Diffusive Gibbs Sampling

Wenlin Chen [*1 2] **Mingtian Zhang** [*3] **Brooks Paige** [3] **José Miguel Hernández-Lobato** [1] **David Barber** [3]

Abstract

The inadequate mixing of conventional Markov Chain Monte Carlo (MCMC) methods for multi-modal distributions presents a significant challenge in practical applications such as Bayesian inference and molecular dynamics. Addressing this, we propose *Diffusive Gibbs Sampling (DiGS)*, an innovative family of sampling methods designed for effective sampling from distributions characterized by distant and disconnected modes. DiGS integrates recent developments in diffusion models, leveraging Gaussian convolution to create an auxiliary noisy distribution that bridges isolated modes in the original space and applying Gibbs sampling to alternately draw samples from both spaces. A novel *Metropolis-within-Gibbs scheme* is proposed to enhance mixing in the denoising sampling step. DiGS exhibits a better mixing property for sampling multi-modal distributions than state-of-the-art methods such as parallel tempering, attaining substantially improved performance across various tasks, including mixtures of Gaussians, Bayesian neural networks and molecular dynamics.

1. Introduction

Generating samples from complex unnormalized probability distributions is an important problem in machine learning, statistics and natural sciences. Consider an unnormalized target distribution of the form

$$p(x) = \frac{\exp(-E(x))}{Z}, \tag{1}$$

where $x \in \mathbb{R}^d$ is the variable of interest, $E : \mathbb{R}^d \to \mathbb{R}$ is a lower-bounded differentiable energy function, and $Z = \int \exp(-E(x))\,\mathrm{d}x$ is the (intractable) normalization constant. We aim to draw independent samples $x \sim p(x)$ from the target distribution and estimate expectations of functions $\mathbb{E}_{p(x)}[h(x)] = \int h(x)p(x)\,\mathrm{d}x$ under the target distribution $p(x)$.

The gradient of the log density of the target distribution is known as the *score function*:

$$\nabla_x \log p(x) = -\nabla_x E(x), \tag{2}$$

which is independent of Z. The score function can be evaluated at any location x since E is assumed to be differentiable. This assumption is commonly satisfied in various practical applications, such as posteriors in Bayesian inference (Welling & Teh, 2011), score/energy networks in generative image modelling (Song & Ermon, 2019), and Boltzmann distributions in statistical mechanics (Noé et al., 2019). Below, we introduce score-based methods for sampling from unnormalized distributions.

1.1. Score-Based MCMC Methods

Unadjusted Langevin Algorithm (ULA) (Grenander & Miller, 1994; Roberts & Tweedie, 1996) follows the transition rule given by a discrete-time Langevin SDE:

$$x_{k+1} = x_k + \eta \nabla_x \log p(x_k) + \sqrt{2\eta}\epsilon_k, \tag{3}$$

where $\epsilon_k \sim N(0, I)$. For an infinitesimal step size η, the Markov chain converges to the target $p(x)$ as $k \to \infty$.

Metropolis-adjusted Langevin Algorithm (MALA) (Roberts & Tweedie, 1996; Roberts & Stramer, 2002) defines a proposal x_{k+1} using the ULA update rule and additionally corrects the bias according to the transition probability given by the Metropolis-Hasting (MH) algorithm:

$$a_{\text{MALA}} \equiv \min\left\{1, \frac{\exp(-E(x_{k+1}))q(x_k|x_{k+1})}{\exp(-E(x_k))q(x_{k+1}|x_k)}\right\}, \tag{4}$$

where the proposal distribution is given by

$$q(x'|x) = \mathcal{N}(x'|x + \eta\nabla_x \log p(x), 2\eta I). \tag{5}$$

Hamiltonian Monte Carlo (HMC) (Duane et al., 1987; Neal et al., 2011) augments the original variable x with an auxiliary momentum variable v, which defines a joint distribution

$$p(x, v) = p(x)p(v) \propto e^{-E(x)-K(v)}, \tag{6}$$

[*]Equal contribution [1]University of Cambridge, Cambridge, UK [2]Max Planck Institute for Intelligent Systems, Tübingen, Germany [3]University College London, London, UK. Correspondence to: Wenlin Chen <wc337@cam.ac.uk>, Mingtian Zhang <m.zhang@cs.ucl.ac.uk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

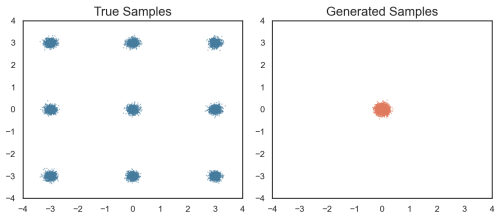


Figure 1. Challenge of multi-modal sampling with score-based MCMC. The true samples represent a mixture of 9 Gaussians and each Gaussian has a standard deviation $\sigma = 0.1$. The generated samples are produced by MALA initialized at the origin.

where $p(v) = \mathcal{N}(v|0, M)$ corresponds to the kinetic energy $K(v) = \frac{1}{2}v^{\mathrm{T}}M^{-1}v$. The total energy, or Hamiltonian, is denoted by $H(x, v) = E(x) + K(v)$. HMC generates samples of x and v by simulating the Hamiltonian equations

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{\partial H}{\partial v} = M^{-1}v, \quad \frac{\mathrm{d}v}{\mathrm{d}t} = -\frac{\partial H}{\partial x} = \nabla_x \log p(x). \quad (7)$$

Accurate numerical simulation of the Hamiltonian equations can be done by the leapfrog algorithm (Neal et al., 2011), with discretization bias corrected by the MH algorithm.

Despite the effort to address the challenge of exploring the whole support of a distribution using Langevin and Hamiltonian dynamics, MALA and HMC can still be ineffective for distributions with disconnected modes in practice since they often struggle to cross low-density barriers that separate the modes, leading to prolonged transitions from one mode to another (Pompe et al., 2020). More generally, score-based sampling methods utilize local gradient information to propose subsequent states, which presents challenges in sampling multi-modal distributions when there is insufficient *bridging density* to connect different modes. This limited connectivity impairs the ability of the resulting samples to accurately represent the entire distribution, a limitation illustrated in Figure 1.

1.2. Convolution-Based Method

A popular approach to bridging disconnected modes in multi-modal distributions is Gaussian convolution, which has been widely used in the recent developments of diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020). For a target distribution $p(x)$ and a Gaussian convolution kernel $p(\tilde{x}|x) = \mathcal{N}(\tilde{x}|\alpha x, \sigma^2 I)$, a convolved distribution $p(\tilde{x})$ can be constructed as follows:

$$p(\tilde{x}) = \int p(\tilde{x}|x)p(x)\,\mathrm{d}x. \quad (8)$$

Since $p(\tilde{x}|x)$ has the full support of $\mathbb{R}^d$, it can effectively create non-negligible density paths between disconnected modes in $p(\tilde{x})$, which makes the modes in $p(\tilde{x})$ exhibit better connectivity compared to those in $p(x)$. This cherished property has made Gaussian convolution a popular remedy to heal the blindness of score matching (Song & Ermon, 2019; Wenliang & Kanagawa, 2020; Zhang et al., 2022) and fix KL divergence training for distributions with disjoint or ill-defined density (Roth et al., 2017; Zhang et al., 2020; 2023b; Brown et al., 2022).

Due to the *mode-bridging* property of convolution, it is generally easier for score-based samplers to explore the whole space of $p(\tilde{x})$ than $p(x)$. If we could obtain numerous samples $\tilde{x} \sim p(\tilde{x})$, then it is more likely that these samples will encapsulate a broader range of modes in $p(\tilde{x})$ which are close to different high-density areas in $p(x)$. Consequently, these samples of $\tilde{x}$ can then serve as initial points for sampling from the original target $p(x)$, which facilitates score-based samplers in capturing different modes in $p(x)$. However, the score function of the convolved noisy distribution $p(\tilde{x})$ has the form

$$\nabla_{\tilde{x}} \log p(\tilde{x}) = \nabla_{\tilde{x}} \log \int \exp\left(-E(x) - \frac{\|\tilde{x} - \alpha x\|^2}{2\sigma^2}\right) \mathrm{d}x \quad (9)$$

which is typically intractable for non-Gaussian targets. This makes score-based sampling infeasible for $p(\tilde{x})$.

2. Diffusive Gibbs Sampling

We introduce a novel sampling method, **Di**ffusive **G**ibbs **S**ampling (*DiGS*), which leverages Gaussian convolution with an innovative *Metropolis-within-Gibbs scheme* to enhance multi-modal sampling, while avoiding the intractability of the convolved score function as shown in Equation 9.

2.1. Sampler Construction

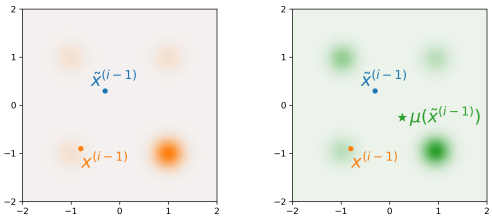
Instead of trying to directly produce samples from the intractable convolved distribution $p(\tilde{x})$, DiGS employs a Gibbs sampler to sample from the joint distribution $p(x, \tilde{x}) = p(\tilde{x}|x)p(x)$. This Gibbs sampling procedure involves alternately drawing samples from the two conditional distributions $p(\tilde{x}|x)$ (the convolution kernel) and $p(x|\tilde{x})$ (the denoising posterior). In each step, for a given clean sample $x^{(i-1)}$ from the target $p(x)$, we draw

1. a noisy sample $\tilde{x}^{(i-1)} \sim p(\tilde{x}|x = x^{(i-1)})$,
2. a new clean sample $x^{(i)} \sim p(x|\tilde{x} = \tilde{x}^{(i-1)})$.

For a Gaussian convolution kernel $p(\tilde{x}|x) = \mathcal{N}(\tilde{x}|\alpha x, \sigma^2 I)$, noisy samples can be easily obtained by corrupting clean samples with Gaussian noises:

$$\tilde{x} = \alpha x + \sigma\epsilon, \quad \epsilon \sim N(0, I), \quad (10)$$

where $\alpha \in [0, 1]$ is a contraction factor inspired by Diffusion models (Ho et al., 2020) and σ determines the level of smoothness in the Gaussian convolution. Intuitively, a small



(a) $p(x)$. (b) $p(x|\tilde{x}^{(i-1)})$.

Figure 2. Visualization of an MoG target with unequal weights $w = [0.1, 0.1, 0.1, 0.7]$ for different components. (a) Density heatmap of the target $p(x)$, a clean sample $x^{(i-1)}$ and a noisy sample $\tilde{x}^{(i-1)}$. (b) Density heatmap of the denoising posterior $p(x|\tilde{x}^{(i-1)})$ with Gaussian convolution parameters $\alpha = 1, \sigma = 1$.

α will compress the distribution and make the modes closer, and a large σ will encourage the sampler to jump out of the local modes. The effects of these hyperparameters will be investigated in Section 2.3, and hyperparameter tuning strategies will be further discussed in Section 2.4.

Unlike the convolved distribution $p(\tilde{x})$ which has an intractable score as shown in Equation 9, the denoising posterior $p(x|\tilde{x}) \propto p(x, \tilde{x}) = p(\tilde{x}|x)p(x)$, taking the form

$$p(x|\tilde{x}) \propto \exp\left(-E(x) - \frac{\|\alpha x - \tilde{x}\|^2}{2\sigma^2}\right), \tag{11}$$

has a tractable score function (Gao et al., 2020; Huang et al., 2023):

$$\nabla_x \log p(x|\tilde{x}) = -\nabla_x E(x) - \frac{\alpha\,(\alpha x - \tilde{x})}{\sigma^2}. \tag{12}$$

Therefore, common score-based methods like MALA, and HMC could be directly applied to sample from the denoising posterior $p(x|\tilde{x})$ in the denoising sampling step. It is worth noting that, in contrast to directly sampling from the original target $p(x) \propto \exp(-E(x))$ using score-based methods, incorporating an additional quadratic term as in Equation 11 improves the Log-Sobolev conditions (i.e., makes it more "Gaussian-like"), which in turn significantly increases the convergence speed of score-based samplers (Vempala & Wibisono, 2019); see Huang et al. (2023) for further in-depth analysis. Below, we show that the DiGS yields a $p(x, \tilde{x})$-irreducible and recurrent Markov Chain under certain regularity conditions.

Theorem 2.1. *For an absolutely continuous target distribution $p(x)$, DiGS with a Gaussian convolution kernel $p(\tilde{x}|x) = \mathcal{N}(\tilde{x}|\alpha x, \sigma^2 I)$ $(\alpha > 0, \sigma > 0)$ yields a $p(x, \tilde{x})$-irreducible and recurrent Markov Chain.*

The proof of Theorem 2.1 can be found in Appendix A. Intuitively, these properties ensure that the chain comprehensively explores the state space from any starting point (irreducibility) and effectively captures the target distribution by infinitely revisiting every state (recurrence) (Robert et al., 1999).

In addition to being a valid MCMC sampler, there are some practical considerations that can impact the performance of DiGS. We discuss these in the following sections.

2.2. Initialization of the Denoising Sampling Step

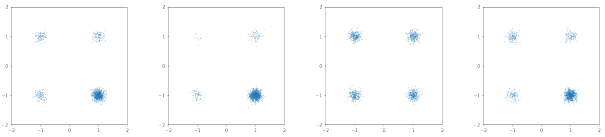
Ideally, one might hope to construct an exact Gibbs sampler where the score-based sampler targeting $p(x|\tilde{x}^{(i-1)})$ draws a true sample $x^{(i)}$ at each iteration. Unfortunately, when the target distribution has very disconnected modes, the resulting denoising posterior $p(x|\tilde{x})$ may still exhibit a multi-modal nature. For instance, Figure 2(a) illustrates the density of an MoG with unbalanced weights. For a given previous clean sample $x^{(i-1)}$, we generate a noisy sample $\tilde{x}^{(i-1)}$ using a Gaussian convolutional kernel with $\alpha = 1$ and $\sigma = 1$. The corresponding denoising density $p(x|\tilde{x}^{(i-1)})$ is depicted in Figure 2(b), which exhibits four distinct modes with varying weights. In such scenarios, selecting an appropriate initial point, $x^{(i)}_{init}$, for the subsequent sampling process $x^{(i)} \sim p(x|\tilde{x} = \tilde{x}^{(i-1)})$ is crucial for score-based samplers.

An ideal initial point for sampling from the denoising posterior $p(x|\tilde{x}^{(i-1)})$ would be the mean of the denoising distribution, defined as $\mu(\tilde{x}) \equiv \int x p(x|\tilde{x})\,\mathrm{d}x$. By Tweedie's lemma (Efron, 2011; Robbins, 1992), the mean can be expressed as a function of the noisy score $\nabla_{\tilde{x}} \log p(\tilde{x})$:

$$\mu(\tilde{x}^{(i-1)}) = \frac{\tilde{x}^{(i-1)} + \sigma^2 \nabla_{\tilde{x}} \log p(\tilde{x}^{(i-1)})}{\alpha}. \tag{13}$$

Figure 2(b) demonstrates that the mean function $\mu(\tilde{x}^{(i-1)})$ provides an initial point positioned in the middle of the four modes according to their weights, slightly favoring the mode with the largest weight. However, a challenge arises since $\nabla_{\tilde{x}} \log p(\tilde{x})$ is generally intractable (see Equation 9), making it impractical to compute.

Alternatively, a more straightforward approach is to initialize the denoising sampler at the previous state $x^{(i-1)}$. However, this strategy has a drawback: since $x^{(i-1)}$ is typically close to one of the modes, the score-based sampler often remains trapped in the vicinity of that mode, thereby hindering effective exploration of the entire distribution. Another heuristic initialization strategy is to use the (scaled) noisy sample $\tilde{x}^{(i-1)}/\alpha$, where the scaling factor α reflects the scale relationship between x and $\tilde{x}$ as suggested by the mean function in Equation 13. This approach exhibits a uniform preference for a random mode (for instance, the upper-left mode in Figure 2(b)) and ignores the underlying weighting between modes, resulting in a bias in representing the true weights of different modes and consequently diminishing the overall quality of the samples.



(a) True. (b) $x^{(i-1)}$. (c) $\tilde{x}^{(i-1)}/\alpha$. (d) *MH*.

Figure 3. Comparison of different initialization techniques for denoising posterior sampling on an unequally weighted MoG target described in Figure 2(a). In each case, we generate 1,000 samples using a Gaussian convolution kernel with $\alpha = 1, \sigma = 1$.

Table 1. MMD between true samples and samples obtained using different initialization techniques for denoising posterior sampling on an unequally weighted MoG target described in Figure 2(a).

Init.	$x^{(i-1)}$	$\tilde{x}^{(i-1)}/\alpha$	*MH*
MMD	0.15 ± 0.01	0.92 ± 0.04	$\mathbf{0.03 \pm 0.01}$

Algorithm 1 Diffusive Gibbs Sampling (DiGS)

Input: target energy $E(x)$; Gaussian convolution hyperparameters α, σ; score-based denoising sampler $\mathcal{S}$; the number of denoising sampling steps L; the number of Gibbs sampling sweeps K; initial clean sample $x^{(0)}$.
for $i \leftarrow 1$ to K **do**
  Draw $\tilde{x}^{(i-1)} \sim p(\tilde{x}|x^{(i-1)})$ using Equation 10.
  Propose $x'_{init} \sim q(x|\tilde{x}^{(i-1)})$ as initialization for the denoising process from the proposal in Equation 14.
  Accept $x^{(i)}_{init} \leftarrow x'_{init}$ with probability a_{init} in Equation 15; otherwise set $x^{(i)}_{init} \leftarrow x^{(i-1)}$.
  Draw $x^{(i)} \sim p(x|\tilde{x}^{(i-1)})$ by running the score-based sampler $\mathcal{S}$ for L steps from the initial point $x^{(i)}_{init}$ using Equation 12.
end for
Output: $x^{(K)}$

2.2.1. A Metropolis-within-Gibbs Scheme

To avoid potential bias, we propose a *Metropolis-within-Gibbs scheme* to facilitate mixing across modes by sampling an initialization position x'_{init} from an additional MCMC transition kernel for the subsequent finite score-based sampling steps. Specifically, to initialize score-based sampling from the denoising posterior $p(x|\tilde{x}^{(i-1)})$, we employ a linear Gaussian proposal $x'_{init} \sim q(x|\tilde{x}^{(i-1)})$ which is centered at the (scaled) noisy sample $\tilde{x}^{(i-1)}/\alpha$:

$$q(x|\tilde{x}^{(i-1)}) = \mathcal{N}(x|\tilde{x}^{(i-1)}/\alpha, (\sigma/\alpha)^2 I), \tag{14}$$

where the mean and the variance are chosen according to $q(x|\tilde{x}) \propto p(\tilde{x}|x)$ inspired by the mean function $\mu(\tilde{x}^{(i-1)})$. Note that this proposal only depends on the noisy sample $\tilde{x}^{(i-1)}$ and is independent of the previous state $x^{(i-1)}$. We use the MH algorithm to calculate the acceptance rate for this proposal:

$$a_{init} = \min\left(1, \frac{p(x'_{init}|\tilde{x}^{(i-1)})q(x^{(i-1)}|\tilde{x}^{(i-1)})}{p(x^{(i-1)}|\tilde{x}^{(i-1)})q(x'_{init}|\tilde{x}^{(i-1)})}\right), \tag{15}$$

where the denoising posterior ratio is tractable since

$$\frac{p(x'_{init}|\tilde{x}^{(i-1)}))}{p(x^{(i-1)}|\tilde{x}^{(i-1)})} = \frac{e^{-E(x'_{init})}p(\tilde{x}^{(i-1)}|x'_{init})}{e^{-E(x^{(i-1)})}p(\tilde{x}^{(i-1)}|x^{(i-1)})}. \tag{16}$$

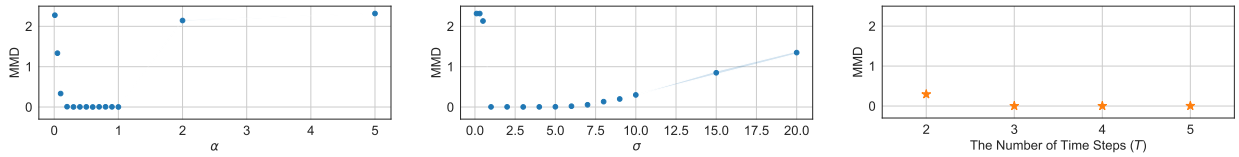
If the proposal is accepted, we will initialize the denoising sampling process with the updated value x'_{init}, rather than the previous state $x^{(i-1)}$. Algorithm 1 summarizes the proposed Diffusion Gibbs Sampling (DiGS) procedure.

To demonstrate the benefits of the additional MCMC kernel updating the initialization, we run DiGS with three different initializations (the previous state $x^{(i-1)}$, a heuristic re-initialization at $\tilde{x}^{(i-1)}/\alpha$, and our MH transition strategy) on an MoG target with different component weights. Figure 3 provides a visual comparison of the samples obtained from these initializations, showing that only the MH transition scheme captures all modes with the correct weightings. For evaluation, we employ the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) throughout all the experiments. The results are shown in Table 1, which demonstrates that the MH strategy outperforms the other two in capturing all modes and accurately representing the true weightings. Detailed experimental setup can be found in Appendix D.1.

2.3. Choosing the Gaussian Convolution Kernels

The performance of DiGS can be influenced by the hyperparameters α and σ in the Gaussian convolution kernel. Intuitively, for a given fixed α, a large σ will enhance the exploration capability of the sampler. However, this also makes the denoising posterior $p(x|\tilde{x})$ closer to $p(x)$ as illustrated in Equation 11, thereby elevating the complexity of denoising sampling. Similarly, with a fixed σ, reducing α will bring the modes closer but will also make $\nabla_x \log p(x|\tilde{x})$ close to $\nabla_x \log p(x)$ as shown in Equation 12, making the denoising sampling challenging. To illustrate the effects of hyperparameters, we apply DiGS to the MoG problem described in Figure 1 with varying values of α, σ and show the results in Figures 4(a) and 4(b), respectively. This demonstrates that within a specific range of α and σ values, DiGS consistently achieves optimal performance indicated by almost zero MMD. However, the quality of samples degrades when hyperparameters deviate beyond certain thresholds.

Although these results suggest that DiGS is robust to a certain range of hyperparameters, selecting an appropriate range remains crucial for each specific target distribution. The optimal range can depend on the support and the shape of the target density, which is typically unknown in practice.



(a) Effects of α (fix $\sigma = 1$). (b) Effects of σ (fix $\alpha = 1$). (c) Effects of T in the VP schedule.

Figure 4. Effects of the hyperparameters α, σ in Gaussian convolution kernels and the number T of noise levels in the variance-preserving (VP) noise scheduling. The y-axis in all three plots is the MMD between true samples and generated samples generated by DiGS with varying hyperparameters. Experimental setups can be found in Appendices D.2 and D.3.

Next section will introduce a multi-level noise scheduling to mitigate the need for precise hyperparameter selection.

2.4. Multi-Level Noise Scheduling

Drawing inspiration from the noise scheduling technique used in diffusion models (Ho et al., 2020; Song & Ermon, 2019; Song et al., 2021; Gao et al., 2018), we propose to use a sequence of Gaussian convolution kernels with $0 < \alpha_T < \cdots < \alpha_1 < 1$ and the corresponding variance $\sigma_t^2 = 1 - \alpha_t^2$ in each Gaussian convolution kernel $p_t(\tilde{x}_t|x) = \mathcal{N}(\tilde{x}|\alpha_t x, \sigma_t^2 I)$. This approach is commonly known as the *variance-preserving* (VP) schedule, where $p_t(\tilde{x}_t) = \int p_t(\tilde{x}_t|x)p(x)\,\mathrm{d}x$ and $p_0(\tilde{x}_0) \equiv p(x)$. As $\alpha_T \to 0$, it follows that $p_T(\tilde{x}_T|x) \to \mathcal{N}(\tilde{x}_T|0, I)$ and $p_T(\tilde{x}_T) \to \mathcal{N}(\tilde{x}_T|0, I)$, which are easy to sample from since they are independent of the characteristics of the target distribution $p(x)$.

We follow the common sampling procedure in the score-based diffusion model literature (Song & Ermon, 2019; 2020) and propose the following sampling procedure: for each t from T to 1, we run DiGS to generate a sample, which is used as an initial point in the subsequent time step $t-1$. For a given number of noise levels T, we apply a simple linear scheduling scheme to determine the values of α_t. Specifically, given the end points α_1 and α_T, we have

$$\alpha_t = \alpha_T + (\alpha_1 - \alpha_T)\frac{T-t}{T-1}, \quad \sigma_t = \sqrt{1-\alpha_t^2}. \tag{17}$$

Despite the similarity in noise scheduling to diffusion models, the fundamental sampling mechanism in DiGS is different. In diffusion models, the sampling process requires a progression from time step T to 0 to yield valid samples. However, DiGS produces a valid sample at any timestep t in principle. This property allows us to set $\alpha_T > 0$ and $\alpha_1 < 1$, thereby enhancing the efficiency of DiGS without the necessity to align with the asymptotic distributions. To illustrate the effect of the number of noise levels T, we implement the VP schedule with a linear noise scheme. We set $\alpha_T = 0.1, \alpha_1 = 0.9$ and vary T from 2 to 5. We test this multi-level DiGS on the MoG problem described in Figure 1. Figure 4(c) shows that an optimal sampler is achieved with $T > 2$ for this problem, circumventing the need for manually selecting Gaussian convolution hyperparameters.

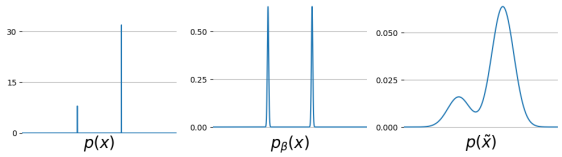


Figure 5. Comparison of a multi-modal target distribution $p(x)$, tempered distribution $p_\beta(x)$, and convolved distribution $p(\tilde{x})$.

3. Comparison to Related Methods

In this section, we explore the relationship between DiGS and related methods, complementing this with empirical comparisons to highlight their distinct characteristics.

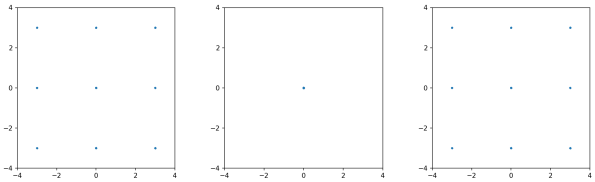
3.1. Tempering-Based Sampling

Tempering-based sampling is a state-of-the-art method for multi-modal target distributions, which samples from smoothed versions of the target distribution and exchanges samples with the original target once in a while. A tempered target distribution is defined as

$$p_\beta(x) \propto p(x)^\beta \propto \exp(-\beta E(x)), \tag{18}$$

where $\beta \equiv 1/\tau < 1$ is the inverse temperature. As $\tau \to \infty$, the tempered target $p_{\beta\to 0}(x)$ converges to a flat distribution, which encourages transitions among different modes. Tempering is a key building block to develop state-of-the-art multi-modal sampling methods such as parallel tempering (PT) (Swendsen & Wang, 1986; Geyer & Thompson, 1995; Syed et al., 2022; Surjanovic et al., 2022) and annealed importance sampling (AIS) (Neal, 2001).

Tempering-based sampling makes transitions between distant modes easier. However, tempering is unable to bridge disconnected modes even with a large temperature as it does not alter the support of a distribution, since $p(x)^\beta = 0$ wherever $p(x) = 0$. Moreover, Figure 5 shows that the tempering-based method is inefficient in connecting modes



(a) True Samples (b) PT (c) *DiGS*

Figure 6. Comparison between parallel tempering (PT) and DiGS on a "Mixture of Deltas" problem in 2D with the same budge. PT consists of 5 temperatures, while DiGS uses one noise level. Each sampler is initialized at the origin and generates 1,000 samples. Detailed experimental setup can be found in Appendix D.4.

even when they are not completely disconnected but far away. Appendix B gives an analytical example, showing the log-density of a point between two modes could approach $-\infty$ in the tempered distribution $p_\beta(x)$, whereas the log-density of that point in the convolved distribution $p(\tilde{x})$ is lower bounded, regardless of the shape of the target $p(x)$.

To empirically compare the mode-bridging property of PT and DiGS, we conduct experiments on an extreme problem, "Mixture of Deltas", which is an MoG with extremely small variance in each Gaussian. Figure 6 shows the comparison of the samples from these two samplers with the same budget: PT is unable to escape the central mode in this extreme setting even with large temperatures τ, whereas DiGS manages to recover all 9 modes, demonstrating its superiority over PT in terms of mode exploration and coverage. To make PT work on this problem, one needs to use more chains with the DEO scheme (Deng et al., 2023); see Appendix D.4 for a detailed description.

3.2. Score-Based Diffusion Model

The convolution technique plays a key role in score-based diffusion models (Song & Ermon, 2019; Song et al., 2021; Song & Ermon, 2020; Ho et al., 2020; Sohl-Dickstein et al., 2015; Gao et al., 2020). These models employ a series of convolutions (indexed by t) to form a distribution sequence from $p(\tilde{x}_0) \equiv p(x)$ to a simple distribution $p(\tilde{x}_T)$. At each time step t, the score function at time $\nabla_{\tilde{x}_t} \log p(\tilde{x}_t)$ is learned from samples $x_1, \cdots, x_N \sim p(x)$ using denoising score matching (Vincent, 2011). To sample from $p(x)$, ULA and Euler discretization are typically applied in reverse ($t = T \to 0$), using samples from each time step t as the initial point at time $t-1$. This scheme has shown state-of-the-art generation quality for complex data like images. However, unlike score-based/diffusion generative models where $\nabla_{\tilde{x}} \log p(\tilde{x})$ is learned directly from samples, our setting only assumes access to an energy function $E(x)$ *without* samples, making the noisy marginal score $\nabla_{\tilde{x}} \log p(\tilde{x})$ intractable, as shown in Equation 9.

Zhang et al. (2023a) proposes a pseudo-Gibbs sampler for the joint $p(x, \tilde{x}) = p(x)p(\tilde{x}|x)$ to obtain clean samples from score-based diffusion models, where $p(\tilde{x}|x) = \mathcal{N}(\tilde{x}|\alpha x, \sigma^2 I)$ and $p(x|\tilde{x})$ is approximated by a full-covariance Gaussian. This approach also requires training data to estimate the score function $\nabla_{\tilde{x}} \log p(\tilde{x})$, which is different from our problem setting.

3.3. Proximal Sampler

DiGS also belong to the proximal sampler family (Chen et al., 2022; Lee et al., 2021), where Gibbs sampling is executed between the Gaussian convolution distribution $p(\tilde{x}|x)$ and the denoising posterior $p(x|\tilde{x})$. In the denoising sampling step, proximal samplers find a mode using gradient-based optimization technique and perform rejection sampling around that mode. We highlight several significant improvements of DiGS over the proximal samplers. First, we employ a score-based sampler such as MALA and HMC in the denoising step instead of rejection sampling around a single mode as in proximal samplers, which helps the method scale to higher-dimensional distributions. Second, we employ an innovative *Metropolis-within-Gibbs scheme* to initialize the denoising sampling step, demonstrating its importance in achieving correct density allocation in multi-modal situations, as discussed in Section 2.2. Third, we emphasize the importance of tuning Gaussian convolution kernel in Section 2.4 and employ a multi-level scheduling to eliminate the need of choosing Gaussian convolution hyper-parameters. These improvements make the DiGS applicable to practical problems with multi-modal target distributions in high dimensional spaces.

Entropy-MCMC (Li & Zhang, 2023) is designed to sample from the flat regions within the posterior of the Bayesian neural network $p(\theta|\mathcal{D})$, where $\mathcal{D}$ represents the training dataset. Specifically, the approach begins by defining a surrogate distribution using Gaussian convolution:

$$p(\theta_a|\mathcal{D}) = \int p(\theta_a|\theta)p(\theta|\mathcal{D})\,\mathrm{d}\theta, \tag{19}$$

where $p(\theta_a|\theta) \propto \exp\left(-\frac{1}{2\eta}\|\theta_a - \theta\|_2^2\right)$ is a Gaussian kernel aimed at *smoothing out* the sharp regions in the posterior. Subsequently, SGLD (Welling & Teh, 2011) is applied to sample within the joint space of (θ, θ_a), where samples of both θ and θ_a are kept. It is crucial to highlight that, while this method aims to seek flat modes in the target distribution $p(\theta|\mathcal{D})$, similar to the approaches described in Chaudhari et al. (2019); Staines & Barber (2012), it is not intended and cannot provide exact samples from the target distribution $p(\theta|\mathcal{D})$ due to its optimization nature. In contrast, DiGS is designed as a valid MCMC sampler that is specifically designed to identify all modes (including both sharp and flat modes) and their corresponding density allocations in the target distribution, thus pursuing a distinct objective.

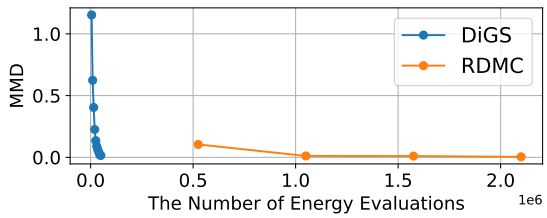


Figure 7. Computational cost comparison between DiGS and RDMC. The x-axis is the number of energy evaluations ($\times 10^6$) and the y-axis is the MMD between true and generated samples. We use DiGS with a single noise level ($\alpha = 1, \sigma = 1$) and run Gibbs sampling for 1-10 sweeps, represented by the 10 blue scatters. The RDMC is experimented with $T \in \{1, 2, 3, 4\}$, represented by the four orange points in the plot. Detailed experimental setup can be found in Appendix D.5.

3.4. Reverse Diffusion Monte Carlo

To emulate a diffusion model sampling process with access to the unnormalized density of the target only, Huang et al. (2023) introduces Reverse Diffusion Monte Carlo (RDMC), which approximates the score function in the noisy space at each time t. Specifically, by rewriting the Tweedie's lemma in Equation 13, the noisy score can be expressed as

$$\nabla_{x_t} \log p(x_t) = (\alpha\mu(x_t) - x_t)/\sigma_t^2, \tag{20}$$

where the denoising posterior mean $\mu(x_t) = \int x p(x|x_t) \approx \frac{1}{K}\sum_{k=1}^{K} x_0^{(t,k)}$ is approximated by the Monte Carlo method, with samples $x_0^{(t,1)}, \cdots, x_0^{(t,K)}$ obtained by running a score-based sampler such as ULA targeting the posterior $p(x_0|x_t) \propto \exp(-E(x_0) - \|x_t - \alpha_t x_0\|^2/2\sigma_t^2)$. Therefore, in RDMC, obtaining a single sample $x_t \sim p(x_t)$ via ULA at each time step $t \in [1, T]$ involves an intermediate step of generating K posterior samples via another ULA, leading to a *nested* MCMC sampling procedure that imposes substantial computational demands. In contrast, DiGS does not have such hierarchy since it requires only one MCMC chain during the denoising sampling step in each Gibbs sweep, significantly reducing the computational burden. Figure 7 compares the computational costs of these two methods when applied to MoG target as shown in Figure 1, demonstrating that DiGS can achieve the same accuracy as RDMC with $10\times$ fewer energy evaluations. The efficiency of DiGS is particularly vital in applications such as molecular configuration sampling (Section 4.3), where even a single energy evaluation is costly.

3.5. Auxiliary Variable MCMC

DiGS belongs to the broader auxiliary variable MCMC family. This family encompasses various notable methods such as the Swendsen-Wang algorithm (Swendsen & Wang, 1987), slice sampling (Neal, 2003), Hamiltonian Monte Carlo (HMC) (Duane et al., 1987; Neal et al., 2011), and auxiliary variational MCMC (Habib & Barber, 2018; Agakov & Barber, 2004); see Barber (2012) for a detailed introduction. Among these, HMC bears the closest resemblance to DiGS. We delve into the similarities and differences below.

As introduced in Section 1.1, HMC generates samples from the joint distribution $p(x, v) = p(x)p(v)$ where the v is the auxiliary variable that represents the momentum. The momentum is usually distributed as a Gaussian $p(v) = \mathcal{N}(v|0, \sigma_v^2 I)$ that is independent of x. However, there are other variants of HMC where the momentum auxiliary variable v depends on x. For example, in Riemannian Manifold HMC (Girolami & Calderhead, 2011), the momentum is distributed as $p(v|x) = \mathcal{N}(v|0, \Sigma_v(x))$, where $\Sigma_v(x)$ is the Fisher information matrix that captures the local curvature of the energy around x. The auxiliary variable in DiGS is the convolved variable $\tilde{x}$, which is a noisy version of x, given by $p(\tilde{x}|x) = \mathcal{N}(\tilde{x}|\alpha x, \sigma^2 I)$. Notably, when $\alpha = 0$, it follows that $p(\tilde{x}|x) = p(\tilde{x}) = \mathcal{N}(\tilde{x}|0, \sigma^2 I)$, and the joint distribution $p(x, \tilde{x}) = p(x)p(\tilde{x})$ in DiGS is identical to $p(x, v)$ in the classic HMC formulation. Despite this similarity, DiGS employs Gibbs sampling which alternately samples from the two conditionals $p(\tilde{x}|x)$ and $p(x|\tilde{x})$, whereas HMC simulates the Hamiltonian equations as discussed in Section 1.1, interleaved with samples from $p(v)$.

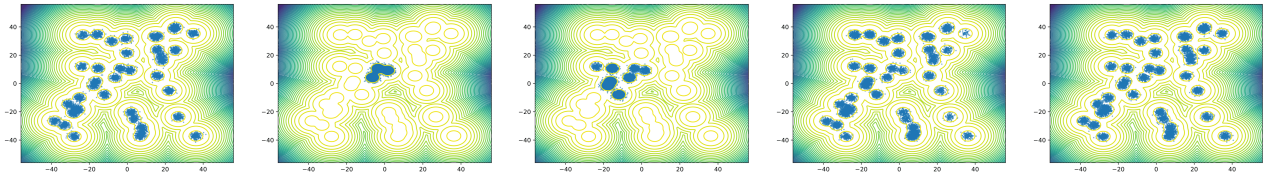
Table 2. Sample quality on MoG-40. MMD is computed between true samples and samples generated by each sampler. MAE is computed between the true and estimated expectation values of a quadratic function under the target and is expressed as the percentage of the true expectation value.

Sampler	MMD	MAE (%)	#energy eval.
MALA	1.73 ± 0.12	93.3 ± 0.73	1.0×10^7
HMC	1.70 ± 0.09	92.8 ± 0.34	1.0×10^7
PT	$(1.89 \pm 0.44) \times 10^{-2}$	7.32 ± 1.85	1.0×10^7
DiGS	$(\mathbf{4.57 \pm 1.10}) \times \mathbf{10^{-4}}$	$\mathbf{0.75 \pm 0.19}$	1.0×10^7

4. Empirical Evaluation

We evaluate DiGS on three complex multi-modal sampling tasks across various domains[1]: a mixture of 40 Gaussians, Bayesian neural network, and molecular dynamics. For DiGS, we employ MALA with the Metropolis-within-Gibbs scheme. We compare DiGS with three baselines: MALA, HMC and PT. In all experiments, the step sizes of MALA and HMC are tuned via trial-and-error so that the acceptance rates are close to 0.574 (Roberts & Rosenthal, 1998) and 0.65 (Neal et al., 2011), respectively. We choose not to compare with RDMC, since it is computationally intractable on these complex tasks as demonstrated in Section 3.4.

[1]The code of our experiments can be found in https://github.com/Wenlin-Chen/DiGS.



(a) True Samples (b) MALA (c) HMC (d) PT (e) *DiGS*

Figure 8. Visualization of 10^4 samples for MoG-40 generated by each sampler. All samplers are initialized at the origin.

4.1. Mixture of 40 Gaussians

We first consider a synthetic problem from Midgley et al. (2023), which is a 2D MoG with 40 mixture components. This is a relatively challenging multi-modal sampling task, yet it allows for visual examination of the mode-coverage property for each method. In this experiment, each method is initialized at the origin and generates 10^4 samples for evaluation. MALA runs 1,000 Langevin steps per sample. HMC runs 1,000 leapfrog steps per sample. PT consists of 5 chains with temperatures $\tau = \{1.0, 5.62, 31.62, 177.83, 1000.0\}$, where each chain is constructed by an HMC sampler with 200 leapfrog steps per sample. DiGS uses $T = 1$ noise level with $\alpha = 0.1$ and $\sigma^2 = 1 - \alpha^2$, 200 Gibbs sweeps, and 5 MALA denoising sampling steps per Gibbs sweep.

Figure 8 shows a visual comparison for 10^4 samples generated by each method. We can see that MALA and HMC fail to explore the modes that are far away from the origin. PT covers all 40 modes but produces significantly less samples for the modes on the top-right and bottom-right corners. DiGS manages to cover all modes with the right amount of samples in each mode. Table 2 shows the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) (computed with 5 kernels with bandwidths $\{2^{-2}, 2^{-1}, 2^0, 2^1, 2^2\}$) between the true samples and samples generated by each sampler and the Mean Absolute Error (MAE) between the true and estimated expectations of a quadratic function under the MoG-40 target. This demonstrates that our method significantly outperforms all baselines on this problem.

4.2. Bayesian Neural Networks

The posterior density of the parameters in a Bayesian neural network (BNN) is known to be complex and multi-modal (Barber & Bishop, 1998; Hernández-Lobato & Adams, 2015; Louizos & Welling, 2017; Izmailov et al., 2018). For a given training dataset $D_{\text{train}} = \{(x_i, y_i)\}_{i=1}^N$, the posterior density can be expressed as $p(\theta|D_{\text{train}}) \propto p(\theta) \prod_i p(y_i|x_i, \theta)$, where $p(\theta)$ is the prior density over the parameters and $p(y|x, \theta)$ is the likelihood given by the NN $f_\theta(x)$ for a data point (x, y). We consider a three-layer neural network with ReLU activation, input-layer size $d_x = 20$, hidden-layer size $d_h = 25$, and output-layer size $d_y = 1$. This results in $d = 550$ parameters in total. We use a Gaussian prior $p(\theta) = \mathcal{N}(\theta|0, \sigma_p^2 I)$ with $\sigma_p = 1/\sqrt{d_{\text{fan-in}}}$ and a Gaussian likelihood $p(y|x, \theta) = \mathcal{N}(y|f_\theta(x), \sigma_n^2)$ with $\sigma_n = 0.1$. We sample the ground-truth parameters $\theta^* \sim p(\theta)$ from the prior and use θ^* to generate $N = 500$ training points and 500 test points for evaluation.

Table 3. Average test predictive NLL for the BNN estimated by 10^3 samples generated by each sampler.

Sampler	NLL	#energy evaluations
MAP	0.548 ± 0.066	5.0×10^6
MALA	0.399 ± 0.014	5.0×10^6
HMC	0.315 ± 0.012	5.0×10^6
PT	0.241 ± 0.005	5.0×10^6
DiGS	$\mathbf{0.189 \pm 0.002}$	5.0×10^6

All methods are initialized at the same random sample from the prior and generate 150 samples from the posterior $p(\theta|D_{\text{train}})$ for evaluation. MAP (Maximum a Posteri) runs 7.5×10^5 full-batch gradient descent steps. MALA runs 5,000 Langevin steps per sample. HMC runs 5,000 leapfrog steps per sample. PT consists of 5 chains with temperatures $\tau = \{1.0, 5.62, 31.62, 177.83, 1000.0\}$, where each chain is constructed by an HMC sampler with 1,000 leapfrog steps per sample. DiGS uses the VP schedule with $T = 5$ noise levels, ranging from $\alpha_T = 0.1$ to $\alpha_1 = 0.9$, each with 100 Gibbs sweeps and 10 MALA denoising sampling steps per Gibbs sweep. Table 3 reports the average predictive negative log-likelihood (NLL) for each method on the test data, which shows that DiGS significantly outperforms other baselines. We speculate that the performance gain comes from the fact that DiGS captures a broader range of modes.

4.3. Molecular Configuration Sampling

Finally, we consider a real-world problem of sampling equilibrium molecular configurations from the Boltzmann distribution of the 22-atom molecule alanine dipeptide in an implicit solvent at temperature 300K, where the potential energy $E(x)$ is a function of 3D atomic coordinates obtained by simulating physical laws (Wu et al., 2020; Dibak et al., 2022; Campbell et al., 2021; Stimper et al., 2022;

Table 4. KL divergences for the Ramachandran plots and the marginals of the dihedral angles ϕ and ψ in alanine dipeptide.

Sampler	$p(\phi)$	$p(\psi)$	Ramachandran	#energy evaluations	#samples
MALA	1.9×10^{-3}	5.3×10^{-4}	1.1×10^{-2}	1.0×10^{9}	10^6
HMC	4.0×10^{-4}	3.1×10^{-4}	7.1×10^{-3}	1.0×10^{9}	10^6
DiGS	$\mathbf{2.3\times10^{-4}}$	$\mathbf{1.2\times10^{-4}}$	$\mathbf{4.3\times10^{-3}}$	1.0×10^{9}	10^6
MD PT (Midgley et al., 2023)	2.5×10^{-4}	1.4×10^{-4}	5.3×10^{-3}	2.3×10^{10}	10^6
Ground-truth (Midgley et al., 2023)	–	–	–	2.3×10^{11}	10^7

Midgley et al., 2023). This is a very challenging problem since $E(x)$ is highly multi-modal with many high energy barriers and is also costly to evaluate. Following the setup in Midgley et al. (2023), we represent the molecule with $d = 60$ roto-translation invariant internal coordinates.

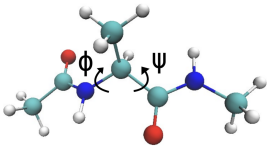
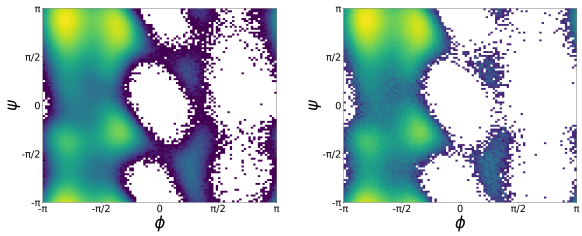


Figure 9. Alanine dipeptide (Midgley et al., 2023)

Each method is initialized at the minimum energy configuration as in Midgley et al. (2023) and generates 10^6 samples of molecular configurations for evaluation. MALA runs 1,000 Langevin steps per sample. HMC runs 1,000 leapfrog steps per sample. DiGS uses the VP schedule with $T = 2$ noise levels ($\alpha_2 = 0.1$ and $\alpha_1 = 0.9$), each with 100 Gibbs sweeps and 5 MALA denoising sampling steps per Gibbs sweep. Note that PT is the gold-standard method in molecular dynamics (MD) simulation, which serves as the ground-truth. The PT samples are taken from Midgley et al. (2023), which is generated using 21 chains starting at temperature 300K and increasing the temperature by 50K for each subsequent chain, where each chain is constructed by an HMC sampler with 1,000 leapfrog steps per sample. We follow Midgley et al. (2023) and treat 10^7 PT samples as the ground-truth. In addition, we consider a baseline of 10^6 PT samples as a reference. The quality of the sampled configurations is assessed by the Ramachandran plot (Ramachandran et al., 1963), which can be used to analyze how the protein folds locally. A Ramachandran plot is a 2D histogram for joint distribution of the two dihedral angles ϕ and ψ in the bonds connecting an amino acid to the protein backbone, as shown in Figure 9. Table 4 shows the KL divergences for the Ramachandran plots $p(\phi, \psi)$ and the marginals $p(\phi)$ and $p(\psi)$ between the ground-truth samples and samples generated by each sampler. We can see that DiGS significantly outperforms MALA and HMC with the same number of energy evaluations. Moreover, DiGS also outperforms the gold-standard MD simulation method PT with 23× less energy evaluations. Figure 10 shows a visual comparison between the ground-truth Ramachandran plots and the one produced by DiGS, confirming that DiGS captures all modes with the correct weightings.



(a) Ground-truth (b) *DiGS* (10^6 samples)

Figure 10. Ramachandran plots for alanine dipeptide.

5. Conclusion

Diffusive Gibbs Sampling (DiGS) is an innovative sampler that combines recent development in diffusion models with a novel Metropolis-within-Gibbs scheme. DiGS offers significant improvements in sampling multi-modal distributions, surpassing traditional methods in both efficiency and accuracy. Its applicability across various fields (e.g., Bayesian inference and molecular dynamics) showcases the potential of DiGS to facilitate sampling complex distributions in numerous scientific applications.

5.1. Limitations and Future Work

In our experiments, we observed that the acceptance rate of the Metropolis-within-Gibbs scheme given by Equation 15 was around 0.15, due to the random walk behavior of the linear Gaussian proposal as in Equation 14. A potential future work direction would be to design more efficient proposals for the Metropolis-within-Gibbs scheme with higher acceptance rate, for instance, combining the Gaussian convolution kernel and the linear Gaussian proposal by marginalizing out the noisy variable: $q(x|x^{(i-1)}) = \int q(x|\tilde{x})p(\tilde{x}|x^{(i-1)})d\tilde{x}$ (Titsias & Papaspiliopoulos, 2018). Empirically, it would be interesting to investigate the the performance of DiGS with different score-based samplers such as HMC in the denoising sampling step in future work. Besides, other multi-modal samplers such as population-based MCMC could be considered as baselines for comparison in future work. We also leave the convergence analysis of DiGS in the general multi-level noise scheduling setting as a future work; see Appendix C for some preliminary discussions.

Acknowledgements

We thank Andi Zhang for useful discussions on RDMC, Ruqi Zhang for highlighting Entropy-MCMC, an anonymous reviewer for discussions on PT, and Arnaud Doucet, Michael Hutchinson and Sam Power for pointing out proximal samplers.

MZ and DB acknowledge funding from the Cisco Centre of Excellence. WC acknowledges funding via a Cambridge Trust Scholarship (supported by the Cambridge Trust) and a Cambridge University Engineering Department Studentship (under grant G105682 NMZR/089 supported by Huawei R&D UK). JMHL acknowledges support from a Turing AI Fellowship under grant EP/V023756/1.

Impact Statement

The goal of this paper is to advance the field of machine learning. There could be many potential positive societal consequences of this methodological work, too numerous to specifically highlight here.

References

Agakov, F. V. and Barber, D. An auxiliary variational method. In *Neural Information Processing: 11th International Conference, ICONIP 2004, Calcutta, India, November 22-25, 2004. Proceedings 11*, pp. 561–566. Springer, 2004.

Barber, D. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

Barber, D. and Bishop, C. M. Ensemble learning in Bayesian neural networks. *Nato ASI Series F Computer and Systems Sciences*, 168:215–238, 1998.

Brown, B. C., Caterini, A. L., Ross, B. L., Cresswell, J. C., and Loaiza-Ganem, G. The union of manifolds hypothesis and its implications for deep generative modelling. *arXiv preprint arXiv:2207.02862*, 2022.

Campbell, A., Chen, W., Stimper, V., Hernandez-Lobato, J. M., and Zhang, Y. A gradient based strategy for Hamiltonian Monte Carlo hyperparameter optimization. In *International Conference on Machine Learning*, pp. 1238–1248. PMLR, 2021.

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.

Chen, Y., Chewi, S., Salim, A., and Wibisono, A. Improved analysis for a proximal algorithm for sampling. In *Conference on Learning Theory*, pp. 2984–3014. PMLR, 2022.

Deng, W., Zhang, Q., Feng, Q., Liang, F., and Lin, G. Non-reversible parallel tempering for deep posterior approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7332–7339, 2023.

Dibak, M., Klein, L., Krämer, A., and Noé, F. Temperature steerable flows and Boltzmann generators. *Physical Review Research*, 4(4):L042005, 2022.

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

Efron, B. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Gao, R., Lu, Y., Zhou, J., Zhu, S.-C., and Wu, Y. N. Learning generative convnets via multi-grid modeling and sampling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9155–9164, 2018.

Gao, R., Song, Y., Poole, B., Wu, Y. N., and Kingma, D. P. Learning energy-based models by diffusion recovery likelihood. *arXiv preprint arXiv:2012.08125*, 2020.

Geyer, C. J. and Thompson, E. A. Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association*, 90 (431):909–920, 1995.

Girolami, M. and Calderhead, B. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(2):123–214, 2011.

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.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Habib, R. and Barber, D. Auxiliary variational MCMC. In *International Conference on Learning Representations*, 2018.

Hernández-Lobato, J. M. and Adams, R. Probabilistic backpropagation for scalable learning of Bayesian neural networks. In *International conference on machine learning*, pp. 1861–1869. PMLR, 2015.

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

Huang, X., Hanze Dong, Y. H., Ma, Y., and Zhang, T. Reverse Diffusion Monte Carlo. *arXiv preprint arXiv:2307.02037*, 2023.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Lee, Y. T., Shen, R., and Tian, K. Structured logconcave sampling with a restricted gaussian oracle. In *Conference on Learning Theory*, pp. 2993–3050. PMLR, 2021.

Li, B. and Zhang, R. Entropy-mcmc: Sampling from flat basins with ease. *arXiv preprint arXiv:2310.05401*, 2023.

Louizos, C. and Welling, M. Multiplicative normalizing flows for variational bayesian neural networks. In *International Conference on Machine Learning*, pp. 2218–2227. PMLR, 2017.

Midgley, L. I., Stimper, V., Simm, G. N. C., Schölkopf, B., and Hernández-Lobato, J. M. Flow annealed importance sampling bootstrap. In *The Eleventh International Conference on Learning Representations*, 2023.

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

Neal, R. M. Slice sampling. *The annals of statistics*, 31(3): 705–767, 2003.

Neal, R. M. et al. MCMC using Hamiltonian dynamics. *Handbook of Markov chain Monte Carlo*, 2(11):2, 2011.

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

Pompe, E., Holmes, C., and Łatuszyński, K. A framework for adaptive MCMC targeting multimodal distributions. *The Annals of Statistics*, 48(5):2930–2952, 2020.

Ramachandran, G., Ramakrishnan, C., and Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7(1):95–99, 1963.

Robbins, H. E. An empirical bayes approach to statistics. In *Breakthroughs in Statistics: Foundations and basic theory*, pp. 388–394. Springer, 1992.

Robert, C. P., Casella, G., and Casella, G. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

Roberts, G. O. and Rosenthal, J. S. Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.

Roberts, G. O. and Smith, A. F. Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic processes and their applications*, 49(2):207–216, 1994.

Roberts, G. O. and Stramer, O. Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and computing in applied probability*, 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.

Roth, K., Lucchi, A., Nowozin, S., and Hofmann, T. Stabilizing training of generative adversarial networks through regularization. *Advances in neural information processing systems*, 30, 2017.

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

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Song, Y. and Ermon, S. Improved techniques for training score-based generative models. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

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

Staines, J. and Barber, D. Variational optimization. *arXiv preprint arXiv:1212.4507*, 2012.

Stimper, V., Schölkopf, B., and Hernández-Lobato, J. M. Resampling base distributions of normalizing flows. In *International Conference on Artificial Intelligence and Statistics*, pp. 4915–4936. PMLR, 2022.

Surjanovic, N., Syed, S., Bouchard-Côté, A., and Campbell, T. Parallel tempering with a variational reference. *Advances in Neural Information Processing Systems*, 35: 565–577, 2022.

Swendsen, R. H. and Wang, J.-S. Replica Monte Carlo simulation of spin-glasses. *Physical review letters*, 57 (21):2607, 1986.

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

Syed, S., Bouchard-Côté, A., Deligiannidis, G., and Doucet, A. Non-reversible parallel tempering: a scalable highly parallel mcmc scheme. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):321–350, 2022.

Titsias, M. K. and Papaspiliopoulos, O. Auxiliary gradient-based sampling algorithms. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4): 749–767, 2018.

Vempala, S. and Wibisono, A. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

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

Wenliang, L. and Kanagawa, H. Blindness of score-based methods to isolated components and mixing proportions. *arXiv preprint arXiv:2008.10087*, 2020.

Wu, H., Köhler, J., and Noé, F. Stochastic normalizing flows. *Advances in Neural Information Processing Systems*, 33: 5933–5944, 2020.

Zhang, M., Hayes, P., Bird, T., Habib, R., and Barber, D. Spread divergence. In *International Conference on Machine Learning*, pp. 11106–11116. PMLR, 2020.

Zhang, M., Key, O., Hayes, P., Barber, D., Paige, B., and Briol, F.-X. Towards healing the blindness of score matching. *arXiv preprint arXiv:2209.07396*, 2022.

Zhang, M., Hawkins-Hooker, A., Paige, B., and Barber, D. Moment matching denoising gibbs sampling. *Advances in Neural Information Processing Systems*, 36, 2023a.

Zhang, M., Sun, Y., Zhang, C., and Mcdonagh, S. Spread flows for manifold modelling. In *International Conference on Artificial Intelligence and Statistics*, pp. 11435–11456. PMLR, 2023b.

A. Proof of Theorem 2.1

Proof. A sufficient condition for irreducibility and recurrency of a Markov chain is that the joint distribution $p(x, \tilde{x})$ should satisfy the *positivity condition* (Robert et al., 1999; Roberts & Smith, 1994), which requires $p(x', \tilde{x}') > 0$ for all $x', \tilde{x}'$ such that $p(x') > 0$ and $\tilde{p}(\tilde{x}') > 0$, where $\tilde{p}(\tilde{x}') = \int p(\tilde{x}'|x)p(x)\,\mathrm{d}x$. This requirement is satisfied in the DiGS. This is because $p(x, \tilde{x}) = p(\tilde{x}|x)p(x)$, where the Gaussian convolution kernel $p(\tilde{x}|x)$ has full support in $\mathbb{R}^d$. Consequently, if $p(x') > 0$, then it follows that $p(\tilde{x}') > 0$ and $p(x', \tilde{x}') > 0$. □

B. An Analytical Example of Tempering v.s. Convolution

Although tempering-based sampling can alleviate the issue of multi-modal sampling in certain contexts, they still struggle in situations where the modes are less connected or completely isolated. Consider a toy example of a mixture of two Gaussians in 1D, given by $p(x) = \frac{1}{2}\mathcal{N}(x|\mu, \sigma_g^2) + \frac{1}{2}\mathcal{N}(x|-\mu, \sigma_g^2)$ where both Gaussian components have the same variance σ_g^2 and symmetric means positioned at $\mu, -\mu$. A point of interest is $x = 0$, which lies in the low-density region between the two modes and acts as a barrier point, hindering the state transition from one mode to another. In this example, the tempered log-density has a closed-form expression up to some constant C_β:

$$\begin{aligned}\log p_\beta(x=0) &= \beta \log p(x=0) + C_\beta \\ &= \beta\left(-\frac{\mu^2}{2\sigma_g^2} - \log \sigma_g\right) + C_\beta.\end{aligned} \tag{21}$$

For any given inverse temperature β and position μ, as $\sigma_g \to 0$, we have $\log p_\beta(x=0) \to -\infty$ and thus $p_\beta(x=0) \to 0$. This "Mixture of Deltas" example shows that tempering methods do not effectively overcome the low-density barrier in such situations. Furthermore, if we consider scenarios where each component of the mixture distribution is entirely disconnected, the tempered density in these regions remains zero, as tempering does not alter the support of a distribution.

In contrast, we define the convolved distribution $p(\tilde{x}) = \int p(\tilde{x}|x)p(x)\,\mathrm{d}x$ for the same target $p(x)$ with a convolution kernel $p(\tilde{x}|x) = \mathcal{N}(\tilde{x}|x, \sigma^2)$. For any given μ and σ_g, without loss of generality, we choose $\sigma \geq \sigma_g/\delta$ with a small constant $\delta > 0$. This leads to a lower bound:

$$\begin{aligned}\log p(\tilde{x}=0) &= -\frac{\mu^2}{2(\sigma_g^2+\sigma^2)} - \frac{1}{2}\log(2\pi(\sigma_g^2+\sigma^2)) \\ &\geq -\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2(1+\delta^2)),\end{aligned} \tag{22}$$

illustrating that the convolved log-density at $\tilde{x} = 0$ is lower-bounded for any $\sigma > 0$ as $\sigma_g \to 0$, since the lower bound of $\log p(\tilde{x} = 0)$ is independent of σ_g and remains finite. This approach effectively guarantees the maintenance of a non-negligible density within the bridges connecting different modes of distribution. Figure 5 presents a one-dimensional comparison between the tempering and convolution methods. However, in cases where the standard deviation of the Mixture of Gaussians (MoG) is exceptionally small, the efficacy of tempering diminishes significantly. In such instances, even with a large temperature T, tempering fails to connect the modes. Conversely, the convolution method continues to effectively bridge the modes, even with a relatively small σ. This comparison underscores the convolution method's robustness in handling scenarios with disconnected modes, a situation that also mirrors challenges often encountered in high-dimensional cases, whereas the tempering method struggles.

C. Discussions of Convergence Guarantees

In some specific cases, the convergence rate of DiGS may be established. Specifically, the convergence rate of DiGS depends on both the convergence of the inner denoising sampling step and the convergence of the outer Gibbs sampler. For instance, when using MALA to sample from the denoising posterior $p(x|\tilde{x})$, the convergence of MALA has been studied under different conditions such as strong log-concavity and log-Sobolev inequality (Vempala & Wibisono, 2019). Compared to direct sampling from the target distribution, the L2 regularizer in the log denoising posterior can improve the log-Sobolev condition:

$$\log p(x|\tilde{x}) = -E(x) - \frac{\|\alpha x - \tilde{x}\|^2}{2\sigma^2} + const, \tag{23}$$

which in turn accelerates the convergence of MALA (Huang et al., 2023). Intuitively, this makes the target distribution more "Gaussian-like". The convergence of the outer Gibbs sampler can be analyzed following the approach in Chen et al. (2022), demonstrating convergence under conditions of strong log-concavity/log-concavity under Wasserstein distance and log-Sobolev inequality under KL divergence. Therefore, the theoretical convergence guarantee of DiGS with a single noise level can be obtained by combining the convergence analyses of the inner MALA denoising sampling step and the outer Gibbs sampler from Vempala & Wibisono (2019); Chen et al. (2022); Huang et al. (2023).

However, in addition to the aforementioned factors, the multi-level noise schedule also plays a part in accelerating the convergence speed. Furthermore, our empirical investigation revealed that the performance of DiGS for practical problems such as molecular dynamics could also be affected by the initial condition (e.g., we followed the convention of molecular dynamics to initialize DiGS at the minimum energy configuration, which turns out to be better than random initialization). These analyses are out of the scope of this work, and we leave the theoretical study of the general DiGS framework as a future research direction.

D. Experimental Details

In all experiments, the step sizes of MALA and HMC are tuned via trial-and-error so that the acceptance rates are close to optimal values 0.574 (Roberts & Rosenthal, 1998) and 0.65 (Neal et al., 2011), respectively.

D.1. Comparison of Initialization Strategies for the Denoising Sampling Step

For the comparison of three initialization strategies for the denoising sampling step in Figure 3 and Table 1 in Section 2.2, we run DiGS with $\alpha = 1, \sigma = 1$ for 200 Gibbs sweeps on an MoG with unbalanced weights. The denoising sampling step begins with an initialization using the Metropolis-Hastings (MH) algorithm, followed by 50 MALA steps with a step size of 1×10^{-3}. For evaluation, we employ the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) that utilizes 5 kernels with bandwidths $\{2^{-2}, 2^{-1}, 2^0, 2^1, 2^2\}$.

D.2. Comparison of Gaussian Convolution Hyperparameters

For the convolution parameter comparison in Section 2.3, we use DiGS to generate 1,000 samples. For each sample, we execute 1,000 Gibbs sweeps. The denoising sampling step begins with an initialization using the Metropolis-Hastings (MH) algorithm, followed by 10 MALA steps. The step size for MALA is set at 1×10^{-3}. We set $\sigma = 1.0$ and vary α across $\{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 2.0, 5.0\}$ for the experiment shown in Figure 4(a). Similarly, we set $\alpha = 1.0$ and vary σ across $\{0.1, 0.3, 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 10, 15, 20\}$ for the experiment shown in Figure 4(b). For evaluation, we employ the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) that utilizes 5 kernels with bandwidths $\{2^{-2}, 2^{-1}, 2^0, 2^1, 2^2\}$.

D.3. Multi-Level Noise Scheduling

For the multi-level noise experiment with the VP schedule in Section 2.4, we define $\alpha_t = \alpha_T + (T - t)\Delta_\alpha$, where $\Delta_\alpha = (\alpha_1 - \alpha_T)/(T - 1)$ and $\sigma_t = \sqrt{1 - \alpha_t^2}$. We set $\alpha_1 = 0.9$ and $\alpha_T = 0.1$, and experiment with $T \in \{2, 3, 4, 5\}$ for the comparison shown in Figure 4(c). For evaluation, we employ the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) that utilizes 5 kernels with bandwidths $\{2^{-2}, 2^{-1}, 2^0, 2^1, 2^2\}$.

D.4. Comparison with Parallel Tempering

For comparison with parallel tempering on the "Mixture of Delta" problem in Figure 6 in Section 3.1, each sampler is initialized at the origin and generates 1,000 samples. DiGS employs a single noise level Gaussian convolution with parameters $\alpha = 1$ and $\sigma = 1$, complemented by 1,000 Gibbs sweeps per sample. The denoising sampling step in each Gibbs sweep begins with MH initialization, followed by 5 steps of MALA with a step size of 1×10^{-3}. PT consists of 5 chains with temperatures τ=$\{1.0, 5.62, 31.62, 177.83, 1000.0\}$, where each chain is constructed by an HMC sampler with 1,000 leapfrog steps per sample and a step size of 1×10^{-2}.

To make PT work on this problem, one needs to use 20 chains with temperatures τ=$\{$1.0, 2.07, 4.28, 8.86, 18.33, 37.93, 78.48, 162.38, 335.98, 695.19, 1432.45, 2976.35, 6158.48, 12742.75, 26366.51, 54555.95, 112883.79, 233572,15, 483293.02, 1000000.0$\}$ and the DEO scheme (Deng et al., 2023), where each chain is constructed by an HMC sampler with 500 leapfrog

steps per sample and a step size of 1×10^{-2}.

D.5. Comparison with RDMC

For the comparison with Reverse Diffusion Monte Carlo (RDMC) in Figure 7 in Section 3.4, we follow the original setting discussed in Huang et al. (2023) and use ULA for generating samples for the posterior distribution. The score function of the reverse distribution using $K = 5$ to construct the score approximation

$$\nabla_{x_t} \log p(x_t) \approx \left(\frac{\alpha}{K} \sum_{k=1}^{K} x_0^{(t,k)} - x_t \right) / \sigma_t^2, \tag{24}$$

where $x_0^{(t,k)} \sim p(x_0|x_t)$ denotes the samples drawn using ULA with $L_{ULA} = 5$ steps and a step size of 1×10^{-2}. These samples are initialized through importance sampling (IS) with $S_{is} = 100$ samples. The discretization step size η is set to Γ/T, with the scaling factor $a_t = e^{-(\Gamma - t\cdot\eta)}$ and the standard deviation $\sigma_t = \sqrt{1-\alpha_t^2}$. Following the procedure in Huang et al. (2023), we further apply ULA for additional L_{ULA} steps after RDMC. The parameter Γ is fixed at 0.1 for our experiments, and we explore different values of $T \in \{1, 2, 3, 4\}$, necessitating $TK(L_{ULA} + S_{is}) + L_{ULA}$ energy evaluations to generate a single sample.

For DiGS, we use DiGS with one noise level Gaussian convolution with parameter $\alpha = 1, \sigma = 1$. For sampling from the denoising distribution, we use the MH initialization followed by the ULA sampling with $L_{ULA} = 5$ steps and a step size of 1×10^{-2}. This MH+ULA scheme ensures a fair comparison to the IS+ULA scheme used in the RDMC. We vary $S_{gibbs} \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ Gibbs sweeps to generate a sample, which takes $S_{gibbs}(L_{ULA} + 2)$ energy evaluations in total, where the constant 2 accounts for the two energy evaluations required by MH. For both DiGS and RDMC, a total of 1,000 samples are generated for comparison.

For evaluation, we employ the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) that utilizes 5 kernels with bandwidths $\{2^{-2}, 2^{-1}, 2^0, 2^1, 2^2\}$.

D.6. Mixture of 40 Gaussians

For the MoG-40 experiment in Section 4.1, each method is initialized at the origin and generates 10^4 samples for evaluation. MALA runs 1,000 Langevin steps per sample with a step size of 1×10^{-1}. HMC runs 1,000 leapfrog steps per sample with a step size of 1×10^{-1}. PT consists of 5 chains with temperatures $\tau = \{1.0, 5.62, 31.62, 177.83, 1000.0\}$, where each chain is constructed by an HMC sampler with 200 leapfrog steps per sample with a step size of 1×10^{-1}. DiGS uses $T = 1$ noise level with $\alpha = 0.1$ and $\sigma^2 = 1 - \alpha^2$, 200 Gibbs sweeps, and 5 MALA denoising sampling steps per Gibbs sweep with a step size of 1×10^{-1}.

D.7. Bayesian Neural Networks

For the BNN experiment in Section 4.2, all methods are initialized at the same random sample from the prior and generate 150 samples from the posterior $p(\theta|D_{\text{train}})$ for evaluation. MAP runs 7.5×10^5 full-batch gradient descent steps with a step size of 3×10^{-2}. MALA runs 5,000 Langevin steps per sample with a step size of 1×10^{-4}. HMC runs 5,000 leapfrog steps per sample with a step size of 5×10^{-4}. PT consists of 5 chains with temperatures $\tau = \{1.0, 5.62, 31.62, 177.83, 1000.0\}$, where each chain is constructed by an HMC sampler with 1,000 leapfrog steps per sample with a step size of 5×10^{-4}. DiGS uses the VP schedule with $T = 5$ noise levels, ranging from $\alpha_T = 0.1$ to $\alpha_1 = 0.9$, each with 100 Gibbs sweeps and 10 MALA denoising sampling steps per Gibbs sweep with a step size of 1×10^{-4}.

D.8. Molecular Configuration Sampling

For the molecular configuration sampling experiment in Section 4.3, each method is initialized at the minimum energy configuration as in Midgley et al. (2023) and generates 10^6 samples of configurations for evaluation. MALA runs 1,000 Langevin steps per sample with a step size of 1×10^{-4}. HMC runs 1,000 leapfrog steps per sample with a step size of 1×10^{-3}. DiGS uses the VP schedule with $T = 2$ noise levels ($\alpha_2 = 0.1$ and $\alpha_1 = 0.9$), each with 100 Gibbs sweeps and 5 MALA denoising sampling steps per Gibbs sweep with a step size of 1×10^{-4}. Note that PT is the gold-standard method in molecular dynamics (MD) simulation, which serves as the ground-truth. The PT samples are taken from Midgley et al. (2023), which is generated using 21 chains starting at temperature 300K and increasing the temperature by 50K for

Table 5. Sample quality on a 200-dimensional mixture of Gaussians problem.

Sampler	MMD	#energy evaluations
MALA	0.123 ± 0.035	10^6
HMC	0.104 ± 0.025	10^6
PT	0.094 ± 0.014	10^6
DiGS	$\mathbf{0.027} \pm \mathbf{0.010}$	10^6

each subsequent chain, where each chain is constructed by an HMC sampler with 1,000 leapfrog steps per sample. We follow Midgley et al. (2023) and treat 10^7 PT samples as the ground-truth. In addition, we consider a baseline of 10^6 PT samples as a reference.

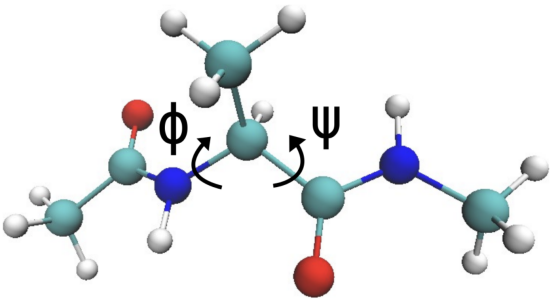
E. Additional Experimental Results

E.1. 200-Dimensional Mixture of Gaussians

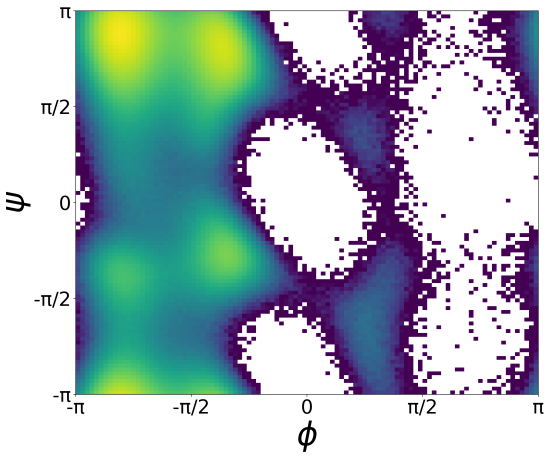
We take the MoG target in Figure 2(a) and generalize it to the 200-dimensional space. For each compared sampler, we draw 1,000 samples within a given compute budget of 10^6 energy evaluations. All samplers are tuned to achieve their optimal acceptance rate as described in Section 4. Table 5 shows that DiGS significantly outperforms all the other samplers even in the high-dimensional space.

E.2. Molecular Configuration Sampling

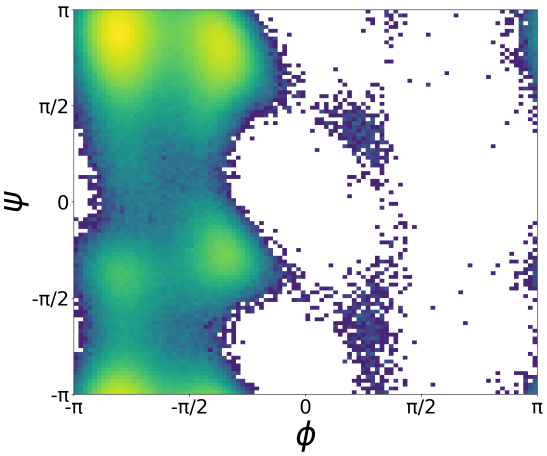
The Ramachandran plots produced by all compared methods for the molecular configuration sampling experiment in Section 4.3 can be found in Figure 11.



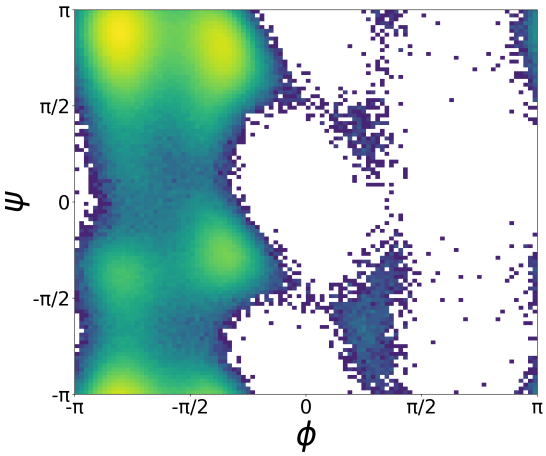
(a) Dihedral angles ϕ, ψ in alanine dipeptide



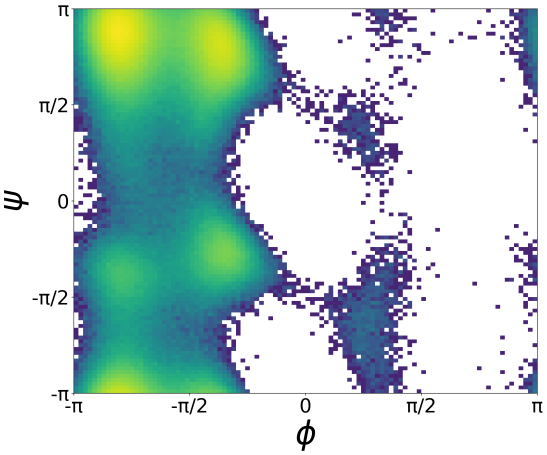
(b) MD (10^7 samples, 2.3×10^{11} energy evaluations)



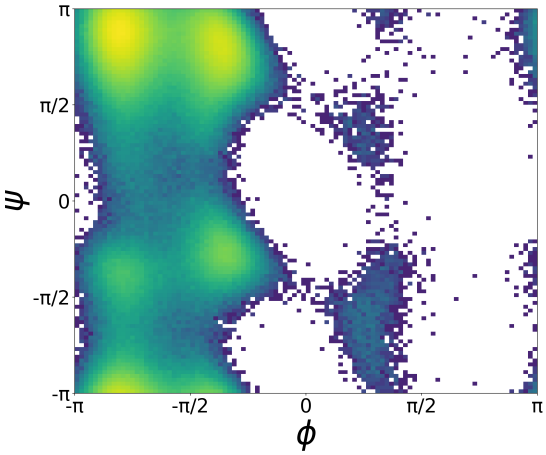
(c) MALA (10^6 samples, 1.0×10^9 energy evaluations)



(d) HMC (10^6 samples, 1.0×10^9 energy evaluations)



(e) PT (10^6 samples, 2.3×10^{10} energy evaluations)



(f) *DiGS* (10^6 samples, 1.0×10^9 energy evaluations)

Figure 11. (a) Visualization of the dihedral angles ϕ and ψ in alanine dipeptide (Midgley et al., 2023). (b)-(d) Ramachandran plots for the dihedral angles ϕ and ψ of alanine dipeptide. MD simulation (PT with 10^7 samples) serves as ground-truth.