Entropy-Guided Sampling of Flat Modes in Discrete Spaces

Anonymous Author(s)
Affiliation
Address
email

Abstract

Sampling from flat modes in discrete spaces is a crucial yet underexplored problem. Flat modes represent robust solutions and have broad applications in combinatorial optimization and discrete generative modeling. However, existing sampling algorithms often overlook the mode volume and struggle to capture flat modes effectively. To address this limitation, we propose *Entropic Discrete Langevin Proposal* (EDLP), which incorporates local entropy into the sampling process through a continuous auxiliary variable under a joint distribution. The local entropy term guides the discrete sampler toward flat modes with a small overhead. We provide non-asymptotic convergence guarantees for EDLP in locally log-concave discrete distributions. Empirically, our method consistently outperforms traditional approaches across tasks that require sampling from flat basins, including Bernoulli distribution, restricted Boltzmann machines, combinatorial optimization, and binary neural networks.

1 Introduction

Discrete sampling is fundamental to many machine learning tasks, such as graphical models, energy-based models, and combinatorial optimization. Efficient sampling algorithms are crucial for navigating the complex probability landscapes of these tasks. Recent advancements in gradient-based methods have significantly enhanced the efficiency of discrete samplers by leveraging gradient information, setting new benchmarks for tasks such as probabilistic inference and combinatorial optimization (Grathwohl et al., 2021; Zhang et al., 2022; Rhodes & Gutmann, 2022; Sun et al., 2022, 2023; Li & Zhang, 2025).

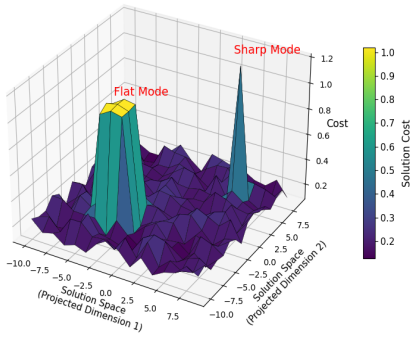


Figure 1: Cost landscape visualization on Traveling Salesman Problem (TSP). Flat modes imply robust solutions under budget, whereas sharp modes are highly sensitive to small changes, leading to abrupt cost increases.

Sampling from flat modes in discrete spaces is a critical yet underexplored challenge. Flat modes, regions where neighboring states have similar probabilities, arise frequently in applications such as energy-based models and neural networks (Hochreiter & Schmidhuber, 1997; Arbel et al., 2021). These regions not only represent mode parameter configurations with high generalization performance (Hochreiter & Schmidhuber, 1997), but they are also important in constrained combinatorial optimization tasks, where finding structurally similar solutions under a budget is required (see Figure 1 for illustration). While there has been growing interest in addressing flat regions in continuous spaces,

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

particularly for tasks like neural network optimization and Bayesian deep learning (Li & Zhang, 2024; Izmailov et al., 2021; Chaudhari et al., 2019), the discrete counterpart remains largely unexplored, highlighting a significant gap.

In this paper, we propose *Entropic Discrete Langevin Proposal* (EDLP), that incorporates the concept of flatness-aware local entropy (Baldassi et al., 2016) into Discrete Langevin Proposal (DLP) (Zhang et al., 2022). By coupling discrete and flat-mode-guided variables, we obtain a broader, entropy-informed joint target distribution that biases sampling towards flat modes. Specifically, while updating the primary discrete variable using DLP, we simultaneously perform continuous Langevin updates on the auxiliary variable. Through the interaction between discrete and auxiliary variables, the discrete sampler will be steered toward flat regions. We summarize our contributions as follows:

- We propose Entropic DLP (EDLP), an entropy-guided, gradient-based proposal for sampling discrete flat modes. EDLP efficiently incorporates local entropy guidance by coupling discrete and continuous variables within a joint distribution.
- We provide non-asymptotic convergence guarantees for EDLP in locally log-concave distributions, offering the first such bound for unadjusted gradient-based discrete sampling.
- Through extensive experiments, we demonstrate that EDLP outperforms existing discrete samplers in capturing flat-mode configurations across various tasks, including Ising models, restricted Boltzmann machines, combinatorial optimization, and binary Bayesian neural networks. We release the code at `https://anonymous.4open.science/r/EDLP-C0E8`.

2 Related Works

Gradient-Based Discrete Sampling. Gradient-based methods have significantly improved sampling efficiency in discrete spaces. Locally informed proposals method by Zanella (2020) leverages probability ratios to explore discrete spaces more effectively. Building on this, Grathwohl et al. (2021) introduced a gradient-based approach to approximate the probability ratio, further improving sampling efficiency. Discrete Langevin Proposal (DLP), introduced by Zhang et al. (2022), adapts the principles of the Langevin algorithm (Grenander & Miller, 1994; Roberts & Tweedie, 1996; Roberts & Rosenthal, 2002), originally designed for continuous spaces, to discrete settings. This algorithm enables parallel updates of multiple coordinates using a single gradient computation, boosting both computational efficiency and scalability.

Flatness-aware Optimization. In early neural network optimization, flatness in energy landscapes emerged as crucial for improving generalization. Hochreiter & Schmidhuber (1994) linked flat minima to better generalization due to their robustness to parameter perturbations. Ritter & Schulten (1988) further emphasized the stability advantages of flat regions. Further, LeCun et al. (1990) linked learning algorithm stability to flatness, suggesting optimization methods to exploit this. Later, Gardner & Derrida (1989) analyzed training algorithms using a statistical mechanics framework, highlighting energy landscape topology's role. In Bayesian deep learning, Li & Zhang (2024) introduced Entropy MCMC (EMCMC) to bias posterior sampling towards flat regions, achieving better generalization of Bayesian neural networks.

Our EDLP differs from existing works by targeting flat modes in discrete distributions. A key algorithmic innovation lies in bridging discrete and continuous spaces. This allows the sampler to explore intermediate regions between discrete states and gain a richer understanding of the discrete landscape, enhancing its ability to sample effectively from flat modes. Further, to our knowledge, we are the first to provide non-asymptotic results for DLP-type algorithms without the MH step, as established in Theorem 5.5, addressing a critical gap in the literature.

3 Preliminaries

Target Distribution. We define a target distribution over a discrete space using an energy function. The target distribution is given by $\pi(\boldsymbol{\theta}) = \frac{1}{Z}\exp(U(\boldsymbol{\theta}))$, where $\boldsymbol{\theta}$ is a d-dimensional discrete variable within domain $\boldsymbol{\Theta}$, $U(\boldsymbol{\theta})$ represents the energy function, and Z is the normalizing constant ensuring $\pi(\boldsymbol{\theta})$ is a proper probability distribution. We make the following assumptions consistent with the literature on gradient-based discrete sampling (Grathwohl et al., 2021; Sun et al., 2022; Zhang et al., 2022): 1. The domain $\boldsymbol{\Theta}$ is factorized coordinatewisely i.e. $\boldsymbol{\Theta} = \Pi_{i=1}^{d}\boldsymbol{\Theta}_i$. 2. 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.

Langevin Algorithm. In continuous spaces, the Langevin algorithm is a powerful sampling method that follows a Langevin diffusion to update variables: $\boldsymbol{\theta}'_{k+1} = \boldsymbol{\theta}_k + \frac{\alpha}{2}\nabla U(\boldsymbol{\theta}_k) + \sqrt{\alpha}\boldsymbol{\epsilon}_k$, where $\boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d\times d})$. The gradient assists the sampler in efficiently exploring high-probability regions.

Discrete Langevin Proposal. The Discrete Langevin Proposal (DLP) is an extension of the Langevin algorithm tailored for discrete spaces, introduced by Zhang et al. (2022). At a given position $\boldsymbol{\theta}$, the proposal distribution $q(\cdot|\boldsymbol{\theta})$ determines the next position. The proposal distribution in DLP 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. DLP can be employed without or with a Metropolis-Hastings (MH) step, resulting in the discrete unadjusted Langevin algorithm (DULA) and the discrete Metropolis-adjusted Langevin algorithm (DMALA), respectively.

Local Entropy. Local entropy is a critical concept in flatness-aware optimization techniques, which is used to understand the geometric characteristics of energy landscapes (Baldassi et al., 2016; Chaudhari et al., 2019; Baldassi et al., 2019). It is defined as:

$$\mathcal{F}(\boldsymbol{\theta}_a; \eta) = \log\left(\sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \exp\left\{U(\boldsymbol{\theta}) - \frac{1}{2\eta}\|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2\right\}\right), \tag{2}$$

where η is a scalar parameter controlling the sensitivity to flatness in the landscape. Local entropy provides a measure of the density of configurations around a point, thus identifying regions with high configuration density and flat energy landscapes.

4 Entropic Discrete Langevin Proposal

4.1 Target Joint Distribution: Coupling Mechanism

We propose leveraging local entropy (Eq.2) to construct an auxiliary distribution that emphasizes flat regions of the target distribution. This auxiliary distribution smoothens the energy landscape, acting as an external force, driving the exploration of flat basins. Figure 4 in the Appendix A illustrates the motivation behind our approach and the impact of the parameter η on the smoothened target distribution.

We start with the original target distribution $p(\boldsymbol{\theta}) \propto \exp(U(\boldsymbol{\theta}))$. By incorporating local entropy, we derive a smoothed target distribution in terms of a new variable $\boldsymbol{\theta}_a$:

$$p(\boldsymbol{\theta}_a) \propto \exp \mathcal{F}(\boldsymbol{\theta}_a; \eta) = \sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \exp\left\{U(\boldsymbol{\theta}) - \frac{1}{2\eta}\|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2\right\} \tag{3}$$

Inspired by the coupling method introduced by Li & Zhang (2024) in their Section 4.1, we couple $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_a$ as follows:

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

$$p(\widetilde{\boldsymbol{\theta}}) = p(\boldsymbol{\theta}, \boldsymbol{\theta}_a) \propto \exp\left\{U(\boldsymbol{\theta}) - \frac{1}{2\eta}\|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2\right\} \tag{4}$$

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).

This result directly follows from Lemma 1 under Section 4.1 in Li & Zhang (2024). 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 $\mathbb{R}^d$. Consequently, the energy function of $\widetilde{\boldsymbol{\theta}}$ becomes $U(\widetilde{\boldsymbol{\theta}}) = U(\boldsymbol{\theta}) - \frac{1}{2\eta}\|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2$, and its gradient is given by:

$$\nabla_{\widetilde{\boldsymbol{\theta}}} U_\eta(\widetilde{\boldsymbol{\theta}}) = \begin{bmatrix} \nabla_{\boldsymbol{\theta}} U_\eta(\widetilde{\boldsymbol{\theta}}) \\ \nabla_{\boldsymbol{\theta}_a} U_\eta(\widetilde{\boldsymbol{\theta}}) \end{bmatrix} = \begin{bmatrix} \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) - \frac{1}{\eta}(\boldsymbol{\theta} - \boldsymbol{\theta}_a) \\ \frac{1}{\eta}(\boldsymbol{\theta} - \boldsymbol{\theta}_a) \end{bmatrix}. \tag{5}$$

4.2 Sampling Algorithm: Local Entropy Guidance in Discrete Langevin Proposals

We propose EDLP, an extension of DLP designed to enhance sampling efficiency from flat modes. In our framework (Algorithm 1), the Langevin update for $\boldsymbol{\theta}_a$ follows the distribution $q_{\alpha_a}(\boldsymbol{\theta}'_a|\widetilde{\boldsymbol{\theta}})$:

$$q_{\alpha_a}(\boldsymbol{\theta}'_a|\widetilde{\boldsymbol{\theta}}) = \frac{1}{\sqrt{2\pi\alpha_a}^d} \exp\left(-\frac{1}{2\alpha_a}\|\boldsymbol{\theta}'_a - \boldsymbol{\theta}_a - \frac{\alpha_a}{2}\nabla_{\boldsymbol{\theta}_a} U_\eta(\widetilde{\boldsymbol{\theta}})\|^2\right). \tag{6}$$

Unlike the standard DLP, where transitions are purely between discrete states, EDLP leverages the current joint variables $\widetilde{\boldsymbol{\theta}} = [\boldsymbol{\theta}^T, \boldsymbol{\theta}_a^T]^T$ to propose the next discrete state. 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:

$$q_\alpha(\boldsymbol{\theta}'|\widetilde{\boldsymbol{\theta}}) \propto \exp\left(-\frac{1}{2\alpha}\|\boldsymbol{\theta}' - \boldsymbol{\theta} - \frac{\alpha}{2}\nabla_{\boldsymbol{\theta}} U_\eta(\widetilde{\boldsymbol{\theta}})\|^2\right). \tag{7}$$

To further simplify, we use coordinate-wise factorization from DLP to obtain $q_\alpha(\boldsymbol{\theta}'|\widetilde{\boldsymbol{\theta}}) = \prod_{i=1}^d q_{\alpha_i}(\theta'_i|\widetilde{\boldsymbol{\theta}})$, where $q_{\alpha_i}(\theta'_i|\widetilde{\boldsymbol{\theta}})$ is a categorical distribution:

$$\text{Cat}\left(\text{Softmax}\left(\frac{1}{2}\nabla_\theta U_\eta(\widetilde{\boldsymbol{\theta}})_i(\theta'_i - \theta_i) - \frac{(\theta'_i - \theta_i)^2}{2\alpha}\right)\right). \tag{8}$$

By synthesizing Equations (6) and (8), we derive the full proposal distribution:

$$q_\gamma(\widetilde{\boldsymbol{\theta}'}|\widetilde{\boldsymbol{\theta}}) \propto q_\alpha(\boldsymbol{\theta}'|\widetilde{\boldsymbol{\theta}}) q_{\alpha_a}(\boldsymbol{\theta}'_a|\widetilde{\boldsymbol{\theta}}) \tag{9}$$

where $\gamma = (\alpha, \alpha_a)$.

This factorized proposal in Eq. (9) is purely a design choice to simplify sampling. The proposal distribution is called the *Entropic Discrete Langevin Proposal* (EDLP). At the current joint position $\widetilde{\boldsymbol{\theta}}$, EDLP generates the next joint position. EDLP can be paired with or without a Metropolis-Hastings step (Metropolis et al., 1953; Hastings, 1970) to ensure the Markov chain's reversibility. These algorithms are referred to as EDULA (Entropic Discrete Unadjusted Langevin Algorithm) and EDMALA (Entropic Discrete Metropolis-Adjusted Langevin Algorithm), respectively. We will collect samples of $\boldsymbol{\theta}$, as the marginal distribution of $p(\widetilde{\boldsymbol{\theta}})$ over $\boldsymbol{\theta}$ yields our desired discrete target distribution.

Alongside the vanilla EDLP, we introduce a computationally efficient *Gibbs-like-update* (GLU) version, in the Appendix B, which involves alternating updates instead of simultaneous updates of our variables. We provide a sensitivity analysis of the hyperparameters in Appendix A.

5 Theoretical Analysis

In this section, we provide a theoretical analysis of the convergence rate of EDLP i.e. EDULA and EDMALA. We make similar assumptions as Pynadath et al. (2024). Those are as follows,

Assumption 5.1. *The function $U(\cdot) \in C^2(\mathbb{R}^d)$ has M-Lipschitz gradient.*

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 $m_{\boldsymbol{\theta}} > 0$.*

Assumption 5.3. *$\boldsymbol{\theta}_a$ is restricted to a compact subset of $\mathbb{R}^d$ labeled $\boldsymbol{\Theta}_a$.*

We define $\text{diam}(\boldsymbol{\Theta}) = \sup_{\boldsymbol{\theta},\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$, and $\text{diam}(\boldsymbol{\Theta}_{\boldsymbol{a}}) = \sup_{\boldsymbol{\theta}_a,\boldsymbol{\theta}'_a \in \boldsymbol{\Theta}_a} \|\boldsymbol{\theta}_a - \boldsymbol{\theta}'_a\|$. Let $\vartheta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{\boldsymbol{a}}) = \inf_{\boldsymbol{\theta},\boldsymbol{\theta}' \in \boldsymbol{\Theta}; \boldsymbol{\theta}_a, \boldsymbol{\theta}'_a \in \boldsymbol{\Theta}_{\boldsymbol{a}}} (\boldsymbol{\theta} - \boldsymbol{\theta}_a)^\top (\boldsymbol{\theta}' - \boldsymbol{\theta}'_a)$ and $\Delta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_a) = \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta},\, \boldsymbol{\theta}_{\boldsymbol{a}} \in \boldsymbol{\Theta}_{\boldsymbol{a}}} \|\boldsymbol{\theta}_{\boldsymbol{a}} - \boldsymbol{\theta}\|$. Let the joint valid bounded space be $\widetilde{\boldsymbol{\Theta}}$ and finally 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, and 5.3 are standard in optimization and sampling literature Bottou et al. (2018); Dalalyan (2017); Durmus & Moulines (2017). Under Assumption 5.2, $U(\cdot)$ is m-strongly concave on conv($\boldsymbol{\Theta}$), following Lemma C.3 from Pynadath et al. (2024). The total variation distance between two probability measures μ and ν, defined on some space $\boldsymbol{\theta} \subset \mathbb{R}^d$ is$\|\mu - \nu\|_{TV} = \sup_{A \subseteq B(\boldsymbol{\theta})} |\mu(A) - \nu(A)|$ where $B(\boldsymbol{\theta})$ is the set of all measurable sets in $\boldsymbol{\theta}$.

Algorithm 1 Entropic Discrete Langevin Proposal: EDULA and EDMALA

Inputs: Main variable $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, Auxiliary variable $\boldsymbol{\theta}_a \in \mathbb{R}^d$, Main stepsize α, Auxiliary stepsize α_a, Flatness parameter η
Initialize: $\boldsymbol{\theta}_a \leftarrow \boldsymbol{\theta}, \mathcal{S} \leftarrow \emptyset$
loop
  Construct $\nabla_{\widetilde{\boldsymbol{\theta}}} U_\eta(\widetilde{\boldsymbol{\theta}})$ as in Equation (5)
  for $i = 1$ **to** d **do**
    Construct $q_{i\alpha}(\cdot|\widetilde{\boldsymbol{\theta}})$ as in Equation (8)
    Sample $\boldsymbol{\theta}_i{}' \sim q_{i\alpha}(\cdot|\widetilde{\boldsymbol{\theta}})$
  end for
  Compute $\boldsymbol{\theta}'_{\boldsymbol{a}} \leftarrow \boldsymbol{\theta}_a + \frac{\alpha_a}{2}\nabla_{\boldsymbol{\theta}_a} U_\eta(\widetilde{\boldsymbol{\theta}}) + \sqrt{\alpha_a}\boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
  ▷ Optionally, do the MH step
  Compute $q_\alpha(\widetilde{\boldsymbol{\theta}'}|\widetilde{\boldsymbol{\theta}}) = \prod_i q_{i\alpha}(\widetilde{\boldsymbol{\theta}'_{\boldsymbol{i}}}|\widetilde{\boldsymbol{\theta}})$
    and $q_\alpha(\widetilde{\boldsymbol{\theta}}|\widetilde{\boldsymbol{\theta}'}) = \prod_i q_{i\alpha}(\widetilde{\boldsymbol{\theta}_{\boldsymbol{i}}}|\widetilde{\boldsymbol{\theta}'})$
  Set $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}'$ and $\boldsymbol{\theta}_a \leftarrow \boldsymbol{\theta}'_{\boldsymbol{a}}$ with probability

$$\min\left(1, \frac{q_\alpha(\boldsymbol{\theta}|\widetilde{\boldsymbol{\theta}'})}{q_\alpha(\boldsymbol{\theta}'|\widetilde{\boldsymbol{\theta}})} \frac{q_{\alpha_a}(\boldsymbol{\theta}_{\boldsymbol{a}}|\widetilde{\boldsymbol{\theta}'})}{q_{\alpha_a}(\boldsymbol{\theta}'_{\boldsymbol{a}}|\widetilde{\boldsymbol{\theta}})} \frac{\pi(\widetilde{\boldsymbol{\theta}'})}{\pi(\widetilde{\boldsymbol{\theta}})}\right)$$

  .
  if *after burn-in* **then**
    Update $\mathcal{S} \leftarrow \mathcal{S} \cup \{\boldsymbol{\theta}\}$
  end if
end loop
Output: $\mathcal{S}$

5.1 Convergence Analysis for EDULA

Since EDULA does not have the target as the stationary distribution, we establish mixing bounds for it in two steps. We first prove that when both the stepsizes (α , α_a) tend to zero, the asymptotic bias of EDULA is zero for target distribution $\tilde{\pi}(\widetilde{\boldsymbol{\theta}}) \propto e^{(U(\boldsymbol{\theta}) - \frac{1}{2\eta}\|\boldsymbol{\theta}-\boldsymbol{\theta}_a\|^2)}$.

Proposition 5.4. *Under Assumptions 5.1, and 5.3, the Markov chain as defined in* (9) *is reversible with respect to some distribution π_γ and π_γ converges weakly to π as $\alpha \to 0$ and $\alpha_a \to 0$. Further, for any $\alpha > 0, \alpha_a > 0$,*

$$\|\pi_\gamma - \tilde{\pi}\|_1 \leq Z \exp\left(\frac{M}{4} - \frac{1}{2\alpha} + \frac{\Delta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{\boldsymbol{a}})^2 - \vartheta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{\boldsymbol{a}})}{2\eta}\right),$$

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

The parameter α_a is consumed during the computation of the stationary distribution π_γ, explicitly not appearing in the bound. However, α_a indirectly influences the geometric terms $\Delta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_a)$ and $\vartheta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_a)$. Larger α_a increases $\Delta^2(\boldsymbol{\Theta}, \boldsymbol{\Theta}_a)$ due to a greater diameter and reduces $\vartheta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_a)$ due to weaker alignment, thereby loosening the bound. In contrast, smaller α_a tightens convergence guarantees. This parallels the observable role of α in the bound i.e. bias vanishes to 0 as $\alpha \to 0$. Next we establish our main result for EDULA which levarages Proposition 5.4 and the ergodicity of the EDULA chain, as a consequence of Lemma D.6 in the Appendix.

Theorem 5.5. *Under Assumptions 5.1, and 5.3 , in Algorithm 1, Markov chain P exhibits,*

$$\|P^k(x, \cdot) - \tilde{\pi}\|_{TV} \leq (1 - \bar{\eta}^*)^k + Z \exp\left(\frac{M}{4} - \frac{1}{2\alpha} + \frac{\Delta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{\boldsymbol{a}})^2 - \vartheta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{\boldsymbol{a}})}{2\eta}\right)$$

where $\bar{\eta}^*$ is a constant that can be explicitly computed (see (18) in the Appendix). In essence, $\bar{\eta}^* = f(\alpha, \alpha_a, \text{diam}(\boldsymbol{\Theta}), \text{diam}(\boldsymbol{\Theta}_a), \Delta(\boldsymbol{\Theta}_a, \boldsymbol{\Theta}))$, where f is increasing exponentially in the first two arguments and decreasing exponentially in the last three arguments. Theorem 5.5 shows that sufficiently small learning rates bring the samples generated by Algorithm 1 closer to the target distribution. However, excessively small rates hinder convergence by limiting exploration, while large rates cause the sampler to overshoot the target. Thus, choosing an appropriate learning rate is critical for balancing exploration and convergence.

5.2 Convergence Analysis for EDMALA

We establish a non-asymptotic convergence guarantee for EDMALA using a uniform minorization argument.

Theorem 5.6. *Under Assumptions 5.1 ,5.2, and 5.3 , and* $\alpha < \frac{2}{M}$ *in Algorithm 1, Markov chain P is uniformly ergodic under,*

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

where, $\epsilon_\gamma = \exp\left\{ -\left(\frac{M}{2} + \frac{1}{\alpha} - \frac{m}{4}\right) diam(\mathbf{\Theta})^2 - \frac{1}{2}\|\nabla U(a)\|\, diam(\mathbf{\Theta}) - \left(\frac{3\alpha_a}{8\eta^2} + \frac{2}{\eta}\right)\Delta(\mathbf{\Theta},\mathbf{\Theta}_a)^2 + \frac{\vartheta(\mathbf{\Theta},\mathbf{\Theta}_a)}{\eta} \right\}$

One notices, ϵ_γ is exponentially decreasing in the size of the set, $\mathbf{\Theta}$, its distance from $\mathbf{\Theta}_a$. Further, as $\alpha \to 0$, $\epsilon_\gamma \to 0$, causing the convergence factor $1-\epsilon_\gamma$ to approach 1. This slows the convergence rate, as the chain takes longer to approach the stationary distribution.

One notices, for $\eta \to \infty$ (weaker coupling), the bounds in Proposition 5.4 and Theorem 5.6 align with those of DULA Zhang et al. (2022) and DMALA (Pynadath et al., 2024), respectively. Note that the convergence of the chains for both EDULA and EDMALA imply convergence of the marginals as the projection maps are continuous. In fact, deriving a rate of convergence for them is also possible, but we omit it here as that is not the goal of this paper.

6 Experiments

We conducted an empirical evaluation of the Entropic Discrete Langevin Proposal (EDLP) to demonstrate its effectiveness in sampling from flat regions compared to existing discrete samplers. Our experimental setups mainly follow Zhang et al. (2022). EDLP is benchmarked against a range of popular baselines, including Gibbs sampling, Gibbs with Gradient (GWG) (Grathwohl et al., 2021), Hamming Ball (HB) (Titsias & Yau, 2017), Discrete Unadjusted Langevin Algorithm (DULA), and Discrete Metropolis-Adjusted Langevin Algorithm (DMALA) (Zhang et al., 2022). For consistency in comparing DLP samplers with their entropic counterparts, we maintain α values across most instances. We retain Zhang et al. (2022)'s notation for consistency: Gibbs-X for Gibbs sampling, GWG-X for Gibbs with Gradient, and HB-X-Y for Hamming Ball. To the best of our knowledge, fBP (Baldassi et al., 2016) is the only algorithm that targets flat regions in discrete spaces. However, it is not directly comparable to EDLP and the other samplers in our study due to methodological and practical reasons (see Appendix C for details).

6.1 Motivational Synthetic Example

We consider sampling from a joint quadrivariate Bernoulli distribution. Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)$ be a 4-dimensional binary random vector, where each $\theta_i \in \{0, 1\}$. 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 $\boldsymbol{\theta}$ then energy function is given by :

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

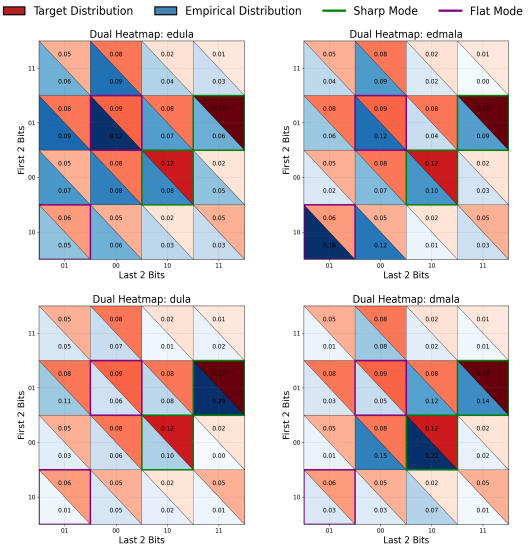


Figure 2: Overlay Heatmaps for EDULA, EDMALA, DULA, and DMALA.

The target distribution over the 4D Joint Bernoulli space contains both sharp and flat modes, each analyzed over their 1-Hamming distance neighborhoods. Sharp modes, such as 0010 and 0111, have high probability mass but are surrounded by neighbors with significantly lower probabilities, indicating steep local gradients. In contrast, flat modes like 0100 and 1001 are characterized by relatively uniform probabilities among their immediate neighbors, reflecting smoother local geometry. For the true target distribution's visualization refer to Figure 10 in Appendix

E.1. We ran 4 chains of DULA, EDULA, DMALA, and EDMALA in parallel for 1000 iterations, with an initial burn of 200. From Figure 2, EDMALA and EDULA demonstrate a strong preference to visit flat modes, without becoming stuck in the high-probability sharp modes. In contrast, DULA and DMALA show a bias toward the sharp modes, showing to be less adept at exploring the flat areas where the probability mass is more evenly distributed. Despite showing flatness bias, entropic samplers still achieve well-matching samples to the target distribution.

6.2 Sampling for Traveling Salesman Problems

In TSP, the objective is to find the shortest route visiting n cities exactly once and returning to the origin, choosing from $n!$ paths. In practical applications, minimal cost and deviation from the optimal route are often essential for operational consistency. For example, in logistics and delivery services, routes that closely follow the optimal sequence improve loading and unloading efficiency and ensure consistent customer experience (Laporte, 2009; Golden et al., 2008). Minimal sensitivity reduces the cognitive load on drivers who rely on established patterns, which is critical in repetitive, high-volume delivery operations Toth & Vigo (2002) Young et al. (2007). Routes with low sensitivity to deviations provide robustness in situations where consistency and predictability are priorities. Thus, sampling from flat modes allows us to propose multiple robust routes that lie within the same cost bracket.

The energy function $U(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents a specific unique route, signifies the weighted sum of the Euclidean distances between consecutive states (cities). In the Traveling Salesman Problem (TSP) and similar optimization problems, $U(\boldsymbol{\theta})$ is designed to capture the total cost of a particular route configuration $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$. The mathematical formulation of $U(\boldsymbol{\theta})$ can be expressed as:

$$U(\boldsymbol{\theta}) = -\left(\sum_{i=1}^{n-1} \left(w_{(\theta_i, \theta_{i+1})} \cdot \|\theta_i - \theta_{i+1}\| \right) + w_{(\theta_n, \theta_1)} \cdot \|\theta_n - \theta_1\| \right),$$

where $w_{(\theta_i, \theta_{i+1})}$ is a directional weight or scaling factor that allows for non-symmetric costs, accounting for the fact that the cost to travel from city θ_i to θ_{i+1} may differ from the reverse direction, and the term $w_{(\theta_n, \theta_1)}$ represents the cost of returning from the last city θ_n back to the starting city θ_1, thereby completing the tour.

The energy function $U(\boldsymbol{\theta})$ quantifies the overall cost associated with a given route, based on the weighted Euclidean distances between consecutive cities. Maximizing $U(\boldsymbol{\theta})$ involves finding the optimal sequence of cities that minimizes the total travel cost. This formulation is particularly useful in real-world applications where different paths may have varying travel costs due to factors like road conditions, transportation constraints, or other contextual variables (Golden et al., 2008; Laporte, 2009).

For our experimental setup, we address the 8-city TSP, where each city is represented as a 3D binary tensor. A valid solution to the TSP ensures that all cities are visited exactly once, and the path returns to the starting city. If a proposed solution violates the uniqueness of city visits, we reject the sample and remain at the current solution.

We employ four samplers: DULA, DMALA, EDULA, and EDMALA, each with a 10,000-iteration run and a 2,000-iteration burn-in period. After the burn-in, we record unique paths and plot their costs (negative of the energy function). Additionally, we identify the best path for each sampler amongst all unique solutions . Consequently, we calculate the average pairwise mismatch count (PMC) of the best path to all other sampled paths (see Figure 3), which quantifies how distinct the explored solutions are from the optimal path (Schiavinotto & Stützle, 2007; Merz & Freisleben, 1997).

Left: EDULA and EDMALA, show clear superiority over their counterparts, DULA and DMALA, by achieving lower variance cost-spreads. This highlights the less variability in their sampling, demonstrating their superiority in efficiently finding consistent, robust solutions for the TSP.

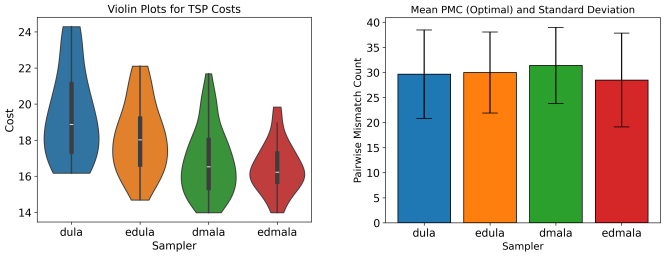


Figure 3: Performance of various samplers on TSP.

Right: To examine the potential variability from the optimal solution, we focus on the upper confidence band, represented as the mean discrepancy plus its standard deviation. While DULA and EDULA have similar upper bounds, EDMALA has a lower upper bound compared to DMALA. We provide additional results in the Appendix E.2.

6.3 Sampling From Restricted Boltzmann Machines

Restricted Boltzmann Machines (RBMs) are a class of generative stochastic neural networks that learn a probability distribution over their input data. The energy function for an RBM, which defines the joint configuration of visible and hidden units, is given by:

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

where $\{\mathbf{W}, a, b\}$ are the weight matrix and bias parameters, respectively, and $\boldsymbol{\theta} \in \{0, 1\}^d$ represents the binary state of the visible units.

When the RBM assigns high probability to specific digit representations, a sharp mode for digit 3 (for instance) might appear as an idealized version without extraneous strokes. This configuration represents the model's interpretation of a quintessential '3' with a prominent probability peak. Any minor alteration, like flipping a single pixel, lowers the altered image's probability. The sampler has thus learned to prioritize exact, pristine versions of each digit, marking any deviation from this high-probability state as unlikely.

For MNIST, this narrow focus limits flexibility. The model assigns high probability to only a few "perfect" digit versions, treating minor variations as less probable. This rigidity makes the generated images sensitive to small changes and limits the RBM's ability to recognize natural, varied handwriting. In the context of RBMs, sampling from flat modes explores a wider range of latent handwritten styles, enhancing the model's ability to capture the underlying data distribution. This reflects a broader representation of possible input variations, crucial for tasks like image generation and data reconstruction Murray et al. (2009). In practice, this means that images generated from flat modes in RBMs are less likely to overfit to sharp, specific patterns in the training data and are instead more reflective of the variability inherent in the dataset.

In our experiments, we generated 5000 images per sampler for the MNIST dataset, applying a thinning factor of 1000 to ensure diversity in the samples. A simple convolutional autoencoder (CAE) was used for image generation and reconstruction, allowing us to evaluate the performance and generalization capability of sampler-generated data. To assess robustness, we trained 5 CAEs on the sampler-generated images and tested them under various conditions. Initially, clean test data was used to establish baseline performance. Subsequently, we introduced Gaussian noise (with a noise factor of 0.1) to evaluate the models' resilience against perturbations, a common method for assessing adversarial robustness (Madry et al., 2018). Additionally, we examined the models with occluded images, where random sections of the images were obscured by zero-valued pixel blocks. This test simulates scenarios with missing or obstructed information, a widely used technique in robustness studies to measure model performance under partial information loss (Zhang et al., 2019).

For quantitative evaluation, we employed several widely accepted metrics: Mean Reconstruction Squared Error (MSE) to measure pixel-level differences between original and reconstructed images, Peak Signal Noise Ratio (PSNR) to measure the fidelity of the reconstructed images, and the Structural Similarity Index (SSIM) to assess the structural integrity of the reconstructions (Wang et al., 2004). Additionally, we computed the log-likelihood to quantify how well the reconstructed images fit the underlying data distribution. These metrics collectively offer a comprehensive assessment of the performance and robustness of the models across clean, noisy, and occluded data.

The results in Table 1 indicate that EDLP methods consistently outperform their non-entropic counterparts across all test settings. Specifically, EDMALA achieves the lowest MSE, highest PSNR, highest SSIM (except for Noisy), and the best log-likelihood values among the samplers tested. These metrics together suggest that EDLP has superior generalization capabilities, making it especially effective for reconstructing unseen data accurately. We provide additional results in the Appendix E.3.

Table 1: Results of different samplers on MNIST under clean, noisy, and occluded conditions.

Sampler	Setting	MSE(↓)	PSNR(↑)	SSIM(↑)	Log-Likelihood(↑)
HB-10-1	Clean	0.0253 ± 0.0005	16.3555 ± 0.0858	0.5303 ± 0.0014	-0.0134 ± 0.0009
	Noisy	0.0267 ± 0.0004	15.9763 ± 0.0697	0.3941 ± 0.0035	0.0165 ± 0.0011
	Occluded	0.0256 ± 0.0004	16.2720 ± 0.0749	0.4963 ± 0.0017	-0.0154 ± 0.0008
BG-1	Clean	0.0257 ± 0.0007	16.2492 ± 0.1125	0.5294 ± 0.0025	-0.0157 ± 0.0014
	Noisy	0.0270 ± 0.0006	15.9086 ± 0.0885	0.3938 ± 0.0038	0.0144 ± 0.0013
	Occluded	0.0260 ± 0.0006	16.1613 ± 0.0992	0.4947 ± 0.0024	-0.0179 ± 0.0013
DULA	Clean	0.0268 ± 0.0006	16.1160 ± 0.1022	0.5114 ± 0.0030	-0.0209 ± 0.0015
	Noisy	0.0280 ± 0.0005	15.7851 ± 0.0815	0.3907 ± 0.0041	0.0097 ± 0.0013
	Occluded	0.0272 ± 0.0006	16.0187 ± 0.0922	0.4766 ± 0.0028	-0.0233 ± 0.0014
DMALA	Clean	0.0256 ± 0.0004	16.3305 ± 0.0709	0.5291 ± 0.0035	-0.0156 ± 0.0011
	Noisy	0.0270 ± 0.0004	15.9547 ± 0.0623	0.3939 ± 0.0032	0.0148 ± 0.0009
	Occluded	0.0259 ± 0.0004	16.2372 ± 0.0632	0.4950 ± 0.0035	-0.0182 ± 0.0010
EDULA	Clean	0.0264 ± 0.0005	16.2135 ± 0.0877	0.5083 ± 0.0052	-0.0179 ± 0.0014
	Noisy	0.0276 ± 0.0004	15.8700 ± 0.0652	**0.3968** ± 0.0030	0.0121 ± 0.0012
	Occluded	0.0268 ± 0.0005	16.1115 ± 0.0797	0.4743 ± 0.0051	-0.0206 ± 0.0014
EDMALA	Clean	**0.0251** ± 0.0005	**16.3974** ± 0.0975	**0.5368** ± 0.0016	**-0.0117** ± 0.0009
	Noisy	**0.0266** ± 0.0004	**15.9938** ± 0.0727	0.3933 ± 0.0029	**0.0177** ± 0.0012
	Occluded	**0.0255** ± 0.0005	**16.3022** ± 0.0839	**0.5019** ± 0.0017	**-0.0141** ± 0.0007

Table 2: Average Test RMSE for various datasets.

Dataset	Gibbs	GWG	DULA	DMALA	EDULA	EDMALA
COMPAS	**0.4752** ±0.0058	0.4756 ±0.0056	0.4789 ±0.0039	0.4773 ±0.0036	0.4778 ±0.0037	0.4768 ±0.0033
News	0.1008 ±0.0011	0.0996 ±0.0027	0.0923 ±0.0037	0.0916 ±0.0040	0.0918 ±0.0036	**0.0915** ±0.0036
Adult	0.4784 ±0.0151	0.4432 ±0.0255	0.3895 ±0.0102	0.3872 ±0.0107	0.3889 ±0.0097	**0.3861** ±0.0110
Blog	0.4442 ±0.0107	0.3728 ±0.0093	0.3236 ±0.0114	0.3213 ±0.0117	0.3218 ±0.0119	**0.3211** ±0.0145

6.4 Binary Bayesian Neural Networks

In alignment with the findings of Li & Zhang (Section 6.3), which highlight the role of flat modes in enhancing generalization in deep neural networks, we explore the training of binary Bayesian neural networks using discrete sampling techniques, leveraging the ability of flat modes to facilitate better generalization. Our experimental design involves regression tasks on four UCI datasets Dua & Graff (2017), with the energy function for each dataset defined as follows:

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

where $D = \{x_i, y_i\}_{i=1}^N$ is the training dataset, and $f_{\boldsymbol{\theta}}$ denotes a two-layer neural network with `Tanh` activation and 500 hidden neurons. Following the experimental setup in Zhang et al. (2022), we report the average test RMSE and its standard deviation. As shown in Table 2, EDMALA and EDULA consistently outperform their non-entropic variants across all datasets, but don't outperform GWG-1 on test RMSE on the COMPAS dataset. This exception can be attributed to overfitting, aligning with prior work Zhang et al. (2022). Overall, these results confirm that our method enhances generalization performance on unseen test data. We provide additional results and hyperparameter settings in the Appendix E.4.

7 Discussion

7.1 Limitations

Since EDLP collects only discrete samples, it produces half as many samples per iteration as EMCMC. The coupling mechanism in Section 4.1 increases the computational load relative to DLP. However, as Li & Zhang states in their Section 4.2, the cost of gradient computation remains the same for d-dimensional models when $\widetilde{\boldsymbol{\theta}}$ resides in a $2d$ dimensional space. EDLP doubles memory usage compared to DLP, but the space complexity remains linear in d, ensuring scalability.

7.2 Conclusion

We propose a simple and computationally efficient gradient-based sampler designed for sampling from flat modes in discrete spaces. The algorithm leverages a guiding variable based on local entropy. We provide non-asymptotic convergence guarantees for both the unadjusted and Metropolis-adjusted versions. Empirical results demonstrate the effectiveness of our method across a variety of applications. We hope our framework highlights the importance of flat-mode sampling in discrete systems, with broad utility across scientific and machine learning domains.

References

Arbel, M., Zhou, L., and Gretton, A. Generalized energy based models. In *International Conference on Learning Representations*, 2021.

Baldassi, C., Borgs, C., Chayes, J. T., Ingrosso, A., Lucibello, C., Saglietti, L., and Zecchina, R. Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes. *Proceedings of the National Academy of Sciences*, 113 (48):E7655–E7662, 2016.

Baldassi, C., Pittorino, F., and Zecchina, R. Shaping the learning landscape in neural networks around wide flat minima. *Proceedings of the National Academy of Sciences*, 117(1):161–170, 2019.

Besag, J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974.

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

Camm, J. D. and Evans, J. R. Constrained optimization models: An illustrative example. *Interfaces*, 27(3):117–127, 1997.

Casella, G. and George, E. I. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

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

Diebolt, J. and Robert, C. P. Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(2):363–375, 1994.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

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

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pp. 265–284. Springer, 2006.

Ekvall, K. O. and Jones, G. L. Convergence analysis of a collapsed Gibbs sampler for Bayesian vector autoregressions. *Electronic Journal of Statistics*, 15(1):691 – 721, 2021. doi: 10.1214/21-EJS1800. URL `https://doi.org/10.1214/21-EJS1800`.

Gardner, E. and Derrida, B. Training and generalization in neural networks. *Journal of Physics A: Mathematical and General*, 22(12):1983, 1989.

Ghosh, A., Roughgarden, T., and Sundararajan, M. Universally optimal privacy mechanisms for minimax agents. *arXiv preprint arXiv:1207.1240*, 2012.

Golden, B., Raghavan, S., and Wasil, E. *The vehicle routing problem: Latest advances and new challenges*. Springer Science & Business Media, 2008.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

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

Grenander, U. and Miller, M. I. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(4):549–581, 1994.

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

Hochreiter, S. and Schmidhuber, J. Simplifying neural nets by discovering flat minima. *Advances in neural information processing systems*, 7, 1994.

Hochreiter, S. and Schmidhuber, J. Flat minima. *Neural computation*, 9(1):1–42, 1997.

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

Jones, G. L. On the Markov chain central limit theorem. *Probability Surveys*, 1(none): 299 – 320, 2004. doi: 10.1214/154957804100000051. URL https://doi.org/10.1214/154957804100000051.

Laporte, G. Fifty years of vehicle routing. *Transportation Science*, 43(4):408–416, 2009.

LeCun, Y., Denker, J. S., and Solla, S. A. Optimal brain damage. In *Advances in neural information processing systems*, pp. 598–605, 1990.

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

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

Li, M. and Zhang, R. Reheated gradient-based discrete sampling for combinatorial optimization. *Transactions on Machine Learning Research*, 2025.

Liang, J. and Chen, Y. A proximal algorithm for sampling. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=CkXOwlhf27.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

Merz, P. and Freisleben, B. Genetic algorithms for the traveling salesman problem. In *Proceedings of the International Conference on Genetic Algorithms (ICGA)*, pp. 321–328. Morgan Kaufmann, 1997. URL https://dl.acm.org/doi/10.5555/285619.285682.

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

Murray, I., Salakhutdinov, R., and Hinton, G. Evaluating rbm approximations: Contrastive divergence vs. alternative approaches. *Neural Computation*, 2009.

Neal, R. M. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

Pereyra, M. Proximal markov chain monte carlo algorithms. *Statistics and Computing*, 26:745–760, 2016.

Pynadath, P., Bhattacharya, R., HARIHARAN, A. N., and Zhang, R. Gradient-based discrete sampling with automatic cyclical scheduling. In *ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2024. URL https://openreview.net/forum?id=aTDId2TrtL.

Rhodes, B. and Gutmann, M. U. Enhanced gradient-based MCMC in discrete spaces. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

Ritter, H. and Schulten, K. Flat minima. *Journal of Physics A: Mathematical and Theoretical*, 21 (10):L745–L749, 1988.

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

Roberts, G. O. and Tweedie, R. L. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pp. 341–363, 1996.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. In *Advances in neural information processing systems*, volume 29, pp. 2234–2242, 2016.

Schiavinotto, T. and Stützle, T. A review of metrics on permutations for search landscape analysis. *Computers & Operations Research*, 34(10):3143–3153, 2007. doi: 10.1016/j.cor.2005.11.023.

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

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

Sun, Y., Wang, Z., Liu, X., and Fan, J. When smart devices collaborate: Context-aware inference in smart homes with edge computing. *IEEE Internet of Things Journal*, 2017.

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

Toth, P. and Vigo, D. The vehicle routing problem. *Society for Industrial and Applied Mathematics*, 2002.

Tsantekidis, A., Passalis, N., Tefas, A., Kanniainen, J., Gabbouj, M., and Iosifidis, A. Using deep learning to forecast stock prices from the limit order book. *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: From error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Young, K., Regan, M. A., and Hammer, M. Driver distraction: A review of the literature. *Accident Analysis & Prevention*, 39(3):562–570, 2007.

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

Zhang, H., Yu, Y., Jiao, J., Xing, E., Ghaoui, L. E., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019.

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

A Analysis of the Effect of Flatness Parameter η

A.1 Intuition

Figure 4 illustrates the effect of varying the flatness parameter η on the probability distribution $p(\boldsymbol{\theta}_a)$ for $\boldsymbol{\theta}$ drawn from a Bernoulli(0.5) distribution. The *layered* curves represent different values of η, showing how the distribution $p(\boldsymbol{\theta}_a)$ changes as η increases.

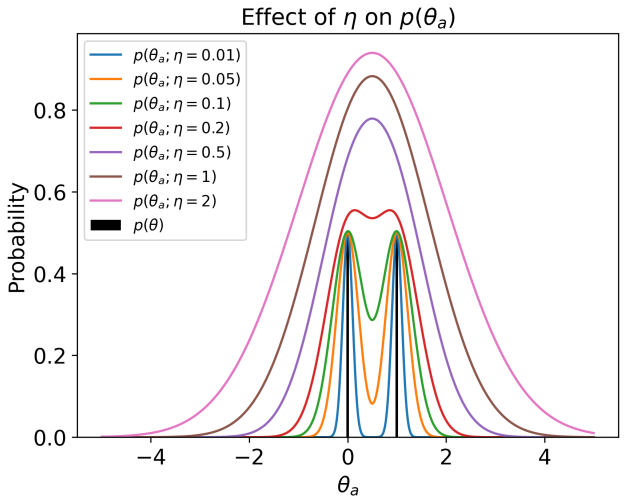


Figure 4: $p(\boldsymbol{\theta_a})$ for $\boldsymbol{\theta} \sim \textit{Bernoulli}(0.5)$

Effect of Small η (Strong Coupling)

For very small values of η (e.g., $\eta = 0.01$, $\eta = 0.05$, $\eta = 0.1$), the curves (blue, orange, and green) are sharply peaked and closely resemble the original $p(\boldsymbol{\theta})$. Small η values imply strong coupling between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_a$. The auxiliary distribution $p(\boldsymbol{\theta}_a)$ remains very close to $p(\boldsymbol{\theta})$, indicating that $\boldsymbol{\theta}_a$ is tightly bound to $\boldsymbol{\theta}$, and the variance is minimal.

Moderate η Values (Moderate Coupling)

As η increases (e.g., $\eta = 0.2$), the curves (red) become wider and smoother. These moderate η values adequately capture the flatness of the landscape. The distribution $p(\boldsymbol{\theta}_a)$ starts to diverge from $p(\boldsymbol{\theta})$, allowing $\boldsymbol{\theta}_a$ to explore a broader region around the peaks.

Large η (Weak Coupling)

For larger values of η (e.g., $\eta = 0.5$, $\eta = 1$, $\eta = 2$), the curves (purple, brown, and magenta) are much wider. Large η values imply weak coupling between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_a$. The auxiliary distribution $p(\boldsymbol{\theta}_a)$ is excessively smoothed out compared to $p(\boldsymbol{\theta})$, indicating that $\boldsymbol{\theta}_a$ can explore a much broader range of values with less influence from $\boldsymbol{\theta}$.

Considerations for η Approaching Infinity

As η approaches infinity, the auxiliary distribution $p(\boldsymbol{\theta}_a)$ flattens, and the gradient $\nabla_{\boldsymbol{\theta}_a} U_\eta(\widetilde{\boldsymbol{\theta}})$ tends toward zero. This results in an extremely weak coupling, effectively causing the EDLP framework to behave similarly to a standard DLP. The parameter η thus plays a critical role in determining the behavior of the sampler, necessitating careful tuning based on the specific requirements of the sampling task.

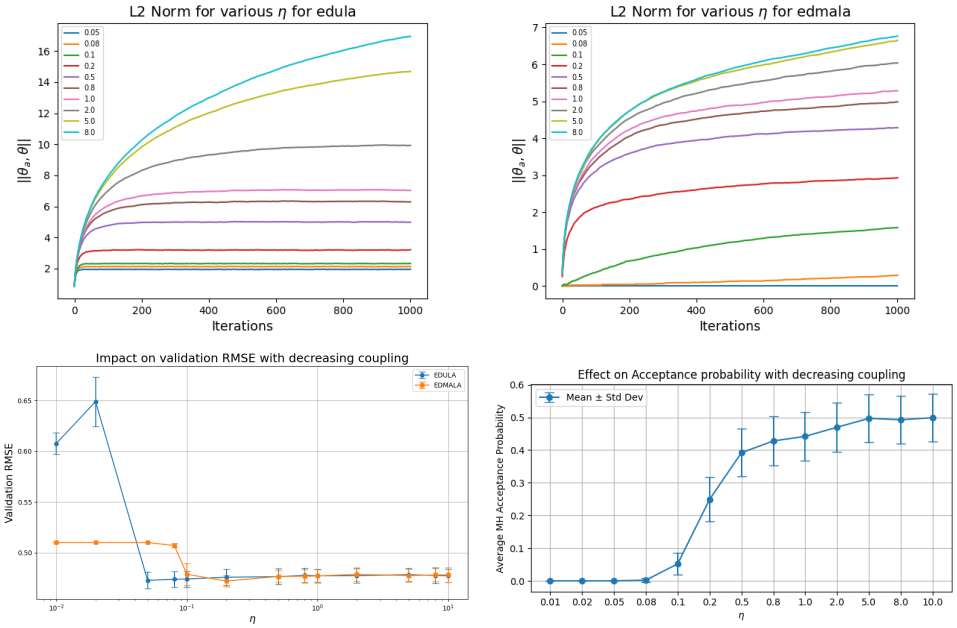


Figure 5: Diagnostics for EDLP

A.2 Sensitivity Analysis

The flatness parameter η is arguably the most crucial hyperparameter to optimize in the EDLP algorithm (Algorithm 1). Similar to the hyperparameter tuning ablation strategies employed in Li & Zhang (2024) (Appendix, Section E), we conduct hyperparameter tuning on the COMPAS dataset's validation data. Specifically, we monitor the L2 norm between sampled pairs of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_a$ for various values of η. Additionally, we plot the validation RMSE for both EDULA and EDMALA across different values of η. Finally, we plot the average MH acceptance ratio for EDMALA to assess the impact of η on the joint MH acceptance step. We maintain $\alpha = 0.1$ for both samplers and $\alpha_a = 0.01$ for EDULA and $\alpha_a = 0.001$ for EDMALA(see Figure 5).

We observe that as η increases, the coupling between the variables weakens, allowing both variables to move more freely, thus increasing the norm. This behavior is consistent across both EDULA and EDMALA. However, EDMALA exhibits a more conservative behavior at the same coupling strength compared to EDULA due to the presence of the joint Metropolis-Hastings (MH) acceptance step, which imposes stricter alignment between the variables, hence maintaining a tighter coupling.

Both samplers demonstrate robustness across a wide range of η, with relatively stable validation RMSE performance. However, EDULA shows slightly less robustness, particularly at extremely small coupling values, resulting in increased variability and higher RMSE. EDMALA maintains a stable, consistent performance, indicating better robustness to changes in the coupling parameter.

The final plot shows how the MH acceptance probability varies with coupling strength η for EDMALA. Initially, with very tight coupling , the acceptance probability is near zero, indicating overly restricted movements due to the strong alignment requirement between the discrete and continuous variables. As η increases (coupling relaxes), the acceptance probability rises significantly, reflecting greater freedom in proposing moves that the joint MH criterion accepts. After a certain coupling threshold (around 0.8 here), the acceptance rate plateaus, suggesting diminishing returns from further relaxation in coupling strength. Thus, an intermediate coupling provides a balance, allowing effective exploration without overly compromising the sampler's consistency.

B Gibbs-like Update Procedure

Gibbs-like updating procedures have been widely employed across various contexts in the sampling literature, particularly within Bayesian hierarchical models, latent variable models, and non-parametric

Bayesian approaches. For instance, Gibbs sampling is a fundamental technique in hierarchical Bayesian models, where parameters are partitioned into blocks and updated conditionally on others to facilitate efficient sampling (Casella & George, 1992). In latent variable models, such as Hidden Markov Models (HMMs) and mixture models, Gibbs-like updates allow for alternating between sampling latent variables and model parameters, thereby simplifying the overall process (Diebolt & Robert, 1994). Additionally, these updates are crucial in non-parametric Bayesian approaches, such as Dirichlet Process Mixture Models (DPMMs), where they enable the efficient sampling of cluster assignments and hyperparameters (Neal, 2000). Gibbs-like updates are also prominently used in spatial statistics, particularly in Conditional Autoregressive (CAR) models, where the value at each spatial location is updated based on its neighbors (Besag, 1974).

Since our goal is to sample from a joint distribution, rather than simultaneously updating $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_a$, we alternatively update these variables iteratively. The conditional distribution for the primary variable $\boldsymbol{\theta}$ is given by:

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

where $Z_{\boldsymbol{\theta}_a} = \exp \mathcal{F}(\boldsymbol{\theta}_a; \eta)$ serves as the normalization constant. Correspondingly, the conditional distribution for the auxiliary variable $\boldsymbol{\theta}_a$ is:

$$p(\boldsymbol{\theta}_a|\boldsymbol{\theta}) \propto \frac{1}{Z_{\boldsymbol{\theta}}} \exp\left\{-\frac{1}{2\eta}\|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2\right\},$$

where $Z_{\boldsymbol{\theta}} = \exp\left(U(\boldsymbol{\theta})\right)$ is the associated normalization constant. This formulation reveals that $\boldsymbol{\theta}_a$ is sampled from $\mathcal{N}(\boldsymbol{\theta}, \eta \boldsymbol{I})$, with the variance η controlling the expected distance between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_a$. During the Metropolis-Hastings (MH) step, the acceptance probability is now calculated as:

$$\min\left(1, \frac{q_\alpha(\boldsymbol{\theta}|\widetilde{\boldsymbol{\theta}'})}{q_\alpha(\boldsymbol{\theta}'|\widetilde{\boldsymbol{\theta}})} \frac{\pi(\widetilde{\boldsymbol{\theta}'})}{\pi(\widetilde{\boldsymbol{\theta}})}\right).$$

This Gibbs-like alternating update scheme offers distinct advantages: (1) exact sampling of $\boldsymbol{\theta}_a$, (2) elimination of the need for the α_a parameter, (3) a less intensive computation of the MH acceptance probability, and (4) reduced overall computational overhead, especially when the proposal step involves an MH correction. This gibbs-like updating also shares similarities with the proximal sampling methods (Pereyra, 2016; Liang & Chen, 2023). This innovation can potentially allow DLP to generalize effectively to more complex, high-dimensional, and non-differentiable discrete target distributions such as the discrete Laplace distribution, which is commonly used in privacy-preserving mechanisms(Dwork et al., 2006; Ghosh et al., 2012). We leave out the theoretical analysis of the GLU versions for future work.

C Considerations for Excluding Focussed Belief Propogation from Benchmarking

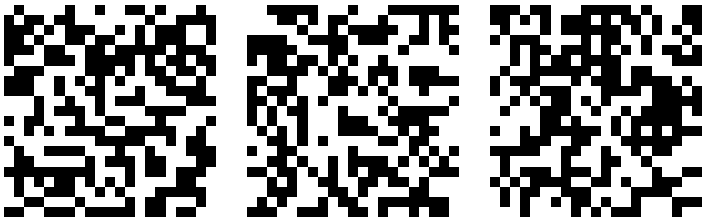
1. Fundamental Differences in Sampling Mechanism: Most of the sampling algorithms we use generate samples sequentially, with each sample x_{t+1} derived from the previous sample x_t. This sequential dependency is essential for building a Markov Chain that explores the distribution space and gradually converges to the target distribution. fBP produces samples sequentially, but instead employs a *message-passing algorithm* aimed at converging to a fixed solution or configuration. It operates to converge deterministically to a solution, rather than generating a sequence of probabilistic samples. Moreover, fBP lacks a formal proof of convergence, relying instead on heuristic principles rooted in replica theory. This absence of theoretical guarantees or established convergence rates means that even if fBP appears to perform well, we cannot interpret or quantify its reliability, efficiency, or consistency across varying datasets and tasks. In contrast, MCMC-based methods like Langevin dynamics and Gibbs sampling come with well-understood convergence properties, enabling meaningful performance evaluations and robust benchmarking. This interpretability gap makes fBP less suitable for our study, where theoretical soundness and predictable behavior are critical.

2. Technical and Practical Constraints with using fBP: While fBP is originally implemented in Julia[1], a Python wrapper[2] is also available. However, this wrapper still depends on the underlying Julia or C++ implementations, introducing potential cross-language communication overhead. This dependency complicates integration in Python workflows and creates an inherent performance disparity when compared to purely Pythonic implementations, making direct runtime comparisons less meaningful. Despite fBP's speed advantage, its execution becomes slow as sample dimensions increase and network ensembles grow larger. The volume of message-passing in high-dimensional contexts limits its scalability. As task complexity increases, fBP faces challenges in achieving stable convergence, further limiting its suitability for our high-dimensional setup. Past studies have excluded computationally expensive methods from experimental evaluations Zhang et al. (2022).

3. Computational Overhead and Efficiency Concerns Resource Demands for Multiple Runs: If we were to use fBP to generate multiple samples, we would need to reinitialize and re-run the algorithm for each sample with a new seed, effectively solving the problem from scratch each time. This is highly inefficient compared to MCMC methods, where each subsequent sample builds on the previous one without needing to restart the entire algorithm. For larger models and datasets, this repeated initialization and execution would result in a significant computational burden.

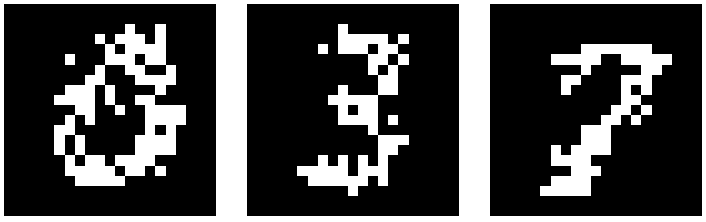
4. Nature of Tasks: In certain structured sampling tasks, such as the TSP, we enforce constraints to ensure that each proposed state is a valid TSP solution. This entails accepting only those configurations that satisfy specific requirements of the TSP. However, fBP does not adhere to such constraints, as it lacks mechanisms for directly enforcing the validity of the sampled states. Consequently, fBP is unsuitable for tasks where such structural constraints are critical, placing it outside the scope for comparison in these applications.

We conducted preliminary experiments using fBP for Restricted Boltzmann Machine (RBM) sampling on the MNIST dataset to assess its effectiveness in image generation. Figure 6 shows random image samples generated by fBP on MNIST, which resemble random unstructured noise rather than recognizable digits, compared to MNIST samples by DMALA and EDMALA in Figures 7, 8 respectively. These outputs suggest that fBP doesn't capture the underlying structure of the MNIST data.



(a) (b) (c)

Figure 6: Random Image Samples for MNIST using fBP

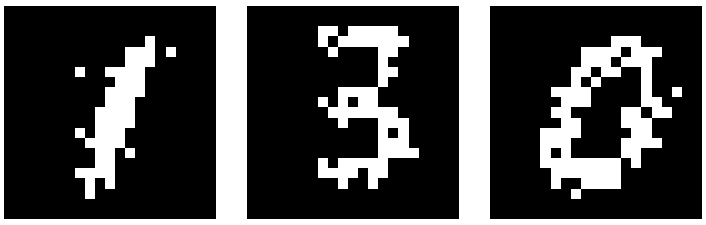


(a) (b) (c)

Figure 7: Random Image Samples for MNIST using DMALA

[1]Carlo Baldassi, *BinaryCommitteeMachinefBP.jl*, GitHub repository, `https://github.com/carlobaldassi/BinaryCommitteeMachinefBP.jl`, accessed November 8, 2024.

[2]Curti, Nico and Dall'Olio, Daniele and Giampieri, Enrico, *ReplicatedFocusingBeliefPropagation*, GitHub repository, `https://github.com/Nico-Curti/rFBP`, accessed November 8, 2024.



(a) (b) (c)

Figure 8: Random Image Samples for MNIST using EDMALA

fBP lacks direct use of the energy function $U(.)$ during optimization, preventing accurate data modeling. Figure 9 illustrates this through a distribution analysis of generated MNIST classes, showing significant mode collapse. Most generated samples cluster around a few classes, with an imbalance favoring certain digits and ignoring others.

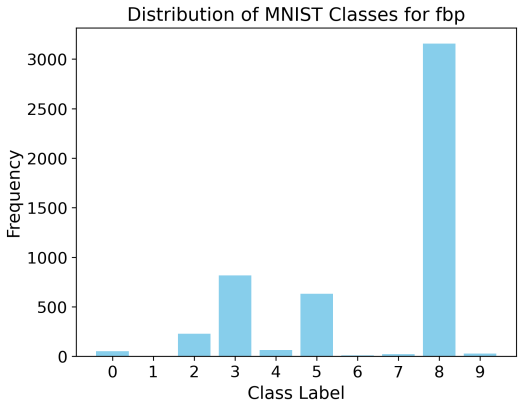


Figure 9: Mode Collapse using fBP

These findings highlight a fundamental issue with fBP in image generation tasks. Mode collapse suggests fBP struggles to explore diverse data regions, making it unsuitable for generating realistic, structured outputs that adhere to specific distribution characteristics, like image data in the MNIST dataset.

In summary, fBP diverges significantly from the MCMC-based sampling methods used in our study due to its deterministic message-passing mechanism, which converges to fixed configurations rather than generating sequential probabilistic samples. While a Python wrapper exists, its reliance on the underlying Julia or C++ implementations introduces potential cross-language communication overhead, creating performance inconsistencies when compared to native Python implementations. Moreover, fBP's lack of constraint adherence and dependence on spin-like variable encoding make it unsuitable for complex, structured sampling tasks like TSP or data-driven applications requiring diverse sampling, such as image generation on MNIST. Our preliminary experiments confirm that fBP struggles with mode collapse and fails to capture essential data distribution characteristics.

D Proofs

D.1 Proof of Lemma 4.1

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

$$p(\widetilde{\boldsymbol{\theta}}) = p(\boldsymbol{\theta}, \boldsymbol{\theta}_a) \propto \exp\left\{U(\boldsymbol{\theta}) - \frac{1}{2\eta}\|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2\right\}. \tag{10}$$

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

$$\begin{aligned} p(\boldsymbol{\theta}) &= \int p(\boldsymbol{\theta}, \boldsymbol{\theta}_a) d\boldsymbol{\theta}_a \\ &= (2\pi\eta)^{-\frac{d}{2}} Z^{-1} \int \exp\left\{ U(\boldsymbol{\theta}) - \frac{1}{2\eta} \|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2 \right\} d\boldsymbol{\theta}_a \\ &= Z^{-1} \exp(U(\boldsymbol{\theta}))(2\pi\eta)^{-\frac{d}{2}} \int \exp\left\{ -\frac{1}{2\eta} \|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2 \right\} d\boldsymbol{\theta}_a \\ &= Z^{-1} \exp(U(\boldsymbol{\theta})), \end{aligned} \tag{11}$$

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

$$\sum_{\boldsymbol{\Theta}} \int \exp\left\{ U(\boldsymbol{\theta}) - \frac{1}{2\eta} \|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2 \right\} d\boldsymbol{\theta}_a = (2\pi\eta)^{\frac{d}{2}} \sum_{\boldsymbol{\Theta}} \exp(U(\boldsymbol{\theta})) := (2\pi\eta)^{\frac{d}{2}} Z. \tag{12}$$

This verifies that the joint posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\theta}_a)$ is mathematically well-defined[3]. Similarly, the marginal distribution for $\boldsymbol{\theta}_a$ is:

$$\begin{aligned} p(\boldsymbol{\theta}_a) &= \sum_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}, \boldsymbol{\theta}_a) \\ &\propto \sum_{\boldsymbol{\Theta}} \exp\left\{ U(\boldsymbol{\theta}) - \frac{1}{2\eta} \|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2 \right\} \\ &= \exp \mathcal{F}(\boldsymbol{\theta}_a; \eta). \end{aligned} \tag{13}$$

D.2 Proof of Proposition 5.4

We follow a similar-style analysis as seen in Theorem 5.1 of Zhang et al. (2022).

Using Equation (9),

$$\begin{aligned} q_\gamma(\widetilde{\boldsymbol{\theta}}'|\widetilde{\boldsymbol{\theta}}) &\propto \exp\left(\frac{1}{2} \nabla_{\boldsymbol{\theta}} U_\eta(\widetilde{\boldsymbol{\theta}})^\top (\boldsymbol{\theta}' - \boldsymbol{\theta}) - \frac{1}{2\alpha} \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2 \right) \cdot \frac{1}{\sqrt{2\pi\alpha_a}^d} \exp\left(-\frac{1}{2\alpha_a} \|\boldsymbol{\theta}'_{\boldsymbol{a}} - \boldsymbol{\theta}_{\boldsymbol{a}} - \frac{\alpha_a}{2} \nabla_{\boldsymbol{\theta}_{\boldsymbol{a}}} U_\eta(\widetilde{\boldsymbol{\theta}}) \|^2 \right) \\ &= \frac{1}{\sqrt{2\pi\alpha_a}^d} \exp\left(\frac{1}{2} \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})^\top (\boldsymbol{\theta}' - \boldsymbol{\theta}) - \frac{1}{2\alpha} \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2 - \frac{1}{2\eta} (\boldsymbol{\theta} - \boldsymbol{\theta}_a)^\top (\boldsymbol{\theta}' - \boldsymbol{\theta}) \right) \cdot \\ &\quad \left(-\frac{1}{2\alpha_a} \|\boldsymbol{\theta}'_a - \boldsymbol{\theta}_a\|^2 + \frac{1}{2\eta} (\boldsymbol{\theta} - \boldsymbol{\theta}_a)^\top (\boldsymbol{\theta}'_a - \boldsymbol{\theta}_a) - \frac{\alpha_a}{8\eta^2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2 \right) \\ &= \frac{1}{\sqrt{(2\pi\alpha_a)^d}} \exp\left(\frac{1}{2}(-U(\boldsymbol{\theta}) + U(\boldsymbol{\theta}')) - (\boldsymbol{\theta} - \boldsymbol{\theta}')^\top \left(\frac{1}{2\alpha} I + \frac{1}{4} \int_0^1 \nabla^2 U((1-s)\boldsymbol{\theta} + s\boldsymbol{\theta}') ds \right) (\boldsymbol{\theta} - \boldsymbol{\theta}') \right. \\ &\quad \left. -\frac{1}{2\eta} (\boldsymbol{\theta} - \boldsymbol{\theta}_a)^\top (\boldsymbol{\theta}' - \boldsymbol{\theta} + \boldsymbol{\theta}_a - \boldsymbol{\theta}'_a) - \frac{1}{2\alpha_a} \|\boldsymbol{\theta}'_a - \boldsymbol{\theta}_a\|^2 - \frac{\alpha_a}{8\eta^2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2 \right) \\ &= \frac{1}{\sqrt{(2\pi\alpha_a)^d}} \exp\left(\frac{1}{2}(-U(\boldsymbol{\theta}) + U(\boldsymbol{\theta}')) - (\boldsymbol{\theta} - \boldsymbol{\theta}')^\top \left(\frac{1}{2\alpha} I + \frac{1}{4} \int_0^1 \nabla^2 U((1-s)\boldsymbol{\theta} + s\boldsymbol{\theta}') ds \right) (\boldsymbol{\theta} - \boldsymbol{\theta}') \right. \\ &\quad \left. -\frac{1}{2\eta} (\boldsymbol{\theta} - \boldsymbol{\theta}_a)^\top (\boldsymbol{\theta}' - \boldsymbol{\theta}'_a) - \frac{1}{2\alpha_a} \|\boldsymbol{\theta}'_a - \boldsymbol{\theta}_a\|^2 + \frac{4\eta - \alpha_a}{8\eta^2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2 \right) \end{aligned}$$

The normalizing constant for Equation (9) $Z_{\widetilde{\boldsymbol{\Theta}}}(\widetilde{\boldsymbol{\theta}})$ is computed by integrating over $\mathbb{R}^d$ and summing over $\boldsymbol{\Theta}$:

$$Z_{\widetilde{\boldsymbol{\Theta}}}(\widetilde{\boldsymbol{\theta}}) = \frac{1}{\sqrt{2\pi\alpha_a}^d} \int_{\boldsymbol{\theta}'_a} \sum_{\boldsymbol{\theta}' \in \Theta} \exp\left(\frac{1}{2} \nabla_{\boldsymbol{\theta}} U_\eta(\widetilde{\boldsymbol{\theta}})^\top (\boldsymbol{\theta}' - \boldsymbol{\theta}) - \frac{1}{2\alpha} \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2 - \frac{1}{2\alpha_a} \|\boldsymbol{\theta}'_a - \boldsymbol{\theta}_a - \frac{\alpha_a}{2} \nabla_{\boldsymbol{\theta}_a} U_\eta(\widetilde{\boldsymbol{\theta}}) \|^2 \right) d\boldsymbol{\theta}'_a \tag{14}$$

We note that since $\nabla^2 U(\cdot)$ is continuous(from Assumption 5.2), we know that

$$\min_{x,y \in \boldsymbol{\Theta}} (x - y)^T \left(\int_0^1 \nabla^2 U((1-s)x + sy) ds \right) (x - y)$$

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

is well-defined.

Consequently, the modified normalizing constant(Equation (14)), $Z_\gamma(\widetilde{\boldsymbol{\theta}})$, becomes

$$Z_\gamma(\widetilde{\boldsymbol{\theta}}) = \frac{1}{\sqrt{(2\pi\alpha_a)^d}} \int_{\boldsymbol{\theta}'_a} \sum_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \exp\left(\frac{1}{2}\big(-U(\boldsymbol{\theta}) + U(\boldsymbol{\theta}')\big) - (\boldsymbol{\theta} - \boldsymbol{\theta}')^\top \left(\frac{1}{2\alpha} I + \frac{1}{4} \int_0^1 \nabla^2 U\big((1-s)\boldsymbol{\theta} + s\boldsymbol{\theta}'\big)\, ds \right) (\boldsymbol{\theta} - \boldsymbol{\theta}') \right.$$
$$\left. - \frac{1}{2\eta}(\boldsymbol{\theta} - \boldsymbol{\theta}_a)^\top (\boldsymbol{\theta}' - \boldsymbol{\theta}'_a) - \frac{1}{2\alpha_a}\|\boldsymbol{\theta}'_a - \boldsymbol{\theta}_a\|^2 + \frac{4\eta - \alpha_a}{8\eta^2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2 \right).$$

Now, we establish that $q(\widetilde{\boldsymbol{\theta}}|\widetilde{\boldsymbol{\theta}'})$ is reversible with respect to π_γ, where

$$\pi_\gamma = \frac{Z_\gamma(\widetilde{\boldsymbol{\theta}}) \exp\{\frac{\alpha_a}{8\eta^2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2\} \pi(\widetilde{\boldsymbol{\theta}})}{\int_y \sum_{x \in \boldsymbol{\Theta}} Z_\gamma([x^\top, y^\top]^\top) \exp\frac{\alpha_a}{8\eta^2}\|x - y\|^2 \pi([x^\top, y^\top]^\top) dy}.$$

Note that,

$$\begin{aligned}
\pi_\gamma(\widetilde{\boldsymbol{\theta}}) q_\gamma(\widetilde{\boldsymbol{\theta}'}|\widetilde{\boldsymbol{\theta}}) &= \frac{Z_\gamma(\widetilde{\boldsymbol{\theta}}) \exp\left(\frac{\alpha_a}{8\eta^2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2\right) \pi(\widetilde{\boldsymbol{\theta}})}{\int_y \sum_{x \in \boldsymbol{\Theta}} Z_\gamma([x^\top, y^\top]^\top) \exp\left(\frac{\alpha_a}{8\eta^2}\|x - y\|^2\right) \pi([x^\top, y^\top]^\top)\, dy} \frac{1}{Z_\gamma(\widetilde{\boldsymbol{\theta}})} \frac{1}{(\sqrt{2\pi\alpha_a})^d} \\
&\quad \exp\left(\frac{1}{2}\big(-U(\boldsymbol{\theta}) + U(\boldsymbol{\theta}')\big) - (\boldsymbol{\theta} - \boldsymbol{\theta}')^\top \left(\frac{1}{2\alpha} I + \frac{1}{4} \int_0^1 \nabla^2 U\big((1-s)\boldsymbol{\theta} + s\boldsymbol{\theta}'\big)\, ds \right) (\boldsymbol{\theta} - \boldsymbol{\theta}') \right. \\
&\quad \left. - \frac{1}{2\eta}(\boldsymbol{\theta} - \boldsymbol{\theta}_a)^\top (\boldsymbol{\theta}' - \boldsymbol{\theta}'_{\boldsymbol{a}}) - \frac{1}{2\alpha_a}\|\boldsymbol{\theta}'_a - \boldsymbol{\theta}_a\|^2 + \frac{4\eta - \alpha_a}{8\eta^2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2 \right) \\
&= \frac{1}{\int_y \sum_{x \in \boldsymbol{\Theta}} Z_\gamma([x^\top, y^\top]^\top) \exp\left(\frac{\alpha_a}{8\eta^2}\|x - y\|^2\right) \pi([x^\top, y^\top]^\top)\, dy} \frac{1}{(\sqrt{2\pi\alpha_a})^d} \\
&\quad \exp\left(\frac{1}{2}\big(U(\boldsymbol{\theta}) + U(\boldsymbol{\theta}')\big) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}')^\top \left(\frac{1}{\alpha} I + \frac{1}{2} \int_0^1 \nabla^2 U\big((1-s)\boldsymbol{\theta} + s\boldsymbol{\theta}'\big)\, ds \right) (\boldsymbol{\theta} - \boldsymbol{\theta}') \right. \\
&\quad \left. - \frac{1}{2\eta}(\boldsymbol{\theta} - \boldsymbol{\theta}_a)^\top (\boldsymbol{\theta}' - \boldsymbol{\theta}'_a) - \frac{1}{2\alpha_a}\|\boldsymbol{\theta}'_a - \boldsymbol{\theta}_a\|^2 \right) \\
&= \pi_\gamma(\boldsymbol{\theta}') q_\gamma(\boldsymbol{\theta}|\boldsymbol{\theta}').
\end{aligned}$$

Chain looks symmetric and reversible with respect to π_γ.

Now, given this, note that $Z'_\gamma(\widetilde{\boldsymbol{\theta}})$ converges to 1 as $\alpha \to 0$ and $\alpha_a \to 0$.

$$\begin{aligned}
Z'_\gamma(\widetilde{\boldsymbol{\theta}}) &= Z_\gamma(\widetilde{\boldsymbol{\theta}}) \exp\left(\frac{\alpha_a}{8\eta^2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2\right) \\
&= \frac{1}{\sqrt{(2\pi\alpha_a)^d}} \int_y \sum_x \exp\left(-\frac{1}{2}\left(U(\boldsymbol{\theta}) - U(x)\right) - (\boldsymbol{\theta} - x)^\top \left(\frac{1}{2\alpha} I + \frac{1}{4} \int_0^1 \nabla^2 U\big((1-s)\boldsymbol{\theta} + s\boldsymbol{\theta}'\big)\, ds \right) (\boldsymbol{\theta} - x) \right. \\
&\quad \left. - \frac{1}{2\alpha_a}\|y - \boldsymbol{\theta}_a\|^2 + \frac{4\eta}{8\eta^2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2 \right) dy \\
&\underset{\alpha \to 0}{=} \frac{1}{\sqrt{(2\pi\alpha_a)^d}} \int_y \sum_x \exp\left(\frac{1}{2}\left(U(x) - U(\boldsymbol{\theta})\right) - \frac{1}{2\alpha_a}\|y - \boldsymbol{\theta}_a\|^2 + \frac{1}{2\eta}\|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2 - \frac{1}{2\eta}(\boldsymbol{\theta} - \boldsymbol{\theta}_a)^\top (x - y) \right) \delta_{\boldsymbol{\theta}}(x)\, dy \\
&= \int_y \exp\left(\frac{1}{2\eta}\|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2 - \frac{1}{2\eta}(\boldsymbol{\theta} - \boldsymbol{\theta}_a)^\top (\boldsymbol{\theta} - y) \right) dy \\
&\underset{\alpha_a \to 0}{=} \int_y \exp\left(\frac{1}{2\eta}\|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2 - \frac{1}{2\eta}(\boldsymbol{\theta} - \boldsymbol{\theta}_a)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_a) \right) dy \\
&= 1.
\end{aligned}$$

where $\delta_\theta(.)$ is a Dirac delta. It follows that π_γ converges pointwisely to $\pi(\widetilde{\boldsymbol{\theta}})$. By Scheffé's Lemma, it immediately implies $\pi_\gamma(\widetilde{\boldsymbol{\theta}}) \to \pi(\widetilde{\boldsymbol{\theta}})$ as $\alpha \to 0$ and $\alpha_a \to 0$.

Let us consider the convergence rate in terms of the L_1-norm

$$\|\pi_\gamma - \pi\|_1 = \int_{\theta_a} \sum_{\theta \in \boldsymbol{\Theta}} \left| \frac{Z'_\gamma(\widetilde{\boldsymbol{\theta}}) \pi(\widetilde{\boldsymbol{\theta}})}{\int_y \sum_{x \in \boldsymbol{\Theta}} Z'_\gamma([x^\top, y^\top]^\top) \pi([x^\top, y^\top]^\top) dy} - \pi(\widetilde{\boldsymbol{\theta}}) \right| d\boldsymbol{\theta}_a$$

We write out each absolute value term

$$\left|\frac{Z'_\gamma(\widetilde{\boldsymbol{\theta}})\pi(\widetilde{\boldsymbol{\theta}})}{\int_y \sum_{x\in\boldsymbol{\Theta}} Z'_\gamma([x^\top, y^\top]^\top)\pi([x^\top, y^\top]^\top)dy} - \pi(\widetilde{\boldsymbol{\theta}})\right| = \pi(\widetilde{\boldsymbol{\theta}})\left|\frac{Z'_\gamma(\widetilde{\boldsymbol{\theta}})}{\int_y \sum_{x\in\boldsymbol{\Theta}} Z'_\gamma([x^\top, y^\top]^\top)\pi([x^\top, y^\top]^\top)dy} - 1\right|$$

First, we note that since U is M-gradient Lipschitz and $\frac{\alpha}{2} < \frac{1}{M}$, the matrix

$$\frac{1}{2\alpha}I - \frac{1}{4}\int_0^1 \nabla^2 U((1-s)\boldsymbol{\theta} + s\boldsymbol{\theta}')ds > \frac{1}{4}\left(\frac{2}{\alpha} - M\right)I$$

is positive definite.

Second, for $x' \in \boldsymbol{\Theta}$ and $y' \in \boldsymbol{\Theta}_a$ (under Assumptions 5.1 and 5.3), we know that the following minimum exists and is well-defined: $\min_{\substack{x\in\boldsymbol{\Theta}\setminus\{x'\}\\ y\in\boldsymbol{\Theta}_a\setminus\{y'\}}} (x-y)^\top(x'-y')$

Thus when, $\dfrac{Z'_\gamma(\widetilde{\boldsymbol{\theta}})}{\int_y \sum_{x\in\boldsymbol{\Theta}} Z'_\gamma\left(\begin{bmatrix}x^\top\\ y^\top\end{bmatrix}\right)\pi\left(\begin{bmatrix}x^\top\\ y^\top\end{bmatrix}\right)dy} - 1 \geq 0$, we get,

$$\begin{aligned}
&\left|\frac{Z'_\gamma(\widetilde{\boldsymbol{\theta}})\pi(\widetilde{\boldsymbol{\theta}})}{\int_y \sum_{x\in\boldsymbol{\Theta}} Z'_\gamma\left(\begin{bmatrix}x^\top\\ y^\top\end{bmatrix}\right)\pi\left(\begin{bmatrix}x^\top\\ y^\top\end{bmatrix}\right)dy} - \pi(\widetilde{\boldsymbol{\theta}})\right| = \pi(\widetilde{\boldsymbol{\theta}})\left|\frac{Z'_\gamma(\widetilde{\boldsymbol{\theta}})}{\int_y \sum_{x\in\boldsymbol{\Theta}} Z'_\gamma\left(\begin{bmatrix}x^\top\\ y^\top\end{bmatrix}\right)\pi\left(\begin{bmatrix}x^\top\\ y^\top\end{bmatrix}\right)dy} - 1\right| \\
&\qquad \leq \pi(\widetilde{\boldsymbol{\theta}})\left(1 + \frac{1}{\sqrt{(2\pi\alpha_a)^d}}\int_{y\neq\theta_a}\sum_{x\neq\theta}\exp\left(\frac{1}{2}(U(x)-U(\boldsymbol{\theta})) - \frac{1}{2}(\boldsymbol{\theta}-x)^\top\left(\frac{1}{\alpha}I + \frac{1}{2}\int_0^1 \nabla^2 U((1-s)\theta + sx)\,ds\right)(\boldsymbol{\theta}-x)\right.\right. \\
&\qquad\quad \left.\left. -\frac{1}{2\alpha_a}\|y-\boldsymbol{\theta}_a\|^2 + \frac{4\eta}{8\eta^2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_a\|^2 - \frac{1}{2\eta}(\boldsymbol{\theta}-\boldsymbol{\theta}_a)^\top(x-y)\right)dy - 1\right) \\
&\qquad \leq \frac{\pi(\widetilde{\boldsymbol{\theta}})}{\sqrt{(2\pi\alpha_a)^d}}\exp\left(\frac{M}{4} - \frac{1}{2\alpha} + \frac{1}{2\eta}\|\boldsymbol{\theta}-\boldsymbol{\theta}_a\|^2 - \frac{\vartheta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)}{2\eta}\right)\cdot\left(\int_{y\neq\theta_a}\sum_{x\neq\theta}\exp\left(\frac{1}{2}U(x) - \frac{1}{2}U(\boldsymbol{\theta}) - \frac{1}{2\alpha_a}\|y-\theta_a\|^2\right)dy\right) \\
&\qquad \leq \pi(\widetilde{\boldsymbol{\theta}})\exp\left(\frac{M}{4} - \frac{1}{2\alpha} + \frac{1}{2\eta}\|\boldsymbol{\theta}-\boldsymbol{\theta}_a\|^2 - \frac{\vartheta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)}{2\eta}\right)\left(\sum_x \exp\left(U(x)\right)\right) \\
&\qquad = \pi(\widetilde{\boldsymbol{\theta}})Z\exp\left(\frac{M}{4} - \frac{1}{2\alpha} + \frac{1}{2\eta}\|\boldsymbol{\theta}-\boldsymbol{\theta}_a\|^2 - \frac{\vartheta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)}{2\eta}\right) \\
&\qquad \leq \pi(\widetilde{\boldsymbol{\theta}})Z\exp\left(\frac{M}{4} - \frac{1}{2\alpha} + \frac{\Delta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)^2 - \vartheta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)}{2\eta}\right).
\end{aligned}$$

Similarly, when $\dfrac{Z'_\gamma(\widetilde{\boldsymbol{\theta}})}{\int_y \sum_{x\in\boldsymbol{\Theta}} Z'_\gamma\left(\begin{bmatrix}x^\top\\ y^\top\end{bmatrix}\right)\pi\left(\begin{bmatrix}x^\top\\ y^\top\end{bmatrix}\right)dy} - 1 < 0$, we get

$$\begin{aligned}
&\left|\frac{Z'_\gamma(\widetilde{\boldsymbol{\theta}})\pi(\widetilde{\boldsymbol{\theta}})}{\int_y \sum_{x\in\boldsymbol{\Theta}} Z'_\gamma\left(\begin{bmatrix}x^\top\\ y^\top\end{bmatrix}\right)\pi\left(\begin{bmatrix}x^\top\\ y^\top\end{bmatrix}\right)dy} - \pi(\widetilde{\boldsymbol{\theta}})\right| \\
&= \pi(\widetilde{\boldsymbol{\theta}})\left(1 - \frac{1 + \frac{1}{\sqrt{(2\pi\alpha_a)^d}}\int_{y\neq\theta_a}\sum_{x\neq\theta}\exp\left(\frac{1}{2}(U(x)-U(\boldsymbol{\theta})) - \frac{1}{2}(\boldsymbol{\theta}-x)^\top\left(\frac{1}{\alpha}I + \frac{1}{2}\int_0^1\nabla^2U((1-s)\theta+sx)ds\right)(\boldsymbol{\theta}-x) - \frac{1}{2\alpha_a}\|y-\boldsymbol{\theta}_a\|^2 + \frac{4\eta}{8\eta^2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_a\|^2 - \frac{1}{2\eta}(\boldsymbol{\theta}-\boldsymbol{\theta}_a)^\top(x-y)\right)dy}{1 + \frac{1}{\sqrt{2\pi\alpha_a}}\int_p\frac{1}{\sqrt{\pi}}\exp(-p^2)\int_{q\neq p}\sum_r\frac{1}{2}\exp(U(r))\sum_{s\neq r}\exp\left(\frac{1}{2}(U(s)-\frac{1}{2}U(r)) - \frac{1}{2}(r-s)^\top\left(\frac{1}{\alpha}I + \frac{1}{2}\int_0^1\nabla^2U((1-l)r+ls)dl\right)(r-s) - \frac{1}{2\alpha_a}\|q-p\|^2 + \frac{4\eta}{8\eta^2}\|r-p\|^2 - \frac{1}{2\eta}(r-p)^\top(s-q)\right)dq\,dp}\right) \\
&\leq \pi(\widetilde{\boldsymbol{\theta}})\left(1 - \frac{1}{1 + \frac{1}{\sqrt{2\pi\alpha_a}}\int_p\frac{1}{\sqrt{\pi}}\exp(-p^2)\int_{q\neq p}\exp\left(-\frac{1}{2\alpha_a}\|q-p\|^2\right)\sum_r\exp\left(\frac{4\eta}{8\eta^2}\|r-p\|^2\right)\frac{1}{Z}\exp(U(r))\sum_{s\neq r}\exp\left(\frac{1}{2}(U(s)-U(r)) - \frac{1}{2}(r-s)^\top\left(\frac{1}{\alpha}I + \frac{1}{2}\int_0^1\nabla^2U((1-l)r+ls)dl\right)(r-s) - \frac{1}{2\eta}(r-p)^\top(s-q)\right)dq\,dp}\right) \\
&= \pi(\widetilde{\boldsymbol{\theta}})\left(\frac{\frac{1}{\sqrt{2\pi\alpha_a}}\int_p\frac{1}{\sqrt{\pi}}\exp(-p^2)\int_{q\neq p}\exp\left(-\frac{1}{2\alpha_a}\|q-p\|^2\right)\sum_r\exp\left(\frac{4\eta}{8\eta^2}\|r-p\|^2\right)\frac{1}{Z}\exp(U(r))\sum_{s\neq r}\exp\left(\frac{1}{2}(U(s)-U(r)) - \frac{1}{2}(r-s)^\top\left(\frac{1}{\alpha}I + \frac{1}{2}\int_0^1\nabla^2U((1-l)r+ls)dl\right)(r-s) - \frac{1}{2\eta}(r-p)^\top(s-q)\right)dq\,dp}{1 + \frac{1}{\sqrt{2\pi\alpha_a}}\int_p\frac{1}{\sqrt{\pi}}\exp(-p^2)\int_{q\neq p}\exp\left(-\frac{1}{2\alpha_a}\|q-p\|^2\right)\sum_r\exp\left(\frac{4\eta}{8\eta^2}\|r-p\|^2\right)\frac{1}{Z}\exp(U(r))\sum_{s\neq r}\exp\left(\frac{1}{2}(U(s)-U(r)) - \frac{1}{2}(r-s)^\top\left(\frac{1}{\alpha}I + \frac{1}{2}\int_0^1\nabla^2U((1-l)r+ls)dl\right)(r-s) - \frac{1}{2\eta}(r-p)^\top(s-q)\right)dq\,dp}\right) \\
&\leq \frac{\pi(\widetilde{\boldsymbol{\theta}})}{\sqrt{2\pi\alpha_a}^d}\left(\int_p\frac{1}{\sqrt{\pi}^d}\exp(-p^2)\int_{q\neq p}\exp\left(-\frac{1}{2\alpha_a}\|q-p\|^2\right)\sum_r\exp\left(\frac{4\eta}{8\eta^2}\|r-p\|^2\right)\frac{1}{Z}\exp(U(r))\sum_{s\neq r}\exp\left(\frac{1}{2}(U(s)-U(r)) - \frac{1}{2}(r-s)^\top\left(\frac{1}{\alpha}I + \frac{1}{2}\int_0^1\nabla^2U((1-l)r+ls)dl\right)(r-s) - \frac{1}{2\eta}(r-p)^\top(s-q)\right)\right)dq\,dp \\
&\leq \frac{\pi(\widetilde{\boldsymbol{\theta}})}{\sqrt{(2\pi\alpha_a)^d}}\exp\left(\frac{M}{4} - \frac{1}{2\alpha}\right)\left(\int_p\frac{1}{\sqrt{\pi}^d}\exp(-p^2)\right)\int_{q\neq p}\exp\left(-\frac{1}{2\alpha_a}\|q-p\|^2\right)\sum_r\exp\left(\frac{1}{2\eta}\|r-p\|^2\right)\frac{1}{Z}\exp(U(r))\sum_{s\neq r}\exp\left(\frac{1}{2}(U(s)-U(r)) - \frac{1}{2\eta}(r-p)^\top(s-q)\right)\Big)dq\,dp \\
&\leq \frac{\pi(\widetilde{\boldsymbol{\theta}})}{\sqrt{2\pi\alpha_a}^d}\exp\left(\frac{M}{4} - \frac{1}{2\alpha} + \frac{\Delta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)^2 - \vartheta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)}{2\eta}\right)\left(\int_p\frac{1}{\sqrt{\pi}^d}\exp(-p^2)\int_{q\neq p}\exp\left(-\frac{1}{2\alpha_a}\|q-p\|^2\right)\sum_r\frac{1}{Z}(U(r))\sum_{s\neq r}\exp\left(\frac{1}{2}(U(s)-U(r))\right)dq\,dp\right) \\
&\leq \frac{\pi(\widetilde{\boldsymbol{\theta}})}{\sqrt{2\pi\alpha_a}^d}Z\exp\left(\frac{M}{4} - \frac{1}{2\alpha} + \frac{\Delta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)^2 - \vartheta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)}{2\eta}\right)\left(\int_p\frac{1}{\sqrt{\pi}^d}\exp(-p^2)\int_{q\neq p}\exp\left(-\frac{1}{2\alpha_a}\|q-p\|^2\right)dq\,dp\right) \\
&= \pi(\widetilde{\boldsymbol{\theta}})Z\exp\left(\frac{M}{4} - \frac{1}{2\alpha} + \frac{\Delta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)^2 - \vartheta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)}{2\eta}\right)\int_p\left(\frac{1}{\sqrt{\pi}^d}\exp(-p^2)\right)dp \\
&= \pi(\widetilde{\boldsymbol{\theta}})Z\exp\left(\frac{M}{4} - \frac{1}{2\alpha} + \frac{\Delta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)^2 - \vartheta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)}{2\eta}\right)
\end{aligned}$$

Therefore, the difference between π_γ and $\tilde{\pi}$ can be bounded as follows

$$\begin{aligned}
\|\pi_\gamma - \tilde{\pi}\|_1 &\leq \int_{\theta_a}\sum_{\theta\in\boldsymbol{\Theta}}\pi(\widetilde{\boldsymbol{\theta}})Z\exp\left(\frac{M}{4} - \frac{1}{2\alpha} + \frac{\Delta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)^2 - \vartheta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)}{2\eta}\right)d\theta_a \\
&\leq Z\exp\left(\frac{M}{4} - \frac{1}{2\alpha} + \frac{\Delta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)^2 - \vartheta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)}{2\eta}\right)
\end{aligned}$$

D.3 Proofs for EDULA

We start by establishing results for a more general case in which Assumption 5.3 is dropped. We establish that in this setting geometric rates of convergence exist. However, in this case proving that the stationary distribution is close to the target remains an open problem. .

Theorem D.1. *Let Assumption 5.1 hold. Then for the Markov chain with transition operator P as in Algorithm 1, the drift condition is satisfied as follows:*

$$PV(\tilde{\boldsymbol{\theta}}) \leq \alpha_a \, d + 2 \left(1 - \frac{\alpha_a}{\eta}\right)^2 V(\tilde{\boldsymbol{\theta}}) + 2 \, \frac{\alpha_a^2}{\eta^2} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\boldsymbol{\theta}\|^2.$$

Proof. We establish an explicit drift and minorization condition for the joint chain, which confirms the convergence rate. Note that

$$p((\boldsymbol{\theta}'_a, \boldsymbol{\theta}') \mid (\boldsymbol{\theta}'_a, \boldsymbol{\theta}')) = p(\boldsymbol{\theta}'_a \mid \boldsymbol{\theta}, \boldsymbol{\theta}_a) \cdot p(\boldsymbol{\theta}' \mid \boldsymbol{\theta}_a, \boldsymbol{\theta}).$$

Now,

$$p(\boldsymbol{\theta}'_a \mid \boldsymbol{\theta}, \boldsymbol{\theta}_a) = \frac{1}{(2\pi\alpha_a)^{d/2}} \exp\left\{ -\frac{1}{2\alpha_a} \left\| \boldsymbol{\theta}'_a - \boldsymbol{\theta}_a \left(1 - \frac{\alpha_a}{\eta}\right) - \frac{\alpha_a}{\eta} \boldsymbol{\theta} \right\|^2 \right\}$$

and

$$p(\boldsymbol{\theta}' \mid \boldsymbol{\theta}_a, \boldsymbol{\theta}) = \frac{\exp\left\{ -\frac{1}{2\alpha} \left\| \boldsymbol{\theta}' - \boldsymbol{\theta} + \alpha \nabla U(\boldsymbol{\theta}) - \frac{\alpha}{\eta} (\boldsymbol{\theta} - \boldsymbol{\theta}_a) \right\|^2 \right\}}{\sum_{\boldsymbol{x} \in \boldsymbol{\Theta}} \exp\left\{ -\frac{1}{2\alpha} \left\| \boldsymbol{x} - \boldsymbol{\theta} + \alpha \nabla U(\boldsymbol{\theta}) - \frac{\alpha}{\eta} (\boldsymbol{\theta} - \boldsymbol{\theta}_a) \right\|^2 \right\}}.$$

Therefore, our Markov transition operator P is given as

$$P((\boldsymbol{\theta}_a, \boldsymbol{\theta}), A) = \int_A p((\boldsymbol{\theta}'_a, \boldsymbol{\theta}') \mid (\boldsymbol{\theta}, \boldsymbol{\theta}_a)) \, d\mu,$$

where $A \in \boldsymbol{\Theta} \times \mathbb{R}^d$ and μ is the product of the counting measure and Lebesgue measure.

We shall first establish a drift condition:

$$PV \leq \lambda V + b,$$

where we choose the Lyapunov function $V(\boldsymbol{x}_1, \boldsymbol{x}_2) = \|\boldsymbol{x}_1\|^2$ and some constant $b > 0$.

We note that

$$\begin{aligned} PV(\boldsymbol{\theta}_a, \boldsymbol{\theta}) = \frac{1}{(2\pi\alpha_a)^{d/2}} \sum_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \int \|\boldsymbol{\theta}'_a\|^2 \exp\left\{ -\frac{1}{2\alpha_a} \left\| \boldsymbol{\theta}'_a - \boldsymbol{\theta}_a \left(1 - \frac{\alpha_a}{\eta}\right) - \frac{\alpha_a}{\eta} \boldsymbol{\theta} \right\|^2 \right\} \\ \cdot \frac{\exp\left\{ -\frac{1}{2\alpha} \left\| \boldsymbol{\theta}' - \boldsymbol{\theta} + \alpha \nabla U(\boldsymbol{\theta}) - \frac{\alpha}{\eta} (\boldsymbol{\theta} - \boldsymbol{\theta}_a) \right\|^2 \right\}}{\sum_{\boldsymbol{x} \in \boldsymbol{\Theta}} \exp\left\{ -\frac{1}{2\alpha} \left\| \boldsymbol{x} - \boldsymbol{\theta} + \alpha \nabla U(\boldsymbol{\theta}) - \frac{\alpha}{\eta} (\boldsymbol{\theta} - \boldsymbol{\theta}_a) \right\|^2 \right\}} \, d\boldsymbol{\theta}_a. \end{aligned}$$

Using a change of variables, we have

$$\begin{aligned} PV(\boldsymbol{\theta}_a, \boldsymbol{\theta}) &= \frac{1}{(2\pi\alpha_a)^{d/2}} \sum_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \int \left\| u + \boldsymbol{\theta}_a \left(1 - \frac{\alpha_a}{\eta}\right) + \frac{\alpha_a}{\eta} \boldsymbol{\theta} \right\|^2 \exp\left\{ -\frac{1}{2\alpha_a} \|u\|^2 \right\} \\ &\quad \cdot \frac{\exp\left\{ -\frac{1}{2\alpha} \left\| \boldsymbol{\theta}' - \boldsymbol{\theta} + \alpha \nabla U(\boldsymbol{\theta}) - \frac{\alpha}{\eta} (\boldsymbol{\theta} - \boldsymbol{\theta}_a) \right\|^2 \right\}}{\sum_{\boldsymbol{x} \in \boldsymbol{\Theta}} \exp\left\{ -\frac{1}{2\alpha} \left\| \boldsymbol{x} - \boldsymbol{\theta} + \alpha \nabla U(\boldsymbol{\theta}) - \frac{\alpha}{\eta} (\boldsymbol{\theta} - \boldsymbol{\theta}_a) \right\|^2 \right\}} \, du \\ &\leq \alpha_a \, d + 2 \left(1 - \frac{\alpha_a}{\eta}\right)^2 \|\boldsymbol{\theta}_a\|^2 + 2 \, \frac{\alpha_a^2}{\eta^2} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\boldsymbol{\theta}\|^2. \end{aligned}$$

Note that when $\lambda = 2 \left(1 - \frac{\alpha_a}{\eta}\right)^2 < 1$, then this is a proper drift condition with $b = \alpha_a \, d + 2 \, \frac{\alpha_a^2}{\eta^2} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\boldsymbol{\theta}\|^2$.

Theorem D.2. *Under Assumption 5.1, the Markov chain with transition operator P as in Algorithm 1 satisfies,*

$$P(\tilde{\boldsymbol{\theta}}, A) \geq \bar{\eta}\mu(A)$$

where $\bar{\eta} > 0$ is defined in (16) *and $\mu(\cdot)$ is the product of Lebesgue measure and counting measure and $\tilde{\boldsymbol{\theta}} \in C_\alpha$ as in* (15) *.*

Proof. We establish a minorization on the set,

$$C_{\alpha_a} = \left\{ x : V(x) \leq \frac{2\left(\alpha_a\, d + 2\,\frac{\alpha_a^2}{\eta^2}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\|\boldsymbol{\theta}\|^2\right)}{\left(1-\frac{\alpha_a}{\eta}\right)^2} \right\} \tag{15}$$

We define

$$\begin{aligned}\bar{\eta} = \frac{1}{(2\pi\alpha_a)^{d/2}}\exp\left\{-\frac{4}{\alpha_a}\frac{\left(\alpha_a\, d + 2\,\frac{\alpha_a^2}{\eta^2}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\|\boldsymbol{\theta}\|^2\right)}{\left(1-\frac{\alpha_a}{\eta}\right)^2}\right\}\cdot\frac{1}{|\boldsymbol{\Theta}|} \\ \cdot\exp\left\{-\frac{1}{2\alpha}\left[\left((\alpha\,M+1)^2+\alpha\,M^2\right)\text{diam}(\boldsymbol{\Theta})^2 + (2\,(M+\alpha)+2\alpha M)\,\|\nabla U(a)\|\text{diam}(\boldsymbol{\Theta}) + \left(\alpha^2+\alpha\right)\|\nabla U(a)\|^2\right.\right. \\ \left.\left. +2\frac{\alpha}{\eta}\left[(\alpha\,M+1)^2\,\text{diam}(\boldsymbol{\Theta})^2 + 2\,(M+\alpha)\|\nabla U(a)\|\text{diam}(\boldsymbol{\Theta}) + \alpha^2\|\nabla U(a)\|^2\right]^{1/2}\text{diam}(\boldsymbol{\Theta})\right]\right\}\end{aligned} \tag{16}$$

We start with considering any $(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) \in C_\alpha$. Further, we also have $(\boldsymbol{\theta_a}, \boldsymbol{\theta}) \in C_{\alpha_a}$. Therefore

$$\begin{aligned}p((\boldsymbol{\theta_1},\boldsymbol{\theta_2}) \mid (\boldsymbol{\theta_a},\boldsymbol{\theta})) = \frac{1}{(2\pi\alpha_a)^{d/2}}\exp\left\{-\frac{1}{2\alpha_a}\left\|\boldsymbol{\theta_1}-\boldsymbol{\theta_a}\left(1-\frac{\alpha_a}{\eta}\right)-\frac{\alpha_a}{\eta}\boldsymbol{\theta}\right\|^2\right\} \\ \cdot\frac{\exp\left\{-\frac{1}{2\,\alpha}\left\|\boldsymbol{\theta_2}-\boldsymbol{\theta}+\alpha\nabla U(\boldsymbol{\theta})-\frac{\alpha}{\eta}\left(\boldsymbol{\theta}-\boldsymbol{\theta_a}\right)\right\|^2\right\}}{\sum_{x\in\boldsymbol{\Theta}}\exp\left\{-\frac{1}{2\,\alpha}\left\|x-\boldsymbol{\theta}+\alpha\nabla U(\boldsymbol{\theta})-\frac{\alpha}{\eta}\left(\boldsymbol{\theta}-\boldsymbol{\theta_a}\right)\right\|^2\right\}}.\end{aligned}$$

For the first term, we note that

$$\begin{aligned}\left\|\boldsymbol{\theta_1}-\boldsymbol{\theta_a}\left(1-\frac{\alpha_a}{\eta}\right)-\frac{\alpha_a}{\eta}\boldsymbol{\theta}\right\|^2 &\leq 2\,\|\boldsymbol{\theta_1}\|^2 + 2\left\|\left(1-\frac{\alpha_a}{\eta}\right)\boldsymbol{\theta_a}+\frac{\alpha_a}{\eta}\boldsymbol{\theta}\right\|^2 \\ &\leq 2\,\|\boldsymbol{\theta_1}\|^2 + 2\left(1-\frac{\alpha_a}{\eta}\right)\|\boldsymbol{\theta_a}\|^2 + 2\,\frac{\alpha_a}{\eta}\,\|\boldsymbol{\theta}\|^2 \\ &\leq 8\,\frac{\left(\alpha_a\, d + 2\,\frac{\alpha_a^2}{\eta^2}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\|\boldsymbol{\theta}\|^2\right)}{\left(1-\frac{\alpha_a}{\eta}\right)^2}.\end{aligned}$$

Therefore, the first term is greater than

$$\begin{aligned}&\frac{1}{(2\pi\alpha_a)^{d/2}}\exp\left\{-\frac{1}{2\alpha_a}\left\|\boldsymbol{\theta_1}-\boldsymbol{\theta_a}\left(1-\frac{\alpha_a}{\eta}\right)-\frac{\alpha_a}{\eta}\boldsymbol{\theta_2}\right\|^2\right\} \\ &\geq \frac{1}{(2\pi\alpha_a)^{d/2}}\exp\left\{-\frac{4}{\alpha_a}\frac{\left(\alpha_a\, d + 2\,\frac{\alpha_a^2}{\eta^2}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\|\boldsymbol{\theta}\|^2\right)}{\left(1-\frac{\alpha_a}{\eta}\right)^2}\right\}.\end{aligned}$$

For the second term, note that

$$\frac{\exp\left\{-\frac{1}{2\,\alpha}\left\|\boldsymbol{\theta_2}-\boldsymbol{\theta}+\alpha\nabla U(\boldsymbol{\theta})-\frac{\alpha}{\eta}\left(\boldsymbol{\theta}-\boldsymbol{\theta_a}\right)\right\|^2\right\}}{\sum_{x\in\boldsymbol{\Theta}}\exp\left\{-\frac{1}{2\,\alpha}\left\|x-\boldsymbol{\theta}+\alpha\nabla U(\boldsymbol{\theta})-\frac{\alpha}{\eta}\left(\boldsymbol{\theta}-\boldsymbol{\theta_a}\right)\right\|^2\right\}} \geq \frac{1}{|\boldsymbol{\Theta}|}\exp\left\{-\frac{1}{2\,\alpha}\left\|\boldsymbol{\theta_2}-\boldsymbol{\theta}+\alpha\nabla U(\boldsymbol{\theta})-\frac{\alpha}{\eta}\left(\boldsymbol{\theta}-\boldsymbol{\theta_a}\right)\right\|^2\right\}.$$

For the numerator, one sees,

$$\left\|\boldsymbol{\theta_2}-\boldsymbol{\theta}+\alpha\nabla U(\boldsymbol{\theta})-\frac{\alpha}{\eta}\left(\boldsymbol{\theta}-\boldsymbol{\theta_a}\right)\right\|^2 \leq \|\boldsymbol{\theta_2}-\boldsymbol{\theta}+\alpha\nabla U(\boldsymbol{\theta})\|^2+\frac{\alpha^2}{\eta^2}\|\boldsymbol{\theta}-\boldsymbol{\theta_a}\|^2 + 2\frac{\alpha}{\eta}\|\boldsymbol{\theta_2}-\boldsymbol{\theta}+\alpha\nabla U(\boldsymbol{\theta})\|\,\|\boldsymbol{\theta}-\boldsymbol{\theta_a}\|.$$

For the first term, we have

$$\|\boldsymbol{\theta_2}-\boldsymbol{\theta}+\alpha\nabla U(\boldsymbol{\theta})\|^2 \leq \|\boldsymbol{\theta_2}-\boldsymbol{\theta}\|^2+\alpha^2\|\nabla U(\boldsymbol{\theta})\|^2+2\,\alpha\,\|\boldsymbol{\theta_2}-\boldsymbol{\theta}\|\,\|\nabla U(\boldsymbol{\theta})\|.$$

Define $a=\text{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\|\nabla U(\boldsymbol{\theta})\|$. Therefore, the above expression is less than

$$\begin{aligned}\|\boldsymbol{\theta_2}-\boldsymbol{\theta}+\alpha\nabla U(\boldsymbol{\theta})\|^2 &\leq \text{diam}(\boldsymbol{\Theta})^2+\alpha^2\left(M^2\text{diam}(\boldsymbol{\Theta})^2+\|\nabla U(a)\|^2+2\,M\,\text{diam}(\boldsymbol{\Theta})\|\nabla U(a)\|\right)\\ &\quad+2\alpha\,\text{diam}(\boldsymbol{\Theta})\left(M\,\text{diam}(\boldsymbol{\Theta})+\|\nabla U(a)\|\right)\\ &\leq (\alpha\,M+1)^2\,\text{diam}(\boldsymbol{\Theta})^2+2\,(M+\alpha)\|\nabla U(a)\|\text{diam}(\boldsymbol{\Theta})+\alpha^2\|\nabla U(a)\|^2.\end{aligned}$$

For the second term, we have

$$\alpha\|\nabla U(\boldsymbol{\theta})\|^2 \leq \alpha M^2\text{diam}(\boldsymbol{\Theta})^2+\alpha\|\nabla U(a)\|^2+2\alpha\,M\,\text{diam}(\boldsymbol{\Theta})\|\nabla U(a)\|$$

and for the final term we have

$$\begin{aligned}2\frac{\alpha}{\eta}\|\boldsymbol{\theta_2}-\boldsymbol{\theta}+\alpha\nabla U(\boldsymbol{\theta})\|\,\|\boldsymbol{\theta}-\boldsymbol{\theta_a}\| \leq 2\frac{\alpha}{\eta}\Big[&(\alpha M+1)^2\,\text{diam}(\boldsymbol{\Theta})^2+2(M+\alpha)\|\nabla U(a)\|\text{diam}(\boldsymbol{\Theta})\\ &+\alpha^2\|\nabla U(a)\|^2\Big]^{1/2}\text{diam}(\boldsymbol{\Theta}).\end{aligned}\tag{17}$$

Therefore we have

$$\begin{aligned}&\frac{\exp\left\{-\frac{1}{2\,\alpha}\left\|\boldsymbol{\theta_2}-\boldsymbol{\theta}+\alpha\nabla U(\boldsymbol{\theta})-\frac{\alpha}{\eta}\left(\boldsymbol{\theta}-\boldsymbol{\theta_a}\right)\right\|^2\right\}}{\sum_{x\in\boldsymbol{\Theta}}\exp\left\{-\frac{1}{2\,\alpha}\left\|x-\boldsymbol{\theta}+\alpha\nabla U(\boldsymbol{\theta})-\frac{\alpha}{\eta}\left(\boldsymbol{\theta}-\boldsymbol{\theta_a}\right)\right\|^2\right\}}\\ &\geq \frac{1}{|\boldsymbol{\Theta}|}\exp\left\{-\frac{1}{2\alpha}\left[\left((\alpha\,M+1)^2+\alpha\,M^2\right)\text{diam}(\boldsymbol{\Theta})^2+(2\,(M+\alpha)+2\alpha M)\,\|\nabla U(a)\|\text{diam}(\boldsymbol{\Theta})+\left(\alpha^2+\alpha\right)\|\nabla U(a)\|^2\right.\right.\\ &\qquad\left.\left.+2\frac{\alpha}{\eta}\left[(\alpha\,M+1)^2\,\text{diam}(\boldsymbol{\Theta})^2+2\,(M+\alpha)\|\nabla U(a)\|\text{diam}(\boldsymbol{\Theta})+\alpha^2\|\nabla U(a)\|^2\right]^{1/2}\text{diam}(\boldsymbol{\Theta})\right]\right\}.\end{aligned}$$

This finally gives $\tilde{\eta}$ as

$$\begin{aligned}\bar{\eta}=&\frac{1}{(2\pi\alpha_a)^{d/2}}\exp\left\{-\frac{4}{\alpha_a}\frac{\left(\alpha_a\,d+2\,\frac{\alpha_a^2}{\eta^2}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\|\boldsymbol{\theta}\|^2\right)}{\left(1-\frac{\alpha_a}{\eta}\right)^2}\right\}\\ &\cdot\frac{1}{|\boldsymbol{\Theta}|}\exp\left\{-\frac{1}{2\alpha}\left[\left((\alpha\,M+1)^2+\alpha\,M^2\right)\text{diam}(\boldsymbol{\Theta})^2+(2\,(M+\alpha)+2\alpha M)\,\|\nabla U(a)\|\text{diam}(\boldsymbol{\Theta})+\left(\alpha^2+\alpha\right)\|\nabla U(a)\|^2\right.\right.\\ &\qquad\left.\left.+2\frac{\alpha}{\eta}\left[(\alpha\,M+1)^2\,\text{diam}(\boldsymbol{\Theta})^2+2\,(M+\alpha)\|\nabla U(a)\|\text{diam}(\boldsymbol{\Theta})+\alpha^2\|\nabla U(a)\|^2\right]^{1/2}\text{diam}(\boldsymbol{\Theta})\right]\right\}\end{aligned}$$

with the reference measure $\mu(\cdot)$ is the product measure of the Lebesgue measure and the counting measure.

Lemma D.3. *The Markov chain defined by Algorithm 1 is irreducible, aperiodic and Harris recurrent.*

Proof. For any Borel measurable A with $\lambda(A)>0$ and any $\boldsymbol{\theta}\in\boldsymbol{\Theta}$, we have

$$\mathbb{P}\left(\boldsymbol{\theta}'_a\in A,\,\boldsymbol{\theta}'=\boldsymbol{\theta}^*\mid\boldsymbol{\theta}_a,\,\boldsymbol{\theta}\right)=\mathbb{P}\left(\boldsymbol{\theta}'_a\in A\mid\boldsymbol{\theta}_a,\,\theta\right)\mathbb{P}\left(\boldsymbol{\theta}'=\boldsymbol{\theta}^*\mid\boldsymbol{\theta}_a,\,\boldsymbol{\theta}\right).$$

Note that both the above terms are positive since the first distribution is Gaussian and the second term is positive by definition. We can similarly establish aperiodicity by noting that there is no partition of $\boldsymbol{\Theta}\times\mathbb{R}^d$ such that the previous probability is 1. Finally, due to the fact that the algorithm satisfies a drift condition, the Markov chain is Harris.

We may leverage the above results to obtain a rate of convergence of the sampler using Ekvall & Jones (2021).

Theorem D.4. *The Markov chain has a stationary distribution dependent on $\gamma = (\alpha, \alpha_a)$, π_γ, and is (M, ρ) geometrically ergodic with*

$$\|P^k(x, \cdot) - \pi_\gamma(\cdot)\|_{TV} \leq M(x)\rho^k$$

where

$$M(x) = 2 + \frac{\tilde{b}}{1 - \tilde{\lambda}} + \tilde{V}(x)$$

and

$$\rho \leq \max\left\{(1 - \bar{\eta})^r, \left(\frac{1 + 2\tilde{b} + \tilde{\lambda} + \tilde{\lambda}d}{1 + d}\right)^{1-r}\left(1 + 2\tilde{b} + 2\tilde{\lambda}d\right)^r\right\}$$

for some free parameter $0 < r < 1$ and where $\bar{\eta}$, b, λ are previously defined.

Proof. The proof follows directly from Theorem D.1, Theorem D.2 and Lemma D.3 Ekvall & Jones (2021).

Theorem D.5. *For any function $f : \mathbb{R}^p \to \mathbb{R}$ with $f^2(x) \leq V(x)$ for all $x \in \mathbb{R}^p$ one has*

$$\sqrt{n}\left(\bar{f} - \mathbb{E}_{\pi_\gamma} f\right) \xrightarrow{d} N(0, \sigma_f^2)$$

as $n \to \infty$, where $\sigma_f^2 \in [0, \infty)$. , where

$$\bar{f} = \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

Proof. The proof follows from Theorem D.1 by noting that $PV \leq \lambda V + b$ implies

$$P(V + 1) \leq \lambda(V + 1) + (b + 1 - \lambda).$$

This implies a drift condition holds with $V : \mathbb{R}^d \to [1, \infty)$. Hence the result follows via Jones (2004). Note that $\sigma_f^2 = 0$ implies convergence to a Gaussian degenerate at 0.

Define

$$\begin{aligned}\bar{\eta}^* = \frac{1}{\Phi_{\alpha_a}(\boldsymbol{\Theta}_a)}\exp\left\{-\frac{1}{\alpha_a}\operatorname{diam}(\boldsymbol{\Theta}_a)^2 - \frac{\alpha_a}{\eta^2}\Delta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_a)^2\right\} \\ \times \frac{1}{|\boldsymbol{\Theta}|}\exp\left\{-\frac{1}{2\alpha}\left[\left((\alpha M + 1)^2 + \alpha M^2\right)\operatorname{diam}(\boldsymbol{\Theta})^2\right.\right. \\ + (2(M + \alpha) + 2\alpha M)\|\nabla U(a)\|\operatorname{diam}(\boldsymbol{\Theta}) \\ + \left(\alpha^2 + \alpha\right)\|\nabla U(a)\|^2 \\ \left.\left. + 2\frac{\alpha}{\eta}\left[(\alpha M + 1)^2\operatorname{diam}(\boldsymbol{\Theta})^2 + 2(M + \alpha)\|\nabla U(a)\|\operatorname{diam}(\boldsymbol{\Theta}) + \alpha^2\|\nabla U(a)\|^2\right]^{1/2}\operatorname{diam}(\boldsymbol{\Theta})\right]\right\}.\end{aligned} \tag{18}$$

Lemma D.6. *Under Assumptions 5.1 and 5.3, the Markov chain with transition operator P as in Algorithm 1 satisfies,*

$$P((\boldsymbol{\theta}_a, \boldsymbol{\theta}), A) \geq \bar{\eta}^*\mu(A)$$

where $\bar{\eta}^* > 0$ is as defined in (18) *and $\mu(\cdot)$ is the product of Lebesgue measure and counting measure.*

Proof. We consider the case where $\boldsymbol{\theta}_a$ is restricted to some compact subset of $\mathbb{R}^d$, which we refer to as $\boldsymbol{\Theta}_a$. In this case, note that the transition kernel changes to

$$\begin{aligned}p((\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \mid (\boldsymbol{\theta}_a, \boldsymbol{\theta})) = \frac{1}{\Phi_{\alpha_a}(\boldsymbol{\Theta}_a)}\exp\left\{-\frac{1}{2\alpha_a}\left\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_a\left(1 - \frac{\alpha_a}{\eta}\right) - \frac{\alpha_a}{\eta}\boldsymbol{\theta}\right\|^2\right\} \\ \times \frac{\exp\left\{-\frac{1}{2\alpha}\left\|\boldsymbol{\theta}_2 - \boldsymbol{\theta} + \alpha\nabla U(\boldsymbol{\theta}) - \frac{\alpha}{\eta}(\boldsymbol{\theta} - \boldsymbol{\theta}_a)\right\|^2\right\}}{\sum_{\boldsymbol{x}\in\boldsymbol{\Theta}}\exp\left\{-\frac{1}{2\alpha}\left\|\boldsymbol{x} - \boldsymbol{\theta} + \alpha\nabla U(\boldsymbol{\theta}) - \frac{\alpha}{\eta}(\boldsymbol{\theta} - \boldsymbol{\theta}_a)\right\|^2\right\}}.\end{aligned}$$

The proof is similar to Theorem D.2. The key difference is that we can minorize on the entire set. Noting that

$$\left\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_a\left(1-\frac{\alpha_a}{\eta}\right) - \frac{\alpha_a}{\eta}\boldsymbol{\theta}\right\|^2 \leq 2\left\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_a\right\|^2 + 2\frac{\alpha_a^2}{\eta^2}\left\|\boldsymbol{\theta}_a - \boldsymbol{\theta}\right\|^2$$
$$\leq 2\,\mathrm{diam}(\boldsymbol{\Theta}_a)^2 + 2\frac{\alpha_a^2}{\eta^2}\,\Delta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_a)^2.$$

Using the same argument as Theorem D.2, we get a uniform minorization with

$$\begin{aligned}\bar{\eta}^* = &\frac{1}{\Phi_{\alpha_a}(\boldsymbol{\Theta}_a)}\exp\left\{-\frac{1}{\alpha_a}\,\mathrm{diam}(\boldsymbol{\Theta}_a)^2 - \frac{\alpha_a}{\eta^2}\,\Delta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_a)^2\right\}\\
&\times\frac{1}{|\boldsymbol{\Theta}|}\exp\left\{-\frac{1}{2\alpha}\left[\left((\alpha\,M+1)^2+\alpha\,M^2\right)\mathrm{diam}(\boldsymbol{\Theta})^2\right.\right.\\
&\quad+\left(2(M+\alpha)+2\alpha M\right)\|\nabla U(a)\|\,\mathrm{diam}(\boldsymbol{\Theta})\\
&\quad+\left(\alpha^2+\alpha\right)\|\nabla U(a)\|^2\\
&\quad\left.\left.+2\frac{\alpha}{\eta}\left[(\alpha\,M+1)^2\,\mathrm{diam}(\boldsymbol{\Theta})^2+2(M+\alpha)\|\nabla U(a)\|\,\mathrm{diam}(\boldsymbol{\Theta})+\alpha^2\|\nabla U(a)\|^2\right]^{1/2}\mathrm{diam}(\boldsymbol{\Theta})\right]\right\}.\end{aligned}$$

with the reference measure $\mu(\cdot)$ is the product measure of the Lebesgue measure and the counting measure.

Proof of Theorem 5.5. Using Lemma D.6 and Proposition 5.4, we further have

$$\|P^k(x,\cdot) - \tilde{\pi}\|_{TV} \leq (1-\bar{\eta}^*)^k + Z\exp\left(\frac{M}{4} - \frac{1}{2\alpha} + \frac{\Delta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_a)^2 - \vartheta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_a)}{2\eta}\right)$$

for all $x \in \mathbb{R}^d$ and $M(x), \rho$ is as defined in Theorem D.1 itself. Hence we are done.

Theorem D.7. *Let assumptions 5.1, 5.3 hold. Then, for any function $f : \mathbb{R}^p \to \mathbb{R}$ with $\|f\|_{\mathbb{L}^2_\pi} < \infty$, one has*

$$\sqrt{n}\left(\bar{f} - \mathbb{E}_{\pi_\gamma} f\right) \xrightarrow{d} N(0, \sigma_f^2)$$

as $n \to \infty$, where $\sigma_f^2 \in [0, \infty)$.

Proof. Using Theorem 5.5, the proof follows directly from Jones (2004).

D.4 Proofs for EDMALA

Proposition D.8. *For EDMALA(EDLP with MH step, refer Algorithm 1) the drift condition is satisfied with drift function $V(x_1, x_2) = \|x_1\|^2$.*

Proof. The proof follows from Theorem D.1 by observing that

$$PV(\theta_a, \theta) \leq \int \|\boldsymbol{\theta}_{a_1}\|^2 q((\boldsymbol{\theta}_a, \boldsymbol{\theta}), (\boldsymbol{\theta}_{a_1}, \boldsymbol{\theta}_1))d\boldsymbol{\theta}_{a_1} + 1$$
$$\leq \lambda V(\boldsymbol{\theta}_a, \boldsymbol{\theta}) + (b+1).$$

Lemma D.9. *Under Assumptions 5.1, 5.2, 5.3, and $\alpha < \frac{2}{M}$, for Markov chain P in Algorithm 1, we have for any $\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}' \in \widetilde{\boldsymbol{\Theta}}$,*

$$p(\widetilde{\boldsymbol{\theta}}|\widetilde{\boldsymbol{\theta}}') \geq \epsilon_\gamma \frac{\exp\left\{\frac{1}{2}U(\boldsymbol{\theta}')\right\}}{\sum_{x\in\boldsymbol{\Theta}}\exp\left(\frac{U(x)}{2}\right)} \cdot \frac{\exp\left\{-\frac{1}{2\alpha_a}\mathit{diam}(\boldsymbol{\Theta}_a)^2\right\}}{\Phi_{\alpha_a}(\boldsymbol{\Theta}_a)}$$

, where

$$\epsilon_\gamma = \exp\left\{\begin{array}{l} -\left(\frac{M}{2} + \frac{1}{\alpha} - \frac{m}{4}\right)\mathit{diam}(\boldsymbol{\Theta})^2 - \frac{1}{2}\|\nabla U(a)\|\,\mathit{diam}(\boldsymbol{\Theta}) \\ -\left(\frac{3\alpha_a}{8\eta^2} + \frac{2}{\eta}\right)\Delta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_a)^2 + \frac{\vartheta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_a)}{\eta}\end{array}\right\},$$

with $a \in \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\|\nabla U(\boldsymbol{\theta})\|$

Proof. We follow a similar minorization proof style as of Lemma 5.3 from Pynadath et al. (2024). Notice,

$$
\begin{aligned}
Z_\gamma(\widetilde{\boldsymbol{\theta}}) &\leq \frac{1}{\sqrt{2\pi\alpha_a}^d}\exp\left(-\frac{U(\boldsymbol{\theta})}{2}-\frac{\alpha_a}{8\eta^2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_a\|^2+\frac{1}{2\eta}\|\boldsymbol{\theta}-\boldsymbol{\theta}_a\|^2\right)\sum_{x\in\boldsymbol{\Theta}}\exp\left(\frac{U(x)}{2}\right)\\
&\int_y\sum_x\exp\left(-\frac{1}{2\alpha_a}\|y-\boldsymbol{\theta}_a\|^2-\frac{1}{2\eta}(\boldsymbol{\theta}-\boldsymbol{\theta}_a)^\top(x-y)\right)dy\\
&\leq\sum_{x\in\boldsymbol{\Theta}}\exp\left(\frac{U(x)}{2}\right)\exp\left(-\frac{U(\boldsymbol{\theta})}{2}+\frac{1}{2\eta}(\|\boldsymbol{\theta}-\boldsymbol{\theta}_a\|^2-\vartheta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a))\right)\\
&\leq\sum_{x\in\boldsymbol{\Theta}}\exp\left(\frac{U(x)}{2}\right)\exp\left(-\frac{U(\boldsymbol{\theta})}{2}+\frac{\Delta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)^2-\vartheta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)}{2\eta}\right)
\end{aligned}
$$

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

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

From this, one notes that

$$
\begin{aligned}
Z_\gamma(\widetilde{\boldsymbol{\theta}}) &\geq \frac{1}{\sqrt{2\pi\alpha_a}^d}\exp\left\{-\frac{U(\boldsymbol{\theta})}{2}-\frac{\alpha_a}{8\eta^2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_a\|^2+\frac{1}{2\eta}\|\boldsymbol{\theta}-\boldsymbol{\theta}_a\|^2\right\}\exp\left\{-\frac{1}{2}\left(\frac{1}{\alpha}-\frac{m}{2}\right)\text{ diam}(\boldsymbol{\Theta})^2\right\}\\
&\sum_{x\in\boldsymbol{\Theta}}\exp\left(\frac{U(x)}{2}\right)\int_y\sum_x\exp\left(-\frac{1}{2\alpha_a}\|y-\boldsymbol{\theta}_a\|^2-\frac{1}{2\eta}(\boldsymbol{\theta}-\boldsymbol{\theta}_a)^\top(x-y)\right)dy\\
&\geq\sum_{x\in\boldsymbol{\Theta}}\exp\left(\frac{U(x)}{2}\right)\exp\left\{-\frac{U(\boldsymbol{\theta})}{2}-\frac{\alpha_a}{8\eta^2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_a\|^2-\frac{1}{2}\left(\frac{1}{\alpha}-\frac{m}{2}\right)\text{ diam}(\boldsymbol{\Theta})^2-\frac{1}{2\eta}\Delta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)^2\right\}\\
&\geq\sum_{x\in\boldsymbol{\Theta}}\exp\left(\frac{U(x)}{2}\right)\exp\left\{-\frac{U(\boldsymbol{\theta})}{2}-\frac{\alpha_a}{8\eta^2}\Delta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)^2-\frac{1}{2}\left(\frac{1}{\alpha}-\frac{m}{2}\right)\text{ diam}(\boldsymbol{\Theta})^2-\frac{1}{2\eta}\Delta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)^2\right\}
\end{aligned}
$$

In other words,

$$\exp\left((-\frac{\alpha_a}{8\eta^2}-\frac{1}{2\eta})\Delta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)^2-\frac{1}{2}\left(\frac{1}{\alpha}-\frac{m}{2}\right)\text{ diam}(\boldsymbol{\Theta})^2\right)\leq\frac{Z_\gamma(\widetilde{\boldsymbol{\theta}})}{\sum_{x\in\boldsymbol{\Theta}}\exp\left(\frac{U(x)}{2}\right)\exp\left(-\frac{U(\boldsymbol{\theta})}{2}\right)}\leq\exp\left(\frac{\Delta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)^2-\vartheta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)}{2\eta}\right)$$

Consequently,

$$\frac{\frac{Z_\gamma(\widetilde{\boldsymbol{\theta}})}{\sum_{x\in\boldsymbol{\Theta}}\exp\left(\frac{U(x)}{2}\right)\exp\left(-\frac{U(\boldsymbol{\theta})}{2}\right)}}{\frac{Z_\gamma(\widetilde{\boldsymbol{\theta}'})}{\sum_{x\in\boldsymbol{\Theta}}\exp\left(\frac{U(x)}{2}\right)\exp\left(-\frac{U(\boldsymbol{\theta}')}{2}\right)}}\geq\frac{\exp\left((-\frac{\alpha_a}{8\eta^2}-\frac{1}{2\eta})\Delta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)^2-\frac{(2-m\alpha)\text{diam}(\boldsymbol{\Theta})^2}{4\alpha}\right)}{\exp\left(\frac{\Delta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)^2-\vartheta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)}{2\eta}\right)}$$

This implies

$$\frac{Z_\gamma(\widetilde{\boldsymbol{\theta}})}{Z_\gamma(\widetilde{\boldsymbol{\theta}'})}\geq\exp\left(\frac{1}{2}(-U(\boldsymbol{\theta})+U(\boldsymbol{\theta}'))\right)\frac{\exp\left((-\frac{\alpha_a}{8\eta^2}-\frac{1}{2\eta})\Delta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)^2-\frac{(2-m\alpha)\text{diam}(\boldsymbol{\Theta})^2}{4\alpha}\right)}{\exp\left(\frac{\Delta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)^2-\vartheta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)}{2\eta}\right)}$$

One notices from (9),

$$
\begin{aligned}
q_\gamma(\widetilde{\boldsymbol{\theta}'}|\widetilde{\boldsymbol{\theta}}) &= \frac{Z_\gamma(\widetilde{\boldsymbol{\theta}})^{-1}}{\sqrt{(2\pi\alpha_a)^d}}\exp\left(\frac{1}{2}(-U(\boldsymbol{\theta})+U(\boldsymbol{\theta}'))-(\boldsymbol{\theta}-\boldsymbol{\theta}')^\top\left(\frac{1}{2\alpha}I+\frac{1}{4}\int_0^1\nabla^2U((1-s)\boldsymbol{\theta}+s\boldsymbol{\theta}')ds\right)(\boldsymbol{\theta}-\boldsymbol{\theta}')\right.\\
&\left.-\frac{1}{2\eta}(\boldsymbol{\theta}-\boldsymbol{\theta}_a)^\top(\boldsymbol{\theta}'-\boldsymbol{\theta}'_a)-\frac{1}{2\alpha_a}\|\boldsymbol{\theta}'_a-\boldsymbol{\theta}_a\|^2+\frac{4\eta-\alpha_a}{8\eta^2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_a\|^2\right)\\
&\geq\frac{Z_\gamma(\widetilde{\boldsymbol{\theta}})^{-1}}{\sqrt{(2\pi\alpha_a)^d}}\exp\left(\frac{1}{2}\left\langle\nabla U(\boldsymbol{\theta}),\boldsymbol{\theta}'-\boldsymbol{\theta}\right\rangle-\frac{1}{2\alpha}\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|^2-\frac{1}{2\eta}(\boldsymbol{\theta}-\boldsymbol{\theta}_a)^\top(\boldsymbol{\theta}'-\boldsymbol{\theta}'_a)\right.\\
&\left.-\frac{1}{2\alpha_a}\|\boldsymbol{\theta}'_a-\boldsymbol{\theta}_a\|^2-\frac{\alpha_a}{8\eta^2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_a\|^2\right)
\end{aligned}
$$

We also note that

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

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

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

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

Combining, we get

$$\begin{aligned}
q_\gamma(\widetilde{\boldsymbol{\theta}}'|\widetilde{\boldsymbol{\theta}}) &\geq \frac{Z_\gamma(\widetilde{\boldsymbol{\theta}})^{-1}}{\sqrt{(2\pi\alpha_a)^d}} \exp\left\{(-\frac{M}{2} - \frac{1}{2\alpha})\text{diam}(\boldsymbol{\Theta})^2 - \frac{1}{2}\|\nabla U(a)\|\text{diam}(\boldsymbol{\Theta}) - \frac{1}{2\eta}(\boldsymbol{\theta} - \boldsymbol{\theta}_a)^\top(\boldsymbol{\theta}' - \boldsymbol{\theta}'_a) - \frac{1}{2\alpha_a}\|\boldsymbol{\theta}'_a - \boldsymbol{\theta}_a\|^2 - \frac{\alpha_a}{8\eta^2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2\right\} \\
&\geq \frac{\frac{1}{\sqrt{(2\pi\alpha_a)^d}} \exp\left\{(-\frac{M}{2} - \frac{1}{2\alpha})\text{diam}(\boldsymbol{\Theta})^2 - \frac{1}{2}\|\nabla U(a)\|\text{diam}(\boldsymbol{\Theta}) - \frac{1}{2\eta}(\boldsymbol{\theta} - \boldsymbol{\theta}_a)^\top(\boldsymbol{\theta}' - \boldsymbol{\theta}'_a) - \frac{1}{2\alpha_a}\|\boldsymbol{\theta}'_a - \boldsymbol{\theta}_a\|^2 - \frac{\alpha_a}{8\eta^2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2\right\}}{\sum_{x\in\boldsymbol{\Theta}} \exp\left(\frac{U(x)}{2}\right)\exp\left(-\frac{U(\boldsymbol{\theta})}{2} + \frac{\Delta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)^2 - \vartheta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)}{2\eta}\right)} \\
&\geq \frac{\exp\left\{-\frac{1}{2\alpha_a}\text{diam}(\boldsymbol{\Theta}_a)^2\right\}}{\Phi_{\alpha_a}(\boldsymbol{\Theta}_a)} \frac{\exp\left\{(-\frac{M}{2} - \frac{1}{2\alpha})\text{diam}(\boldsymbol{\Theta})^2 - \frac{1}{2}\|\nabla U(a)\|\text{diam}(\boldsymbol{\Theta}) + (-\frac{1}{2\eta} - \frac{\alpha_a}{8\eta^2})\Delta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)^2\right\}}{\sum_{x\in\boldsymbol{\Theta}} \exp\left(\frac{U(x)}{2}\right)\exp\left(-\frac{U(\boldsymbol{\theta})}{2} + \frac{\Delta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)^2 - \vartheta(\boldsymbol{\Theta},\boldsymbol{\Theta}_a)}{2\eta}\right)}
\end{aligned}$$

Acceptance Ratio,

$$\begin{aligned}
\rho(\widetilde{\boldsymbol{\theta}'} \mid \widetilde{\boldsymbol{\theta}}) &= \left(\frac{\pi(\widetilde{\boldsymbol{\theta}'})q_\gamma(\widetilde{\boldsymbol{\theta}} \mid \widetilde{\boldsymbol{\theta}'})}{\pi(\widetilde{\boldsymbol{\theta}})q_\gamma(\widetilde{\boldsymbol{\theta}'} \mid \widetilde{\boldsymbol{\theta}})}\right) \\
&= \exp\left\{U(\boldsymbol{\theta}') - U(\boldsymbol{\theta}) + \frac{1}{2\eta}(\|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2 - \|\boldsymbol{\theta}' - \boldsymbol{\theta}'_a\|^2)\right\}\frac{\tilde{Z}}{\tilde{Z}}\cdot \\
&\exp\left\{U(\boldsymbol{\theta}) - U(\boldsymbol{\theta}') - \frac{1}{2\eta}(\|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2 - \|\boldsymbol{\theta}' - \boldsymbol{\theta}'_a\|^2) - \frac{\alpha_a}{8\eta^2}(\|\boldsymbol{\theta}' - \boldsymbol{\theta}'_a\|^2 - \|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2)\right\}\frac{Z_\gamma(\widetilde{\boldsymbol{\theta}})}{Z_\gamma(\widetilde{\boldsymbol{\theta}'})} \\
&= \exp\left\{-\frac{\alpha_a}{8\eta^2}(\|\boldsymbol{\theta}' - \boldsymbol{\theta}'_a\|^2 - \|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2)\right\}\frac{Z_\gamma(\widetilde{\boldsymbol{\theta}})}{Z_\gamma(\widetilde{\boldsymbol{\theta}'})}
\end{aligned}$$

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

with Acceptance Probability

$$\mathcal{A}(\widetilde{\boldsymbol{\theta}'} \mid \widetilde{\boldsymbol{\theta}}) = \left(\rho(\widetilde{\boldsymbol{\theta}'} \mid \widetilde{\boldsymbol{\theta}}) \wedge 1\right)$$

and consider the transition kernel as

$$p(\widetilde{\boldsymbol{\theta}'} \mid \widetilde{\boldsymbol{\theta}}) = \left(\mathcal{A}(\widetilde{\boldsymbol{\theta}'} \mid \widetilde{\boldsymbol{\theta}})\right) q_\gamma(\widetilde{\boldsymbol{\theta}'} \mid \widetilde{\boldsymbol{\theta}}) + \left(1 - L(\widetilde{\boldsymbol{\theta}})\right)\delta_{\widetilde{\boldsymbol{\theta}}}(\widetilde{\boldsymbol{\theta}'})$$

where $\delta_{\widetilde{\boldsymbol{\theta}}}(\widetilde{\boldsymbol{\theta}'})$ is the Kronecker delta function and $L(\widetilde{\boldsymbol{\theta}})$ is the total acceptance probability from the point $\widetilde{\boldsymbol{\theta}}$ with

$$L(\widetilde{\boldsymbol{\theta}}) = \int_{\boldsymbol{\theta}'_a \in \boldsymbol{\Theta}_a} \sum_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \left(\rho([\boldsymbol{\theta}'^T, \boldsymbol{\theta}'^T_a]^T \mid \widetilde{\boldsymbol{\theta}}) \wedge 1\right) q_\gamma([\boldsymbol{\theta}'^T, \boldsymbol{\theta}'^T_a]^T|\widetilde{\boldsymbol{\theta}}) \quad d\boldsymbol{\theta}'_a$$

We note that

$$
\begin{aligned}
p(\widetilde{\boldsymbol{\theta}}' \mid \widetilde{\boldsymbol{\theta}}) &= \left(\mathcal{A}(\widetilde{\boldsymbol{\theta}}' \mid \widetilde{\boldsymbol{\theta}})\right) q_\gamma(\widetilde{\boldsymbol{\theta}}' \mid \widetilde{\boldsymbol{\theta}}) + \left(1 - L(\widetilde{\boldsymbol{\theta}})\right) \delta_{\widetilde{\boldsymbol{\theta}}}(\widetilde{\boldsymbol{\theta}}') \\
&\geq \left(\mathcal{A}(\widetilde{\boldsymbol{\theta}}' \mid \widetilde{\boldsymbol{\theta}})\right) q_\gamma(\widetilde{\boldsymbol{\theta}}' \mid \widetilde{\boldsymbol{\theta}}) \\
&= \left(\rho(\widetilde{\boldsymbol{\theta}}' \mid \widetilde{\boldsymbol{\theta}}) \wedge 1\right) q_\gamma(\widetilde{\boldsymbol{\theta}}' \mid \widetilde{\boldsymbol{\theta}}) \\
&= \exp\left\{-\frac{\alpha_a}{8\eta^2}\left(\|\boldsymbol{\theta}' - \boldsymbol{\theta}'_a\|^2 - \|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2\right)\right\} \frac{Z_\gamma(\widetilde{\boldsymbol{\theta}})}{Z_\gamma(\widetilde{\boldsymbol{\theta}}')} q_\gamma(\widetilde{\boldsymbol{\theta}}' \mid \widetilde{\boldsymbol{\theta}}) \\
&\geq \exp\left\{-\frac{\alpha_a}{8\eta^2}\|\boldsymbol{\theta}' - \boldsymbol{\theta}'_a\|^2\right\} \frac{Z_\gamma(\widetilde{\boldsymbol{\theta}})}{Z_\gamma(\widetilde{\boldsymbol{\theta}}')} q_\gamma(\widetilde{\boldsymbol{\theta}}' \mid \widetilde{\boldsymbol{\theta}}) \\
&\geq \exp\left\{-\frac{\alpha_a}{8\eta^2}\Delta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_a)^2 + \frac{1}{2}(-U(\boldsymbol{\theta}) + U(\boldsymbol{\theta}'))\right\} \frac{\exp\left(-\frac{\alpha_a}{8\eta^2} - \frac{1}{2\eta})\Delta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_a)^2 - \frac{(2-m\alpha)\text{diam}(\boldsymbol{\Theta})^2}{4\alpha}\right)}{\exp\left(\frac{\Delta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_a)^2 - \vartheta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_a)}{2\eta}\right)} q_\gamma(\widetilde{\boldsymbol{\theta}}' \mid \widetilde{\boldsymbol{\theta}}) \\
&\geq \exp\left\{-\frac{\alpha_a}{8\eta^2}\Delta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_a)^2 + \frac{1}{2}(-U(\boldsymbol{\theta}) + U(\boldsymbol{\theta}'))\right\} \frac{\exp\left(-\frac{\alpha_a}{8\eta^2} - \frac{1}{2\eta})\Delta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_a)^2 - \frac{(2-m\alpha)\text{diam}(\boldsymbol{\Theta})^2}{4\alpha}\right)}{\exp\left(\frac{\Delta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_a)^2 - \vartheta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_a)}{2\eta}\right)} \\
&\quad \cdot \frac{\exp\left\{-\frac{1}{2\alpha_a}\text{diam}(\boldsymbol{\Theta}_a)^2\right\}}{\Phi_{\alpha_a}(\boldsymbol{\Theta}_a)} \frac{\exp\left\{(-\frac{M}{2} - \frac{1}{2\alpha})\text{diam}(\boldsymbol{\Theta})^2 - \frac{1}{2}\|\nabla U(a)\|\text{diam}(\boldsymbol{\Theta}) + \left(-\frac{1}{2\eta} - \frac{\alpha_a}{8\eta^2}\right)\Delta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_a)^2\right\}}{\sum_{x \in \boldsymbol{\Theta}} \exp\left(\frac{U(x)}{2}\right) \exp\left(-\frac{U(\boldsymbol{\theta})}{2} + \frac{\Delta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_a)^2 - \vartheta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_a)}{2\eta}\right)} \\
&= \frac{\exp\left\{-\frac{1}{2\alpha_a}\text{diam}(\boldsymbol{\Theta}_a)^2\right\}}{\Phi_{\alpha_a}(\boldsymbol{\Theta}_a)} \frac{\exp\left\{\frac{1}{2}U(\boldsymbol{\theta}')\right\}}{\sum_{x \in \boldsymbol{\Theta}} \exp\left(\frac{U(x)}{2}\right)} \exp\left\{(-\frac{3\alpha_a}{8\eta^2} - \frac{2}{\eta})\Delta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_a)^2 + \frac{\vartheta(\boldsymbol{\Theta}, \boldsymbol{\Theta}_a)}{\eta}\right\} \\
&\quad \cdot \exp\left\{(-\frac{M}{2} - \frac{1}{\alpha} + \frac{m}{4})\text{diam}(\boldsymbol{\Theta})^2 - \frac{1}{2}\|\nabla U(a)\|\text{diam}(\boldsymbol{\Theta})\right\} \\
&= \epsilon_\gamma \frac{\exp\left\{\frac{1}{2}U(\boldsymbol{\theta}')\right\}}{\sum_{x \in \boldsymbol{\Theta}} \exp\left(\frac{U(x)}{2}\right)} \frac{\exp\left\{-\frac{1}{2\alpha_a}\text{diam}(\boldsymbol{\Theta}_a)^2\right\}}{\Phi_{\alpha_a}(\boldsymbol{\Theta}_a)}
\end{aligned}
$$

Proof. Proof follows from using Lemma $D.9$.

E Additional Experimental Results

E.1 4D Joint Bernoulli

To provide additional insights into the functionality of EDLP samplers, 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:

Flatness Diagnostics

Under the experimental setup outlined in Section 6, we present the true Eigenspectrum of the Hessian, derived from the discrete samples collected for EDULA, EDMALA, DULA, and DMALA (Figure 11).We manually tune the stepsizes for EDULA and EDMALA to 0.1 and 0.4 respectively. This visualization is inspired by Section 6.3 of (Li & Zhang, 2024), where diagonal Fisher information matrix approximation was used to plot the Eigenvalues. The alignment of the Eigenvalues closer to 0 indicates that the sampled data corresponds to a flatter curvature of the energy function.

EDMALA and EDULA, specifically designed with entropy-aware flatness optimization, exhibit eigenvalue distributions that are notably tighter and more concentrated around zero compared to their non-entropic counterparts, DMALA and DULA.

$$P_{\Theta}(\boldsymbol{\theta}) = \begin{cases} 0.07688 & \text{if } \boldsymbol{\theta} = 0000, \\ 0.04725 & \text{if } \boldsymbol{\theta} = 0001, \\ 0.12500 & \text{if } \boldsymbol{\theta} = 0010, \\ 0.01667 & \text{if } \boldsymbol{\theta} = 0011, \\ 0.08688 & \text{if } \boldsymbol{\theta} = 0100, \\ 0.07688 & \text{if } \boldsymbol{\theta} = 0101, \\ 0.07688 & \text{if } \boldsymbol{\theta} = 0110, \\ 0.16756 & \text{if } \boldsymbol{\theta} = 0111, \\ 0.04725 & \text{if } \boldsymbol{\theta} = 1000, \\ 0.05825 & \text{if } \boldsymbol{\theta} = 1001, \\ 0.01667 & \text{if } \boldsymbol{\theta} = 1010, \\ 0.04725 & \text{if } \boldsymbol{\theta} = 1011, \\ 0.07688 & \text{if } \boldsymbol{\theta} = 1100, \\ 0.04725 & \text{if } \boldsymbol{\theta} = 1101, \\ 0.01900 & \text{if } \boldsymbol{\theta} = 1110, \\ 0.01335 & \text{if } \boldsymbol{\theta} = 1111. \end{cases}$$

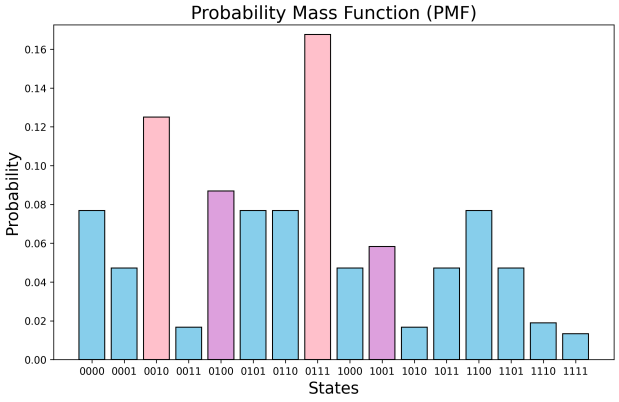


Figure 10: Target Distribution for 4D Joint Bernoulli

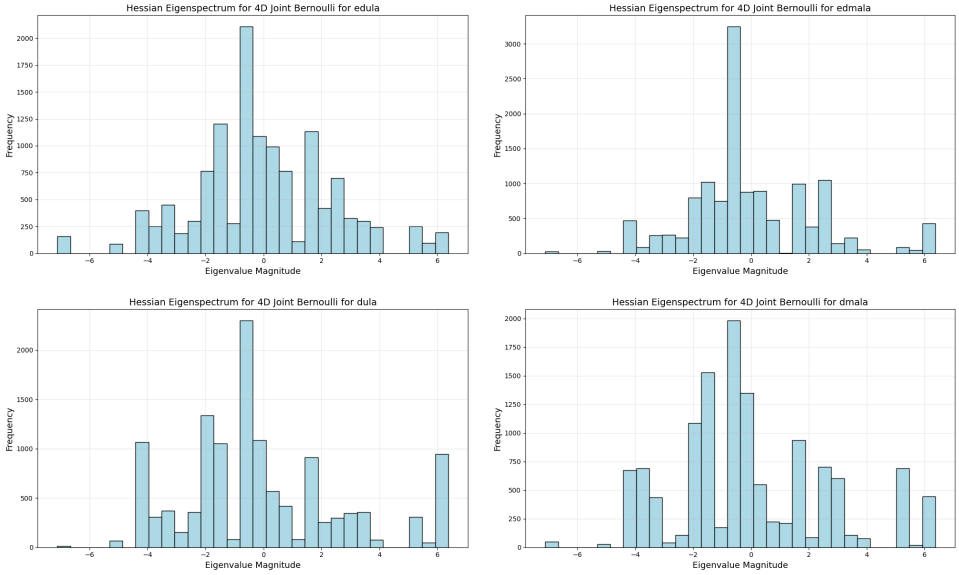


Figure 11: Eigenspectra of EDULA, EDMALA, DULA, and DMALA's performance on a Bernoulli distribution.

Quantitatively, EDULA demonstrates a lower spectral dispersion, evidenced by a lower standard deviation (std = 2.401) and narrower interquartile range (IQR = 3.031), relative to DULA (std = 2.832, IQR = 3.466). Similarly, EDMALA outperforms DMALA in terms of spectral concentration, achieving a standard deviation of 2.197 and IQR of 2.747, compared to DMALA's standard deviation of 2.700 and IQR of 3.224. Furthermore, visual inspection corroborates these quantitative findings; EDMALA and EDULA feature fewer extreme eigenvalues and outliers, reflecting biasing into sampling from flatter regions. Collectively, these results affirm that our entropy-guided methods (EDMALA, EDULA) effectively traverse flatter, aligning well with their intended design objectives.

E.2 TSP

Figure 12 presents the average PMC between solutions generated by each sampler, along with their standard deviations. DULA and EDULA exhibit nearly identical mean swap distances, whereas EDMALA demonstrates a notably lower mean swap distance compared to DMALA. This suggests

that the solutions proposed by EDMALA are structurally more similar, indicating a higher degree of consistency across its sampled solutions.

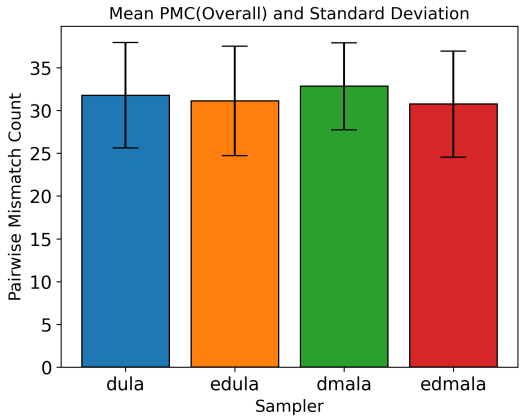


Figure 12: Variation in Solutions

Figure 13 showcases the performance characteristics of different samplers in terms of cost and solution diversity for the TSP. EDMALA and EDULA exhibit a narrower cost distribution, suggesting that they consistently identify solutions within a tighter range of costs. This stability implies a focused exploration within a particular solution quality band Camm & Evans (1997). In contrast, DMALA and DULA have a broader cost spread, indicating more variability in the quality of solutions they find.

When examining diversity in relation to the best solution, both DULA and DMALA maintain a similar spread, signifying comparable exploration depths relative to optimality. However, EDMALA stands out with a significantly smaller diversity spread compared to DMALA, indicating that EDMALA tends to produce solutions that are closer to the optimal path. This characteristic suggests that EDMALA is better suited for tasks requiring proximity to optimal solutions.

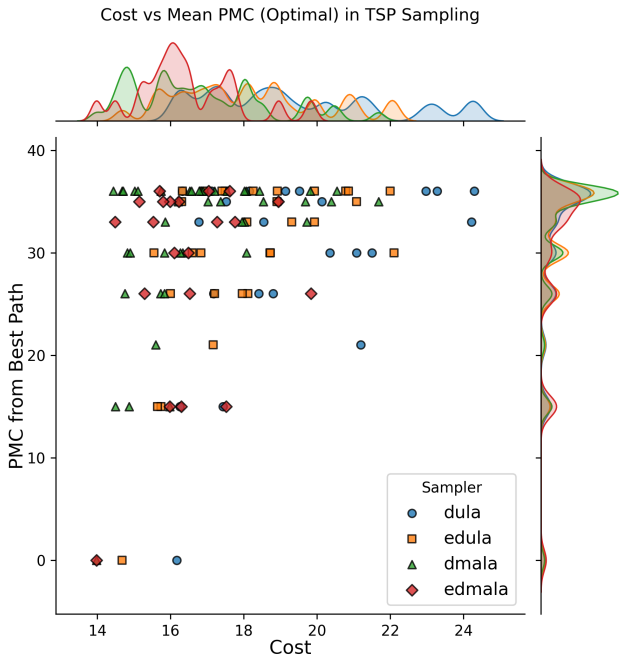


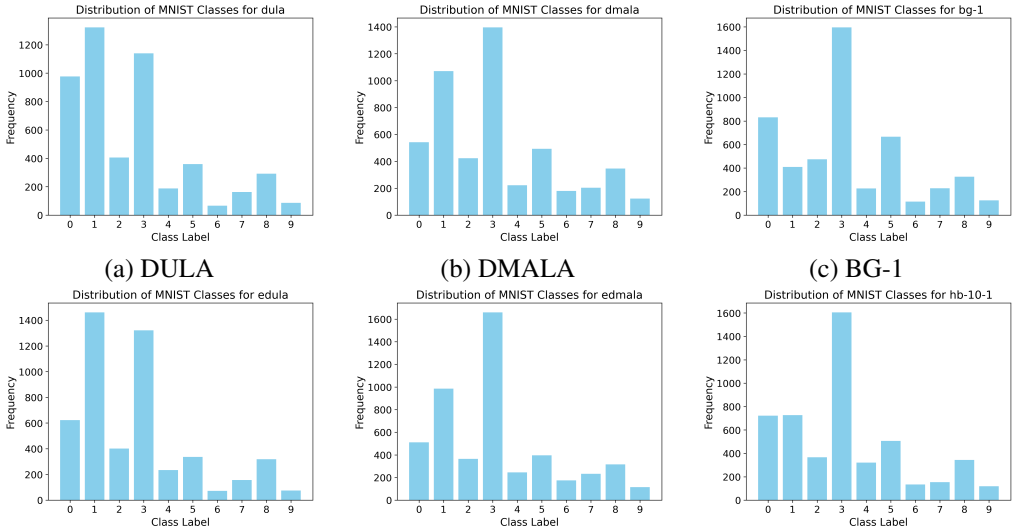
Figure 13: Marginal Plot

E.3 RBM

Mode Analysis

We performed mode analysis to validate the diversity and quality of MNIST digit samples generated by various samplers. Mode analysis assesses whether each sampler can capture the full range of MNIST digit classes (0-9) without falling into *mode collapse*, a phenomenon where a generative model fails to represent certain data modes, thus limiting diversity. We leveraged a *LeNet-5 convolutional neural network* LeCun et al. (1998) trained on MNIST to classify each generated sample and produce a class distribution for each sampler. The choice of LeNet-5, a reliable architecture for digit recognition, ensures accurate class predictions, thus providing a robust method to assess the representativeness of the samples. We train the model for 10 epochs, and achieve a 98.85% accuracy on test data.

The results(Figure 14) from our analysis indicated that all samplers produced samples across all digit classes, showing no evidence of mode collapse. Although certain samplers exhibited a preference for specific classes these biases did not reach the level of complete mode omission. Each class was represented in the generated samples, confirming that the samplers achieved an acceptable level of *mode diversity*. By confirming that all classes are covered, we demonstrate that each sampler can adequately approximate the diversity of the MNIST dataset, assuring the samples' representativeness Salimans et al. (2016); Goodfellow et al. (2014).



(a) DULA (b) DMALA (c) BG-1

(d) EDULA (e) EDMALA (f) HB-10-1

Figure 14: Mode Analysis

E.4 BBNN

We report the Average Training Log-Likelihood for our experiments in Table 3. Across all datasets, the EDLP samplers consistently outperform other samplers, demonstrating their ability to maintain or improve log-likelihood values. Importantly, when EDLP does not yield a substantial improvement, it still manages to avoid significantly impacting the training log-likelihood negatively.

Table 3: Average Training Log-Likelihood

Dataset	**Gibbs**	**GWG**	**DULA**	**DMALA**	**EDULA**	**EDMALA**
COMPAS	-0.3473 ±0.0337	-0.3304 ±0.0302	-0.3385 ±0.0101	-0.3149 ±0.0145	-0.3385 ±0.0110	**-0.3145** ±0.0149
News	-0.2156 ±0.0003	-0.2138 ±0.0010	-0.2101 ±0.0012	**-0.2097** ±0.0011	**-0.2097** ±0.0012	-0.2098 ±0.0012
Adult	-0.4310 ±0.0166	-0.3869 ±0.0325	-0.3044 ±0.0149	-0.2988 ±0.0158	-0.3032 ±0.0141	**-0.2987** ±0.0162
Blog	-0.4009 ±0.0072	-0.3414 ±0.0028	-0.2732 ±0.0128	-0.2705 ±0.0129	**-0.2699** ±0.0128	**-0.2699** ±0.0163

The computational burden associated with sampling can be a major bottleneck in scenarios requiring fast training and prediction, such as online systems or real-time applications. Such requirements are seen in financial modeling and stock market prediction, where models must adapt to real-time data to ensure accuracy Tsantekidis et al. (2017). Similarly, industrial IoT systems rely on real-time predictions to optimize maintenance and reduce downtime, where fast retraining is key Sun et al. (2017).

In Figure 15, we present the measured elapsed time per sample for the adult dataset to demonstrate these computational efficiencies, under the same settings as in Section 6, extending to include the GLU versions of the EDLP framework(Section B), alongside the results for the standard DLP and EDLP methods.

As illustrated, the EDLP versions exhibit an increase in runtime compared to DLP, due to the modifications discussed in Section 4.1. While the runtime difference between the DULA and EDULA algorithms (without MH correction) is negligible, the time difference between DMALA and EDMALA is more pronounced. This can be attributed to the more complex joint acceptance probability calculation required by EDMALA. Despite these variations, the overall runtime overhead for EDLP samplers is not substantial and remains practical.

For the EDLP-GLU variants, we maintained the same η and α values as their corresponding vanilla DLP samplers. The EDLP-GLU variants naturally achieve an approximate 50% reduction in runtime compared to EDLP. This efficiency stems from the alternating updates between sampling from a modified isotropic Gaussian and conditional DLP, designed to match the conditional distributions more effectively. However, this approach also introduces a higher standard deviation in runtime. The variability is primarily attributed to the contrasting computational costs between the two update types: sampling from the modified Gaussian is relatively lightweight, whereas the conditional DLP update is computationally intensive. As a result, the EDLP-GLU variants exhibit greater fluctuations in runtime compared to other samplers. Furthermore, the negative lower bounds are not physically meaningful and stem from the high variability in runtime measurements.

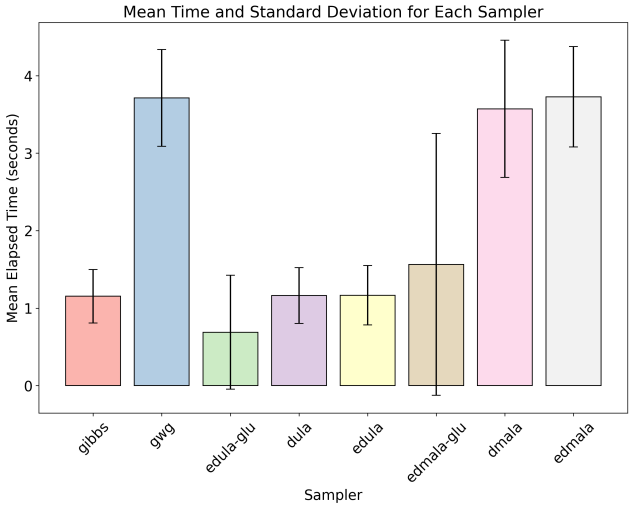


Figure 15: Runtime Analysis on Adult Dataset

For details of datasets used, refer to the Appendix of Zhang et al. (2022).

We fix α to 0.1 for DULA, DMALA, EDULA, and EDMALA. For more details on hyperparameters see Table 4.

All experiments in the paper were run on a single RTX A6000.

Table 4: Hyper-parameter Settings

Hyperparameters for EDLP				
Dataset	**EDULA**		**EDMALA**	
	α_a	η	α_a	η
COMPAS	0.0100	4.0	0.0010	4.0
News	0.0100	2.0	0.0001	0.8
Adult	0.0001	2.0	0.0001	4.0
Blog	0.0100	1.0	0.0001	1.0

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.