

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

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

[64] G. Wang and S. Shan. Mode pursuing sampling method for discrete variable optimization on expensive black-box functions. *Journal of Mechanical Design*, 130(2):021402, 2008.

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

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

[67] R. Xiang, F. Wang, and J. Yang. Efficient sampling of discrete distributions via bridging continuous relaxations. *arXiv preprint arXiv:2305.10217*, 2023.

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

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

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

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

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 [6]. The logistic distribution, on the other hand, exhibits slower tail decay ($e^{-|x|}$) than Gaussian (e^{-x^2}) [20], 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 [26].

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 8, 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 *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:

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

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 9 for intuition).

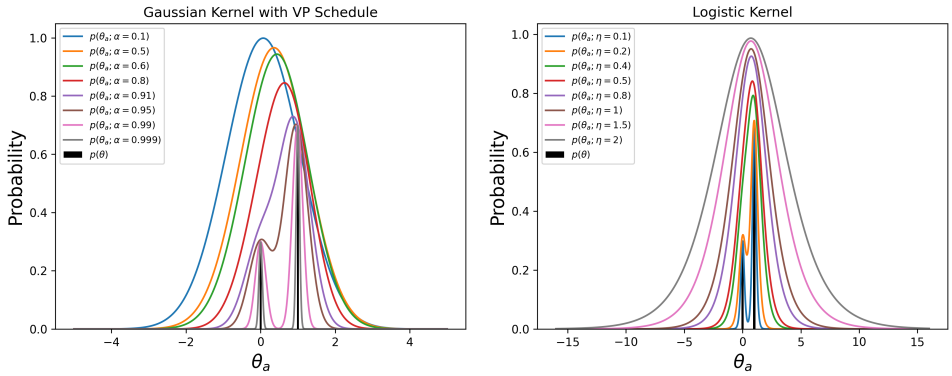

Figure 8: Mode Bridging in Bernoulli Distribution

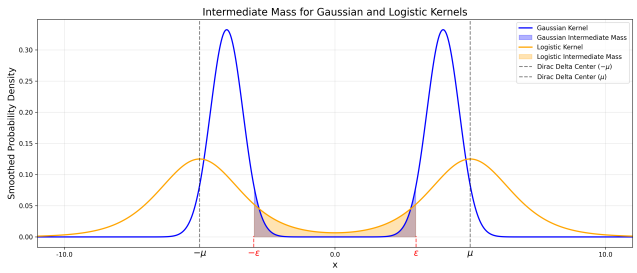

Figure 9: Intermediate Mass for Kernels

Mathematically,

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

Gaussian Kernel

Under the VP schedule inspired by Diffusion Models [54, 55, 25], 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 $x = 0$,

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

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

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

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

Thus,

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

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

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 Key Differences

We compare our approach with Diffusive Gibbs Sampling (DiGS) [10] and Entropic Discrete Langevin Proposal with Gibbs-like Update (EDLP-GLU) [42], highlighting the key differences.

Both EDLP and HiSS employ an innovative coupling mechanism between a discrete primary variable and an auxiliary continuous variable, jointly sampling from them under a unified joint distribution while ultimately retaining only the discrete variable. However, their objectives and underlying principles differ significantly. The contribution of $\boldsymbol{\theta}_a$ brings about this change in both cases: In EDMALA, $\boldsymbol{\theta}_a$ guides $\boldsymbol{\theta}$ under the joint distribution to target and enhance sampling from flat modes, while ensuring detailed balance for the joint distribution. In contrast, 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, making it particularly effective in scenarios with distant or disconnected high-probability regions. Another fundamental distinction lies in their theoretical guarantees. While EDLP's convergence bounds pertain to the joint distribution (Theorems 5.5 and 5.6 in [42]), Theorem 5.5 establishes convergence strictly for the marginal distribution.

HiSS and DiGS enhance Markov chain mixing efficiency through auxiliary structures, but extending DiGS to discrete distributions is not as 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. Since we work in a hybrid setting, noising and denoising steps involve state-switching strategies, designing which are non-trivial and require careful thinking. Additionally, HiSS introduces a logistic convolutional kernel, diverging from DiGS's Gaussian convolutional kernel. This structural shift is detailed further in Section A. HiSS further distinguishes itself through simplicity: unlike DiGS, which relies on an extremely sensitive and complex hyperparameter schedule (the VP schedule) involving multiple noise levels(Section 2.3 and Section 2.4 [10]), HiSS employs a static hyperparameter design. The resulting sampler is not only easier to tune but also offers competitive performance across a wide range of applications. Furthermore, the static hyperparameter choice, makes it easier to provide convergence guarantees(Theorem 5.5) to substantiate the reliability of our proposed approach, something DiGS lacks. Additionally, our empirical analysis is more rigorous and comprehensive: we extend our evaluation to include runtime analysis, accounting for potential overhead introduced by our modifications, as well as mode coverage, which is fundamentally important in discrete settings[43, 64].

By integrating these advanced concepts and mechanisms, HiSS represents a significant advancement in discrete sampling for multimodal spaces.

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

D Proof of Theorems

D.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:

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

□

D.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. $p(\widetilde{\boldsymbol{\theta}}) > 0$[53, 51]. 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[41]. 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[45].

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

□

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

D.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 ith Gibbs sweep, where $i = 1, 2, 3, \cdots, G$, from $\boldsymbol{\theta}^{(i-1)}$ to $\boldsymbol{\theta}^{(i)}$:

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

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

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

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

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

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

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

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

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

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

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

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

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 $\pi(\boldsymbol{\theta})$.[51, 61]. 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 [3].

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 [3, 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 D.1. *Under Assumptions 5.1 and 5.2, and for step size $\alpha < \frac{2}{M}$ in Algorithm 1, for the ith 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.

D.4 Proof of Lemma D.1

Proof. We follow a similar minorization proof style as of Lemma 5.3 from [47]. 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 tth refinement.

We see,

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

Consequently, the modified normalizing constant becomes

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

We see,

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

Also,

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

In other words,

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

Consequently,

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

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

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

Thus,

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

We also note that

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

This is because, from Assumption 5.1 (U is M-gradient Lipschitz), we have

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

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

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

The acceptance ratio is given as:

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

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

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

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

□

D.5 Proof of Theorem 5.5

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

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

In order to lower bound $\kappa_{(\text{HiSS})}(\boldsymbol{\theta}^{(G)} \mid \boldsymbol{\theta}^{(0)})$, we derive a lower bound for $\int p_{\text{MwG}}(\boldsymbol{\theta}_{\text{init}}^{(i)}, \boldsymbol{\theta}_a^{(i-1)} \mid \boldsymbol{\theta}^{(i-1)}) d\boldsymbol{\theta}_a^{(i-1)}$ and use Lemma (D.1) to lower bound $p_{\text{DMALA}}(\boldsymbol{\theta}^{(i)} \mid [{\boldsymbol{\theta}_{\text{init}}^{(i)}}^T, {\boldsymbol{\theta}_a^{(i-1)}}^T]^T)$ 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,

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

We know by definition,

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

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

Similarly,

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

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

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,

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

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

$$\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}_{\text{init}}^{(i)}, \boldsymbol{\theta}_a^{(i-1)} \mid \boldsymbol{\theta}^{(i-1)}) \cdot p_{\text{DMALA}}(\boldsymbol{\theta}^{(i)} \mid [{\boldsymbol{\theta}_{\text{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})\text{diam}^2(\boldsymbol{\Theta}) - (\frac{L}{2}\|\nabla U(a)\| + \frac{3L\sqrt{d}+L}{\eta})\text{diam}(\boldsymbol{\Theta})\right\}\cdot \\
&\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)} \\
&\geq \prod_{i=1}^{G} \left[\nu_i(A) \exp\left\{(-\frac{ML}{2} - \frac{L}{\alpha} + \frac{mL}{4})\text{diam}^2(\boldsymbol{\Theta}) - (\frac{L}{2}\|\nabla U(a)\| + \frac{3L\sqrt{d}+L}{\eta})\text{diam}(\boldsymbol{\Theta})\right\}\cdot \right. \\
&\left. \frac{\exp\left(-M\text{diam}^2(\boldsymbol{\Theta}) - \|\nabla U(a)\|\text{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})\text{diam}^2(\boldsymbol{\Theta}) - (\frac{L}{2}\|\nabla U(a)\| \right.\right. \\
&\left.\left. + \frac{3L\sqrt{d}+L}{\eta})\text{diam}(\boldsymbol{\Theta})\right\} \cdot \exp\left(-M\text{diam}^2(\boldsymbol{\Theta}) - \|\nabla U(a)\|\text{diam}(\boldsymbol{\Theta}) - \frac{2\sqrt{d}}{\eta}\text{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. $A' \subseteq \boldsymbol{\Theta}^{L\times G}$.

□

Theorem D.2. *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 $n \to \infty$.

D.6 Proof of Theorem D.2

Proof. Direct consequence of using Theorem 5.5 and [29][Corollary 5]. □

E Additional Experimental Results

Coverage

Let $\mathbf{\Theta}$ denote the set of all possible discrete states with cardinality $|\mathbf{\Theta}|$, 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.
- $\mathcal{V}(S) \subseteq \mathbf{\Theta}$ represent the subset of states in $\mathbf{\Theta}$ visited at least once by the sampler.

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

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

where:

- $|\mathcal{V}(S)|$ is the number of unique states visited by the sampler.
- $|\mathbf{\Theta}|$ 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., $|\mathcal{V}(S)| = |\mathbf{\Theta}|$.
- **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.

E.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:

The state $\boldsymbol{\theta} = 0000$ has the highest probability ($P_{\mathbf{\Theta}}(0000) = 0.588204$), 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 11, 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.

E.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

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

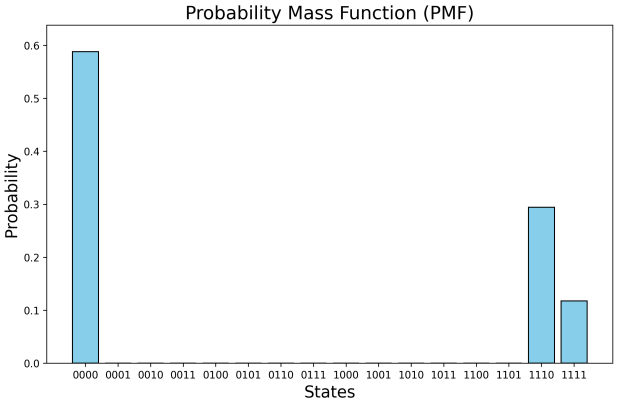

Figure 10: Target Distribution for 4D Joint Bernoulli

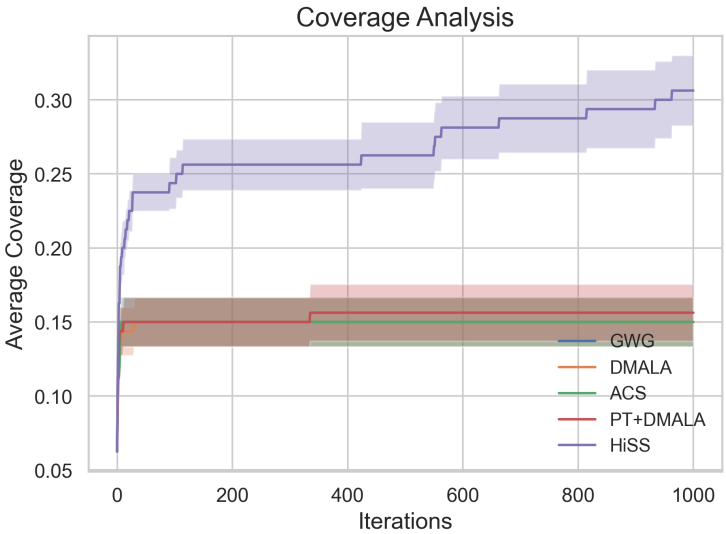

Figure 11: Coverage Analysis for 4D Bernoulli

sparse, breaking the symmetry-like effect. $\mathbf{W}$'s construction is motivated by the study of frustrated and anisotropic systems[18, 9], 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 12, 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.

Coverage Analysis

In Figure 13, 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

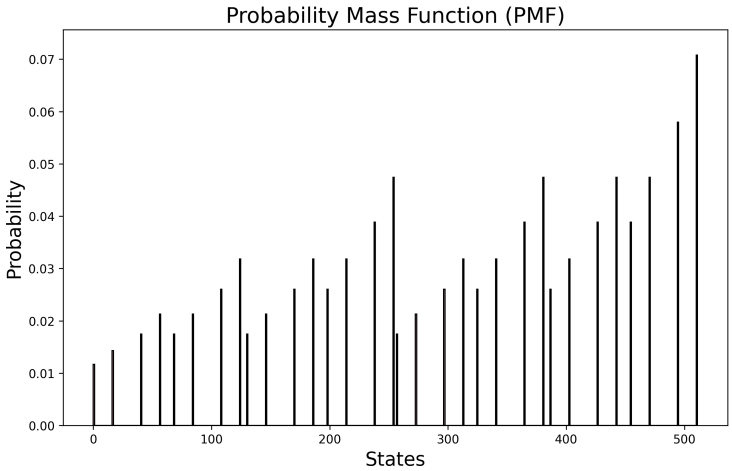

Figure 12: Ising Model Distribution

sufficient mixing. These methods stall early, underscoring their inability to effectively explore the multimodal nature of the Ising model.

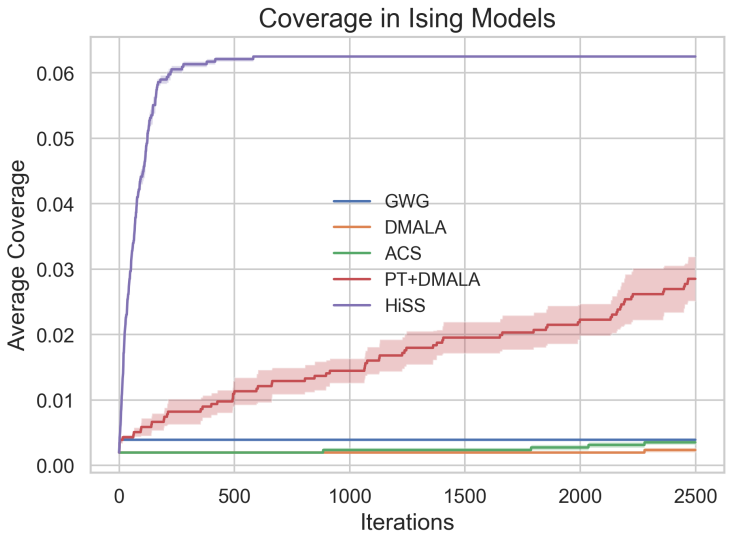

Figure 13: Coverage Analysis for Ising Model

E.2.1 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 3, HiSS does not consistently achieve the best average training metrics compared to other samplers. Specifically, 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 $\eta = 2$), HiSS achieves the lowest training RMSE, though DMALA remains superior in training log-likelihood.

This dataset-specific performance can be partially explained by the selected η values. A smaller η in the Breast Cancer dataset implies less exploration, possibly limiting HiSS's initial fitting capability. In contrast, the larger η in the COMPAS dataset facilitates greater exploration, allowing HiSS to better navigate and fit complex landscapes, thus yielding improved RMSE performance even at the training stage.

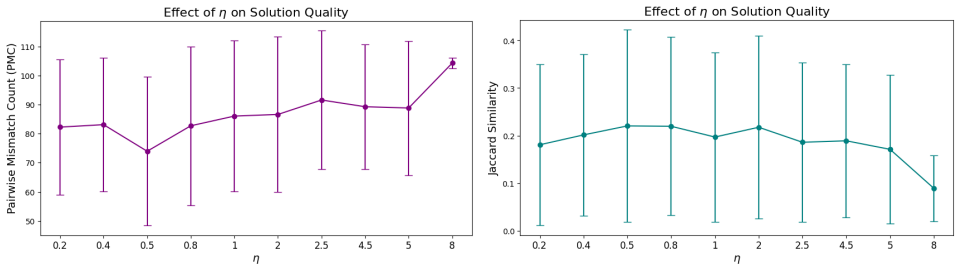

Figure 14: Impact of scale of logistic noise on solution quality.

Interestingly, despite mixed training results, HiSS consistently demonstrates superior performance on test sets (see Section 6). This trend suggests that the samplers outperforming HiSS during training may be prone to overfitting. Conversely, HiSS's comparatively balanced training performance combined with its stronger test performance highlights its robust generalization capability, reflecting its strength in efficient mixing and effective mode discovery in complex posterior distributions.

Table 3: 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

Dataset Details

- **Breast Cancer** [65]: 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** [28]: 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.

E.2.2 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 14, 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.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the introduction, we present four fundamental assertions. Section 4 introduces our discrete sampler. Section 5 delves into the theoretical underpinnings, including the requisite assumptions. Section 6 presents the comprehensive experimental results pertaining to 4D Bernoulli, Ising Model, BBNNs, and TSP.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We do so in Section 7.1.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We do so in Section 5 and Appendix Section D

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We include a lengthy appendix that provides additional results and details all the experimental configuration, along with the hyperparameters used.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will provide a link to an anonymized repository that contains all the code required to execute the necessary experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We include the experimental details in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report the standard error or standard deviation for all the readings.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We do so right at the beginning in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read the Ethics Guidelines, and our submission aligns with all the points listed.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This is a MCMC sampling technique which does not have a direct societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This foundational research that does not directly have a societal impact, as it is primarily an MCMC algorithm for discrete spaces

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: We were unable to locate the license for the datasets we utilized. Nevertheless, these datasets are widely recognized and popular, and we cite the pertinent paper whenever necessary.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This research does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: There are no study participants.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The paper does not describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.