Beyond ELBOs: A Large-Scale Evaluation of Variational Methods for Sampling

Denis Blessing[1] Xiaogang Jia[1] Johannes Esslinger[1] Francisco Vargas[2] Gerhard Neumann[1 3]

Abstract

Monte Carlo methods, Variational Inference, and their combinations play a pivotal role in sampling from intractable probability distributions. However, current studies lack a unified evaluation framework, relying on disparate performance measures and limited method comparisons across diverse tasks, complicating the assessment of progress and hindering the decision-making of practitioners. In response to these challenges, our work introduces a benchmark that evaluates sampling methods using a standardized task suite and a broad range of performance criteria. Moreover, we study existing metrics for quantifying mode collapse and introduce novel metrics for this purpose. Our findings provide insights into strengths and weaknesses of existing sampling methods, serving as a valuable reference for future developments. The code is publicly available here.

1. Introduction

Sampling methods are designed to address the challenge of generating approximate samples or estimating the intractable normalization constant Z for a probability density π on $\mathbb{R}^d$ of the form

$$\pi(\mathbf{x}) = \frac{\gamma(\mathbf{x})}{Z}, \quad Z = \int_{\mathbb{R}^d} \gamma(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{1}$$

where $\gamma : \mathbb{R}^d \to \mathbb{R}^+$ can be pointwise evaluated. This formulation has broad applications in fields such as Bayesian statistics and the natural sciences (Liu & Liu, 2001; Stoltz et al., 2010; Frenkel & Smit, 2023; Mittal et al., 2023).

Monte Carlo (MC) methods (Hammersley, 2013), including Annealed Importance Sampling (AIS) (Neal, 2001) and its Sequential Monte Carlo (SMC) extensions (Del Moral et al., 2006), have traditionally been considered the gold standard for addressing the sampling problem. An alternative approach is Variational Inference (VI) (Blei et al., 2017), where a tractable family of distributions is parameterized, and optimization tools are employed to maximize similarity to the intractable target distribution π.

In recent years, there has been a surge of interest in the development of sampling methods that merge MC with VI techniques to approximate complex, potentially multimodal distributions (Wu et al., 2020a; Zhang & Chen, 2021; Arbel et al., 2021; Matthews et al., 2022; Jankowiak & Phan, 2022; Midgley et al., 2022; Berner et al., 2022; Richter et al., 2023; Vargas et al., 2023a;b; Akhound-Sadegh et al., 2024).

However, the evaluation of these methods faces significant challenges, including the absence of a standardized set of tasks and diverse performance criteria. This diversity complicates meaningful comparisons between methods. Existing evaluation protocols, such as the evidence lower bound (ELBO), often rely on samples from the model, restricting their evaluation capabilities to the model's support. This limitation becomes especially problematic when assessing the ability to mitigate mode collapse on target densities with well-separated modes. To overcome this challenge, others propose the use of integral probability metrics (IPMs), like maximum mean discrepancy (Arenz et al., 2018) or Wasserstein distance (Richter et al., 2023; Vargas et al., 2023a), leveraging samples from the target density to assess performance beyond the model's support. However, these metrics often involve subjective design choices such as kernel selection or cost function determination, potentially leading to biased evaluation protocols.

To address these challenges, our work introduces a comprehensive set of tasks for evaluating variational methods for sampling. We explore existing evaluation criteria and propose a novel metric specifically tailored to quantify mode collapse. Through this evaluation, we aim to provide valuable insights into the strengths and weaknesses of current sampling methods, contributing to the future design of more effective techniques and the establishment of standardized evaluation protocols.

[1] Autonomous Learning Robots, Karlsruhe Institute of Technology, Karlsruhe, Germany [2] University of Cambridge, Cambridge, United Kingdom [3] FZI Research Center for Information Technology, Karlsruhe, Germany. Correspondence to: Denis Blessing <denis.blessing@kit.edu>.

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

2. Preliminaries

We provide an overview of Monte Carlo methods, Variational Inference, and combinations. The notation introduced in this section is used throughout the remainder of this work.

Monte Carlo Methods. A variety of Monte Carlo (MC) techniques have been developed to tackle the sampling problem and estimation of Z. In particular sequential importance sampling methods such as Annealed Importance Sampling (AIS) (Neal, 2001) and its Sequential Monte Carlo (SMC) extensions (Del Moral et al., 2006) are often regarded as a gold standard to compute Z. These approaches construct a sequence of distributions $(\pi_t)_{t=1}^T$ that 'anneal' smoothly from a tractable proposal distribution π_0 to the target distribution $\pi_T = \pi$. One typically uses the geometric average, that is, $\gamma_t(\mathbf{x}) = \pi_0(\mathbf{x})^{1-\beta_t}\gamma(\mathbf{x})^{\beta_t}$, with $\pi_t \propto \gamma_t$ for $0 = \beta_0 < \beta_1 < ... < \beta_T = 1$. Approximate samples from π are then obtained by starting from $\mathbf{x}_0 \sim \pi_0(\cdot)$ and running a sequence of Markov chain Monte Carlo (MCMC) transitions that target $(\pi_t)_{t=1}^T$.

Variational Inference. Variational inference (VI) (Blei et al., 2017) is a popular alternative to MCMC and SMC where one considers a flexible family of easy-to-sample distributions q^θ whose parameters θ are optimized by minimizing the reverse Kullback–Leibler (KL) divergence, i.e.,

$$D_{\text{KL}}(q^\theta(\mathbf{x})\|\pi(\mathbf{x})) = \underbrace{-\,\mathbb{E}_{q^\theta(\mathbf{x})}\Big[\log\frac{\gamma(\mathbf{x})}{q^\theta(\mathbf{x})}\Big]}_{\text{ELBO}} + \log Z \quad (2)$$

It is well known that minimizing the reverse KL is equivalent to maximizing the evidence lower bound (ELBO) and that ELBO $\leq \log Z$ with equality if and only if $q^\theta = \pi$. Later, VI was extended to other variational objectives such as α-divergences (Li & Turner, 2016; Midgley et al., 2022), log-variance loss (Richter et al., 2020), trajectory balance, (Malkin et al., 2022a) or general f- divergences (Wan et al., 2020). Typical choices for q^θ include mean-field approximations (Wainwright & Jordan, 2008), mixture models (Arenz et al., 2022) or normalizing flows (Papamakarios et al., 2021). To construct more flexible variational distributions (Agakov & Barber, 2004) modeled $q^\theta(\mathbf{x})$ as the marginal of a latent variable model, i.e. $q^\theta(\mathbf{x}) = \int q^\theta(\mathbf{x},\mathbf{z})\mathrm{d}\mathbf{z}$ [1]. As this marginal is typically intractable, θ is then learned by minimizing a discrepancy measure between $q^\theta(\mathbf{x},\mathbf{z})$ and an extended target $p^\theta(\mathbf{x},\mathbf{z}) = \pi(\mathbf{x})p^\theta(\mathbf{z}|\mathbf{x})$ where $p^\theta(\mathbf{z}|\mathbf{x})$ is an auxiliary conditional distribution. Using the chain rule for the KL-divergence (Cover, 1999) one obtains an extended version of the ELBO, that is,

$$D_{\text{KL}}(q^\theta(\mathbf{x})\|\pi(\mathbf{x})) \leq \underbrace{-\,\mathbb{E}_{q^\theta(\mathbf{x},\mathbf{z})}\Big[\log\frac{\gamma(\mathbf{x})p^\theta(\mathbf{z}|\mathbf{x})}{q^\theta(\mathbf{x},\mathbf{z})}\Big]}_{\overline{\text{ELBO}}} + \log Z. \quad (3)$$

Although the extended ELBO is often referred to as ELBO, latent variables $\mathbf{z}$ introduce additional looseness, i.e., $\overline{\text{ELBO}} \leq$ ELBO with equality if $p^\theta(\mathbf{z}|\mathbf{x}) = q^\theta(\mathbf{x},\mathbf{z})/q^\theta(\mathbf{x})$. To compute expectations with respect to $q^\theta(\mathbf{x},\mathbf{z})$, one typically chooses tractable distributions $q^\theta(\mathbf{x}|\mathbf{z})$ and $q^\theta(\mathbf{z})$ and performs a Monte Carlo estimate using ancestral sampling.

Variational Monte Carlo Methods. Over recent years, the idea of using extended distributions has been further explored (Wu et al., 2020b; Geffner & Domke, 2021; Thin et al., 2021; Zhang et al., 2021; Doucet et al., 2022b; Geffner & Domke, 2022). In particular, these ideas marry Monte Carlo with variational techniques by constructing the variational distribution and extended target as Markov chains, i.e., $q^\theta(\mathbf{x}_{0:T}) = \pi_0(\mathbf{x}_0)\prod_{t=1}^T F_t^\theta(\mathbf{x}_t|\mathbf{x}_{t-1})$ and $p^\theta(\mathbf{x}_{0:T}) = \pi(\mathbf{x}_T)\prod_{t=0}^{T-1} B_t^\theta(\mathbf{x}_t|\mathbf{x}_{t+1})$ with $\mathbf{x} = \mathbf{x}_T$, $\mathbf{z} = (\mathbf{x}_0, \ldots, \mathbf{x}_{T-1})$ and tractable π_0. Common choices of transition kernels F_t^θ, B_t^θ include Gaussian distributions (Doucet et al., 2022b; Geffner & Domke, 2022) or normalizing flow maps (Wu et al., 2020a; Arbel et al., 2021; Matthews et al., 2022) and can be optimized by e.g. maximization of the extended ELBO via stochastic gradient ascent. Recently, Vargas et al. (2023a); Zhang & Chen (2021); Vargas et al. (2023b; 2024); Richter et al. (2023); Berner et al. (2022) explored the limit of $T \to \infty$ in which case the Markov chains converge to forward and backward time stochastic differential equations (SDEs) (Anderson, 1982; Song et al., 2020) inducing the path distributions $\mathbb{Q}^\theta$ and $\mathbb{P}^\theta$ which can be thought of as continuous time analogous of q^θ and p^θ respectively. Zhang & Chen (2021); Berner et al. (2022); Richter et al. (2023); Vargas et al. (2024) leveraged the continuous-time perspective to establish connections with Schrödinger bridges (Léonard, 2013) and stochastic optimal control (Dai Pra, 1991), resulting in the development of novel sampling algorithms.

Performance Criteria. Several performance criteria have been proposed for evaluating sampling methods, notably, those comparing the density ratio between the target and model density and integral probability metrics (IPMs).

Density ratio-based criteria make use of the ratio between the (unnormalized) target density $\gamma(\mathbf{x})$ and the model $q^\theta(\mathbf{x})$. Due to the intractability of $q^\theta(\mathbf{x})$ for methods that work with latent variables, the density ratio between the joint distributions of $\mathbf{x}$ and $\mathbf{z}$ is considered, i.e.,

$$w = \frac{\gamma(\mathbf{x})}{q^\theta(\mathbf{x})}, \quad \text{and} \quad \overline{w} = \frac{\gamma(\mathbf{x})p^\theta(\mathbf{z}|\mathbf{x})}{q^\theta(\mathbf{x},\mathbf{z})}, \quad (4)$$

respectively. Note that w and $\overline{w}$ are also referred to as

[1] Agakov & Barber (2004) coined the term 'augmentation' for $\mathbf{z}$. We adopt the more established terminology and refer to $\mathbf{z}$ as a latent variable.

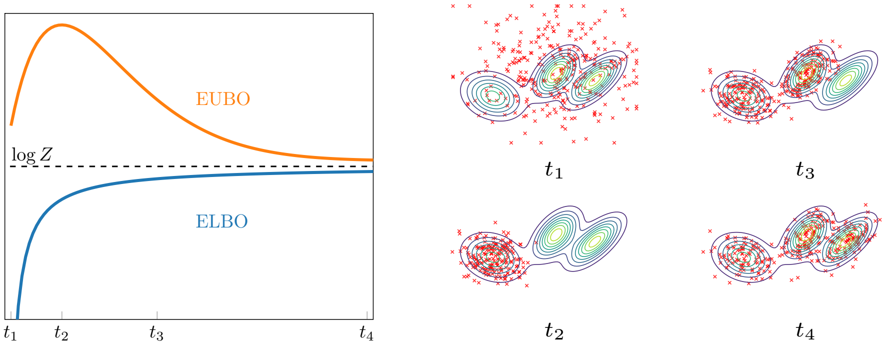


Figure 1: Illustration of the evidence upper (EUBO) and lower bound (ELBO). The mode-seeking nature of reverse KL results in ELBO $\ll \log Z$ if the model density q^θ (indicated by the samples $\times$) averages over the target π (indicated by the level plot) (t_1) and ELBO $\approx \log Z$ if $\pi \geq 0$ whenever $q^\theta \geq 0$ ($t_2 - t_4$). As a result, the ELBO is not sensitive to mode collapse. In contrast, the mass-covering nature of the forward KL ensures that EUBO $\gg \log Z$ if $q^\theta \approx 0$ whenever $\pi > 0$ (t_2) and EUBO $\approx \log Z$ if $q^\theta \geq 0$ whenever $\pi \geq 0$ (t_1). Consequently, the EUBO is well suited to quantify mode collapse.

(unnormalized) importance weights. Using this notation, we can recover commonly used metrics such as the reverse effective sample size (ESS$_r$) or the ELBO, that is,

$$\text{ESS}_r = \frac{(\mathbb{E}_{q^\theta}[w])^2}{\mathbb{E}_{q^\theta}[w^2]} \quad \text{and} \quad \text{ELBO} = \mathbb{E}_{q^\theta}[\log w], \tag{5}$$

respectively. Here, 'reverse' is used to denote that expecations are computed with respect to q^θ. In addition, if the true normalization constant is known, an importance-weighted reverse estimate of $\log Z$ is often employed to report the esimation bias, i.e., $\Delta \log Z_r = |\log Z - \log \hat{Z}_r|$ with

$$\log \hat{Z}_r = \log \mathbb{E}_{q^\theta}[w]. \tag{6}$$

Please note that extended versions of these criteria are obtainable by replacing w with the extended version $\overline{w}$ and taking expectations under the joint distribution $q^\theta(\mathbf{x}, \mathbf{z})$.

3. Quantifying Mode-Collapse

Quantifying the ability to avoid mode collapse is difficult as identifying all modes of the target density π and determining whether a model captures them accurately is inherently challenging. In particular, methods that are optimized using the reverse KL divergence are forced to assign high probability to regions with non-negligible probability in the target distribution π. This is referred to as mode-seeking behavior and can result in an overemphasis on a limited set of modes, leading to mode collapse. Consequently, performance criteria that use expectations under the model q^θ, such as ELBO, (reverse) ESS, or $\Delta \log Z_r$, are influenced by the mode-seeking nature of the reverse KL divergence, making them less sensitive to mode collapse.

Here, we aim to explore criteria that are sensitive to mode collapse such as density-ratio based 'forward' criteria, that leverage expectations under π and integral probability metrics (IPMs). Furthermore, we introduce *entropic mode coverage*, a novel criterion that leverages prior knowledge about the target to heuristically quantify mode coverage.

Forward Criteria. We discuss the 'forward' versions of the criteria discussed in Section 2. First, evidence upper bounds (EUBOs) are based on the forward KL divergence and have already been leveraged as learning objectives in VI (Ji & Shen, 2019). Here, we explore them as performance criteria that are sensitive to mode collapse. Formally, the EUBO is the sum of the forward KL and $\log Z$, that is,

$$D_{\text{KL}}(\pi(\mathbf{x}) \| q^\theta(\mathbf{x})) = \underbrace{\mathbb{E}_{\pi(\mathbf{x})}[\log w]}_{\text{EUBO}} - \log Z, \tag{7}$$

with importance weights $w = \gamma(\mathbf{x})/q^\theta(\mathbf{x})$. Due to the non-negativeness of the KL divergence, it is easy to see that EUBO $\geq \log Z$ with equality if and only if $q^\theta = \pi$. Hence, a lower EUBO means that q^θ is closer to π in a D_{KL} sense. The mass-covering nature of the forward KL leads to high EUBO values if the model fails to cover regions of non-negligible probability in the target distribution π and is therefore well suited to quantify mode-collapse. This is further illustrated in Figure 1. We can again leverage the chain rule for the KL-divergence (Cover, 1999) to obtain an extended version of the EUBO, i.e., $\mathbb{E}_{\pi(\mathbf{x},\mathbf{z})}[\log \overline{w}]$ that satisfies $\overline{\text{EUBO}} \geq$ EUBO, where the introduction of latent variables introduce additional looseness. The extended EUBO requires computing the importance weights $\overline{w}$ and expectations under $\pi(\mathbf{x}, \mathbf{z})$. The former depends on the specific choice of sampling algorithm and is further discussed in Section 4 when introducing the methods considered for evaluation. The latter can be approximated by propagating target samples $\mathbf{x}$ back to $\mathbf{z}$ using $\pi(\mathbf{z}|\mathbf{x})$. Additionally, having access to samples from π allows for computing forward versions of Z and ESS which have already been used to quantify mode collapse by e.g. (Midgley et al., 2023). Formally, they are defined as

$$Z_f = 1/\mathbb{E}_\pi[w^{-1}], \quad \text{and} \quad \mathrm{ESS}_f = Z_f/\mathbb{E}_\pi[w], \tag{8}$$

where expectations are taken with respect to the target π. For a detailed discussion see Appendix A.1.

Integral Probability Metrics. Alternatively, IPMs are often employed if samples from the target distribution π are available (Arenz et al., 2018; Richter et al., 2023; Vargas et al., 2023a; 2024). Common IPMs for assessing sample quality are 2-Wasserstein distance ($\mathcal{W}_2$) (Peyré et al., 2019) or the maximum mean discrepancy (MMD) (Gretton et al., 2012). The former uses a cost function to calculate the minimum cost required to transport probability mass from one distribution to another while the latter assesses distribution dissimilarity by examining the differences in their mean embeddings within a reproducing kernel Hilbert space (Aronszajn, 1950). For further details see Appendix A.2.

Entropic Mode Coverage. Inspired by inception scores and distances from generative modelling (Salimans et al., 2016; Heusel et al., 2017) we propose a heuristic approach for detecting mode collapse by introducing the entropic mode coverage (EMC). To compute EMC, we partition $\mathbb{R}^d$ into disjoint subsets $\xi_i, i \in \{1, \dots, M\}$ that describe different modes of the target density π. Moreover, we introduce an auxiliary distribution that measures the probability of a sample $\mathbf{x}$ being element of a mode descriptor, i.e., $p(\xi_i|\mathbf{x}) = p(\mathbf{x} \in \xi_i)$. We then compute the expected entropy of the auxiliary distribution, that is,

$$\begin{aligned} \mathrm{EMC} &:= \mathbb{E}_{q^\theta(\mathbf{x})}\mathcal{H}\left(p(\xi|\mathbf{x})\right) \\ &\approx -\frac{1}{N}\sum_{\mathbf{x}\sim q^\theta}\sum_{i=1}^{M} p(\xi_i|\mathbf{x})\log_M p(\xi_i|\mathbf{x}), \end{aligned} \tag{9}$$

where the expectation is approximated using a Monte Carlo estimate. Here, N denotes the number of samples drawn from q^θ. Please note that we employ the logarithm with a base of M to ensure that EMC $\in [0, 1]$. This choice of base facilitates a straightforward interpretation: A value of 0 signifies a model that consistently produces samples that are elements of the same mode descriptor. In contrast, a value of 1 represents a model that can produce samples from all mode descriptors with equal probability.

Clearly, EMC is limited to targets where mode descriptors are available which is further discussed in Section 5. Moreover, a suitable criterion for cases where mode descriptors are not equally probable is discussed in Appendix A.3.

4. Benchmarking Methods

In this section, we elaborate on the methods included in this benchmark, categorizing them into three distinct groups based on the computation of importance weights. Please refer to Table 1 for an overview of these methods and to Appendix B for further details.

Acronym	Method	Reference
MFVI	Gaussian Mean-Field VI	(Bishop, 2006)
GMMVI	Gaussian Mixture Model VI	(Arenz et al., 2022)
NFVI†	Normalizing Flow VI	(Rezende & Mohamed, 2015)
SMC	Sequential Monte Carlo	(Del Moral et al., 2006)
AFT	Annealed Flow Transport	(Arbel et al., 2021)
CRAFT	Continual Repeated AFT	(Matthews et al., 2022)
FAB	Flow Annealed IS Bootstrap	(Midgley et al., 2022)
ULA†	Uncorrected Langevin Annealing	(Thin et al., 2021)
MCD	Monte Carlo Diffusion	(Doucet et al., 2022a)
UHA†	Uncorrected Hamiltonian Annealing	(Geffner & Domke, 2021)
LDVI	Langevin Diffusion VI	(Geffner & Domke, 2022)
CMCD†	Controlled MCD	(Vargas et al., 2024)
PIS	Path Integral Sampler	(Zhang & Chen, 2021)
DIS	Time-Reversed Diffusion Sampler	(Berner et al., 2022)
DDS	Denoising Diffusion Sampler	(Vargas et al., 2023a)
GFN†	Generative Flow Networks	(Lahlou et al., 2023)
GBS	General Bridge Sampler	(Richter et al., 2023)

Table 1: Sampling Methods. For methods marked with '†', implementation is available, but the results are either not included or only partially presented in this work.

Tractable Density Models. Tractable density models allow for computing the model likelihood $q^\theta(\mathbf{x})$. It is therefore straightforward to compute performance criteria associated with importance weights $w = \gamma(\mathbf{x})/q^\theta(\mathbf{x})$. Notable works include factorized ('mean-field') Gaussian distributions (MFVI), Normalizing Flows (NFVI) (Rezende & Mohamed, 2015) and full rank Gaussian mixture models (GMMVI) (Arenz et al., 2022).

Sequential Importance Sampling Methods. Sequential importance sampling (SIS) methods define $\overline{w}$ in terms of incremental importance sampling (IS) weights, that is, $\overline{w} = \prod_{t=1}^{T} G_t(\mathbf{x}_{t-1}, \mathbf{x}_t)$ with

$$G_t(\mathbf{x}_{t-1}, \mathbf{x}_t) = \frac{\gamma_t(\mathbf{x}_t)B^\theta_{t-1}(\mathbf{x}_{t-1}|\mathbf{x}_t)}{\gamma_{t-1}(\mathbf{x}_{t-1})F^\theta_t(\mathbf{x}_t|\mathbf{x}_{t-1})}, \tag{10}$$

with annealed versions γ_t of γ. For example, choosing $B^\theta_{t-1}(\mathbf{x}_{t-1}|\mathbf{x}_t) = \pi_t(\mathbf{x}_{t-1})F^\theta_t(\mathbf{x}_t|\mathbf{x}_{t-1})/\pi_t(\mathbf{x}_t)$ recovers AIS (Neal, 2001). Midgley et al. (2022) proposed to parameterize the proposal distribution π_0 with normalizing flows and, in combination with AIS, to minimize the α-divergence, resulting in the Flow Annealed Importance Sampling Bootstrap (FAB) algorithm. Additionally, when AIS is coupled with resampling, it gives rise to Sequential Monte Carlo (SMC) as originally proposed by Del Moral et al. (2006).

Recent advancements include the development of Stochastic Normalizing Flows (Wu et al., 2020a), Annealed Flow Transport (AFT) (Arbel et al., 2021), and Continual Repeated AFT (CRAFT) (Matthews et al., 2022). These methods extend Sequential Monte Carlo by employing sets of

	Funnel	Credit	Seeds	Cancer	Brownian	Ionosphere	Sonar	Digits	Fashion	LGCP	MoG	MoS
True $\log Z$	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗	✓	✓
Samples from π	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗	✓	✓
Mode descriptors ξ	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✓	✓
Dimensionality D	10	25	26	31	32	35	61	196	784	1600	$\mathbb{N}_+$	$\mathbb{N}_+$

Table 2: Target densities $\pi(\mathbf{x}) = \gamma(\mathbf{x})/Z$ considered in this work.

normalizing flows that define deterministic transport maps between neighboring distributions γ_t. For further details on $F_t^\theta, B_{t-1}^\theta$ and the corresponding G_t see Table 7. For an in-depth exploration of the commonalities and distinctions among these methods, please refer to (Matthews et al., 2022).

Diffusion-Based Methods. Diffusion-based methods typically build on stochastic differential equations (SDEs) with parameterized drift terms (Tzen & Raginsky, 2019), i.e.,

$$\begin{aligned} \mathrm{d}\mathbf{x}_t &= f_t^\theta(\mathbf{x}_t)\mathrm{d}t + \sigma_t \mathrm{d}\mathbf{w}_t, && \mathbf{x}_0 \sim \pi_0,\\ \mathrm{d}\mathbf{x}_t &= b_t^\theta(\mathbf{x}_t)\mathrm{d}t + \sigma_t \mathrm{d}\bar{\mathbf{w}}_t, && \mathbf{x}_T \sim \pi_T, \end{aligned} \tag{11}$$

with diffusion coefficient σ_t and standard Brownian motion $\mathbf{w}_t, \bar{\mathbf{w}}_t$. Using the Euler-Maruyama method (Särkkä & Solin, 2019), their discretized counterparts can be characterized by Gaussian forward-backward transition kernels

$$\begin{aligned} F_{t+1}^\theta(\mathbf{x}_{t+1}|\mathbf{x}_t) &= \mathcal{N}(\mathbf{x}_{t+1}|\mathbf{x}_t + f_t^\theta(\mathbf{x}_t)\Delta_t, \sigma_t^2\Delta_t) \text{ and}\\ B_{t-1}^\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) &= \mathcal{N}(\mathbf{x}_{t-1}|\mathbf{x}_t + b_t^\theta(\mathbf{x}_t)\Delta_t, \sigma_t^2\Delta_t), \end{aligned} \tag{12}$$

with discretization step size Δ_t. The extended (unnormalized) importance weights $\bar{w}$ can then be constructed as

$$\frac{p^\theta(\mathbf{x}_{0:T})}{q^\theta(\mathbf{x}_{0:T})} \propto \bar{w} = \frac{\gamma(\mathbf{x}_T)\prod_{t=1}^T B_{t-1}^\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{\pi_0(\mathbf{x}_0)\prod_{t=0}^{T-1} F_{t+1}^\theta(\mathbf{x}_{t+1}|\mathbf{x}_t)}. \tag{13}$$

One line of work considers annealed Langevin dynamics to model Eq. (11). Works include Unadjusted Langevin Annealing (ULA) (Thin et al., 2021), Monte Carlo Diffusions (MCD) (Doucet et al., 2022a), Controlled Monte Carlo Diffusion (CMCD) (Vargas et al., 2024), Uncorrected Hamiltonian Annealing (UHA) (Geffner & Domke, 2021) and Langevin Diffusion Variational Inference (LDVI) (Geffner & Domke, 2022). A second line of work describes diffusion-based sampling from a stochastic optimal control perspective (Dai Pra, 1991). Works include methods such as Path Integral Sampler (PIS) (Zhang & Chen, 2021; Vargas et al., 2023b), Denoising Diffusion Sampler (DDS) (Vargas et al., 2023a), Time-Reversed Diffusion Sampler (DIS) (Berner et al., 2022) and Generative Flow Networks (GFN) (Lahlou et al., 2023; Malkin et al., 2022b; Zhang et al., 2023). Furthermore, Richter et al. (2023) identify several of these methods as special cases of a General Bridge Sampler (GBS) where both processes in Eq. 11 are freely parameterized. Specific choices of $\pi_0, F_{t+1}^\theta, B_{t-1}^\theta$ are detailed in Table 6. Lastly, we refer the interested reader to (Sendera et al., 2024) which concurrently benchmarked diffusion-based sampling methods.

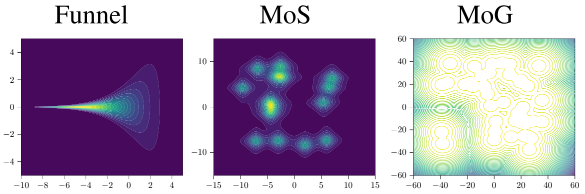


Figure 3: Synthetic target densities. Left: First two dimensions of the funnel density. Middle: Mixture of Student-t distribution with 15 components (MoS). Right: Mixture of 40 isotropic Gaussian distributions (MoG).

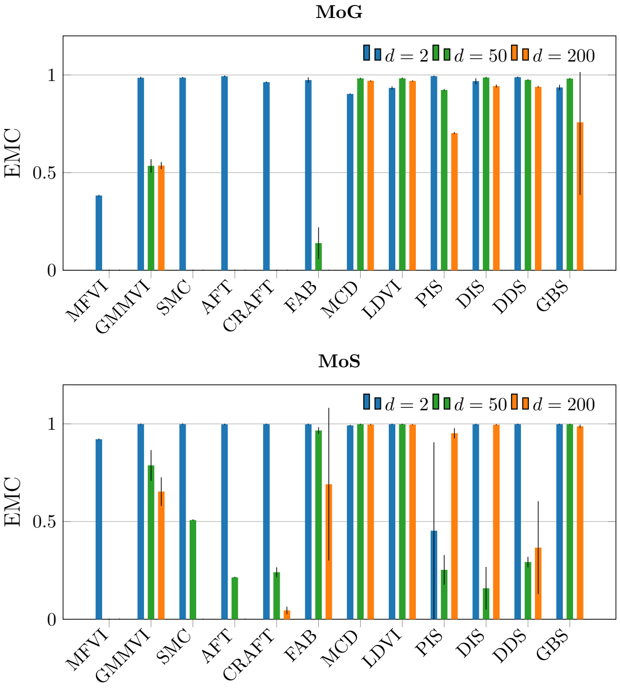


Figure 2: Mean and standard deviation of EMC values for MoG and MoS across varying dimensions d.

5. Benchmarking Target Densities

Here, we briefly summarize the target densities π considered in this work. The dimensionality of the problem, if we have access to the log normalizer $\log Z$, target samples, or mode descriptors for computing the entropic mode coverage is presented in Table 2. Further details and formal definitions of the target densities can be found in Appendix C.

Bayesian Logistic Regression. We consider four experiments where we perform inference over the parameters of a Bayesian logistic regression model for binary classification. The datasets *Credit* and *Cancer* were taken from Nishihara et al. (2014). The former distinguishes individuals as either good or bad credit risks, while the latter deals with the classification of recurrence events in breast cancer. The *Ionosphere* dataset (Sigillito et al., 1989) involves classifying radar signals passing through the ionosphere as either good or bad. Similarly, the *Sonar* dataset (Gorman & Sejnowski, 1988) tackles the classification of sonar signals bounced off a metal cylinder versus those bounced off a roughly cylindrical rock.

Random Effect Regression. The *Seeds* data was collected by (Crowder, 1978). The goal is to perform inference over the variables of a random effect regression model that models the germination proportion of seeds arranged in a factorial layout by seed and type of root.

Time Series Models. We consider the *Brownian* time series model obtained by discretizing a stochastic differential equation, modeling a Brownian motion with a Gaussian observation model, developed by (Sountsov et al., 2020).

Spatial Statistics. The log Gaussian Cox process (*LGCP*) (Møller et al., 1998) is a probabilistic model commonly used in statistics to model spatial point patterns. In this work, the log Gaussian Cox process is applied to modeling the positions of pine saplings in Finland.

Synthetic Targets. We additionally consider synthetic target densities as they commonly give access to the true normalization constant Z, target samples, and mode descriptors. The *Funnel* target was introduced by (Neal, 2003) and provides a complex 'funnel'-shaped distribution. Moreover, we consider two different types of mixture models: a mixture of isotropic Gaussians (*MoG*) as proposed by Midgley et al. (2022), and Student-t distributions (*MoS*). To obtain mode descriptors for a mixture model with K components, i.e., $\pi(\mathbf{x}) = \sum_k \pi_k(\mathbf{x})/K$ we compute the density per component $\pi_k(\mathbf{x})$ and say that $\mathbf{x} \in \xi_i$ if $i = \text{argmax}_k \{\pi_k(\mathbf{x})\}_{k=1}^K$. Lastly, we follow Doucet et al. (2022a) and train NICE (Dinh et al., 2014) on a down-sampled 14×14 variant of MNIST (*Digits*) (LeCun et al., 1998) and a 28×28 variant of Fashion MNIST (*Fashion*) and use the trained model as target density. Here, we obtain the mode descriptors by training a classifier $p(\mathbf{c}|\mathbf{x})$ on samples from π where the classes $\mathbf{c}$ are represented by ten different digits. If $i = \text{argmax}_{\mathbf{c}}\ p(\mathbf{c}|\mathbf{x})$ we conclude $\mathbf{x} \in \xi_i$.

6. Hyperparameters and Tuning

In this section, we provide details on hyperparameter tuning. For further information, please refer to Appendix D.

Tractable Density Methods. For MFVI, we used a batch size of 2000 and performed 100k gradient steps, tuning the learning rate via grid search. For targets with known support, we adjusted the initial model variance accordingly. For GMMVI, we adhered to the default settings from (Arenz et al., 2022), utilizing 100 samples per mixture component. We initialized with 10 components and employed an adaptive scheme to add and remove components heuristically. The initial variance of the components was set based on the target support, and we conducted 3000 training iterations.

Sequential Importance Sampling Methods. In SIS methods, we employed 2000 particles for training. All methods except FAB used 128 annealing steps; FAB followed the original 12 steps as proposed by its authors. The choice and parameters of the MCMC transition kernel significantly impacted performance. Hamilton-Monte Carlo (Duane et al., 1987) generally outperformed Metropolis-Hastings (Chib & Greenberg, 1995) (see Appendix F.3). Step sizes for $\beta_t \geq 0.5$ and $\beta_t < 0.5$ were tuned using grid search. For AFT and CRAFT, we used diagonal affine flows (Papamakarios et al., 2021), which yielded more robust results than complex flows like inverse autoregressive flows (Kingma et al., 2016) or neural spline flows (Durkan et al., 2019) (see Appendix F.6). FAB employed RealNVP (Dinh et al., 2016) for the proposal distribution π_0. Learning rates for these flows were also tuned via grid search. For targets with known support, the variance of $\pi_0(\mathbf{x}) = \mathcal{N}(0, \sigma_0^2 I)$ was set accordingly, otherwise, a grid search was performed. We used multinomial resampling with a threshold of 0.3 (Douc & Cappé, 2005).

Diffusion-based Methods. Training involved a batch size of 2000 and 40k gradient steps. SDEs were discretized with 128 steps, $T = 1$, and a fixed Δt. The diffusion coefficient was chosen as $\sigma_t = \sigma_{\text{max}} \cos^2(\pi(T-t)/2T)$, following (Vargas et al., 2023a) for better performance compared to linear or constant schedules. We used the architecture from (Zhang & Chen, 2021) with 2 layers of 64 hidden units each. For targets with known prior support, the initial model support was set accordingly. For all methods except PIS, this involved setting the variance of the prior distribution $\pi_0(\mathbf{x}) = \mathcal{N}(0, \sigma_0^2 I)$. For PIS, σ_{max} was carefully chosen. In MCD and LDVI, we learned the annealing schedule β_t and σ_{max} end-to-end by maximizing the ELBO.

	FUNNEL		MOG ($d = 50$)		MOS ($d = 50$)		14×14 DIGITS		28×28 FASHION	
	$\mathcal{W}_2 \downarrow$	MMD ↓	$\mathcal{W}_2 \downarrow$	MMD ↓	$\mathcal{W}_2 \downarrow$	MMD ↓	$\mathcal{W}_2 \downarrow$	MMD ↓	$\mathcal{W}_2 \downarrow$	MMD ↓
MFVI	$178.264_{\pm 0.271}$	$0.303_{\pm 0.002}$	$39360.196_{\pm 12.49}$	$0.209_{\pm 0.000}$	$2462.260_{\pm 1.009}$	$0.215_{\pm 0.000}$	$254.179_{\pm 0.025}$	$0.351_{\pm 0.000}$	$1327.517_{\pm 0.845}$	$0.285_{\pm 0.000}$
GMMVI	$\mathbf{105.620}_{\pm \mathbf{3.472}}$	$\mathbf{0.031}_{\pm \mathbf{0.000}}$	$32004.968_{\pm 1069.}$	$0.203_{\pm 0.013}$	$1255.216_{\pm 296.9}$	$0.135_{\pm 0.017}$	$207.163_{\pm 14.60}$	$0.373_{\pm 0.042}$	$1343.495_{\pm 136.9}$	$0.462_{\pm 0.033}$
SMC	$149.353_{\pm 2.973}$	$0.162_{\pm 0.015}$	$46351.236_{\pm 4.795}$	$0.631_{\pm 0.000}$	$3297.640_{\pm 1372.}$	$0.431_{\pm 0.161}$	$159.255_{\pm 1.877}$	$1.168_{\pm 0.008}$	$6696.287_{\pm 250.4}$	$1.556_{\pm 0.008}$
AFT	$145.138_{\pm 6.061}$	$0.159_{\pm 0.010}$	$44914.194_{\pm 1154.}$	$0.622_{\pm 0.009}$	$2648.410_{\pm 301.3}$	$0.395_{\pm 0.082}$	$172.685_{\pm 3.661}$	$1.180_{\pm 0.004}$	$6413.147_{\pm 548.6}$	$1.538_{\pm 0.010}$
CRAFT	$134.335_{\pm 0.663}$	$0.115_{\pm 0.003}$	$43412.038_{\pm 420.9}$	$0.604_{\pm 0.002}$	$1893.926_{\pm 117.3}$	$0.257_{\pm 0.024}$	$151.791_{\pm 11.02}$	$0.139_{\pm 0.032}$	$1413.303_{\pm 11.20}$	$0.562_{\pm 0.002}$
FAB	$153.894_{\pm 3.916}$	$\mathbf{0.032}_{\pm \mathbf{0.000}}$	$9567.319_{\pm 626.1}$	$0.073_{\pm 0.005}$	$\mathbf{1204.160}_{\pm \mathbf{147.7}}$	$\mathbf{0.093}_{\pm \mathbf{0.014}}$	$\mathbf{126.863}_{\pm \mathbf{0.581}}$	$\mathbf{0.129}_{\pm \mathbf{0.003}}$	$1186.967_{\pm 263.4}$	$0.347_{\pm 0.007}$
MCD	$163.317_{\pm 0.101}$	$0.228_{\pm 0.001}$	$5026.147_{\pm 40.03}$	$0.043_{\pm 0.000}$	$6418.981_{\pm 22.15}$	$0.256_{\pm 0.000}$	$220.710_{\pm 5.547}$	$0.252_{\pm 0.007}$	$1898.472_{\pm 3.783}$	$0.327_{\pm 0.002}$
LDVI	N/A	N/A	$5038.420_{\pm 73.77}$	$0.043_{\pm 0.000}$	$2919.688_{\pm 103.4}$	$0.182_{\pm 0.003}$	$154.167_{\pm 0.816}$	$0.133_{\pm 0.000}$	$3432.724_{\pm 406.2}$	$0.284_{\pm 0.016}$
PIS	N/A	N/A	$10495.164_{\pm 83.20}$	$0.083_{\pm 0.000}$	$2113.172_{\pm 31.17}$	$0.218_{\pm 0.007}$	$186.007_{\pm 0.466}$	$0.193_{\pm 0.001}$	$1484.598_{\pm 5.125}$	$0.240_{\pm 0.000}$
DIS	$118.947_{\pm 12.81}$	$0.159_{\pm 0.036}$	$\mathbf{3044.733}_{\pm \mathbf{464.7}}$	$\mathbf{0.034}_{\pm \mathbf{0.003}}$	$2200.590_{\pm 18.73}$	$0.155_{\pm 0.001}$	$220.392_{\pm 11.69}$	$0.194_{\pm 0.011}$	$3927.754_{\pm 858.9}$	$0.282_{\pm 0.019}$
DDS	$142.890_{\pm 9.552}$	$0.172_{\pm 0.031}$	$5551.107_{\pm 116.4}$	$0.046_{\pm 0.001}$	$2154.884_{\pm 3.861}$	$0.131_{\pm 0.001}$	$188.789_{\pm 2.297}$	$0.173_{\pm 0.003}$	$1811.685_{\pm 24.47}$	$\mathbf{0.208}_{\pm \mathbf{0.006}}$
GBS	$178.075_{\pm 0.103}$	$0.305_{\pm 0.002}$	$5080.413_{\pm 125.8}$	$0.043_{\pm 0.001}$	$5722.074_{\pm 22.71}$	$0.232_{\pm 0.000}$	$186.436_{\pm 1.834}$	$0.176_{\pm 0.005}$	$\mathbf{1137.399}_{\pm \mathbf{1.819}}$	$0.246_{\pm 0.003}$
	$\Delta \log Z_r \downarrow$	$\Delta \log Z_f \downarrow$	$\Delta \log Z_r \downarrow$	$\Delta \log Z_f \downarrow$	$\Delta \log Z_r \downarrow$	$\Delta \log Z_f \downarrow$	$\Delta \log Z_r \downarrow$	$\Delta \log Z_f \downarrow$	$\Delta \log Z_r \downarrow$	$\Delta \log Z_f \downarrow$
MFVI	$0.612_{\pm 0.101}$	$0.036_{\pm 0.001}$	$3.658_{\pm 0.040}$	$0.185_{\pm 0.002}$	$3.009_{\pm 0.291}$	$\mathbf{0.048}_{\pm \mathbf{0.002}}$	$7.388_{\pm 0.107}$	$5.866_{\pm 0.016}$	$34.389_{\pm 0.757}$	$108.379_{\pm 0.438}$
GMMVI	$\mathbf{0.001}_{\pm \mathbf{0.000}}$	$\mathbf{0.001}_{\pm \mathbf{0.000}}$	$\mathbf{1.715}_{\pm \mathbf{0.119}}$	$\mathbf{0.048}_{\pm \mathbf{0.007}}$	$1.282_{\pm 0.221}$	$0.084_{\pm 0.055}$	$3.098_{\pm 0.140}$	$\mathbf{0.124}_{\pm \mathbf{0.079}}$	$\mathbf{8.099}_{\pm \mathbf{1.919}}$	$\mathbf{11.676}_{\pm \mathbf{4.041}}$
SMC	$0.187_{\pm 0.054}$	$2.676_{\pm 0.000}$	$690.721_{\pm 11.21}$	$161.796_{\pm 0.000}$	$3.880_{\pm 1.105}$	$80.992_{\pm 0.000}$	$80.184_{\pm 0.162}$	$375.676_{\pm 0.000}$	$11742.014_{\pm 139.2}$	$1530.824_{\pm 0.000}$
AFT	$0.181_{\pm 0.106}$	$414.619_{\pm 141.5}$	$765.624_{\pm 108.0}$	$110.955_{\pm 18.37}$	$4.081_{\pm 1.579}$	$205.297_{\pm 23.91}$	$16.726_{\pm 2.511}$	$163.871_{\pm 6.557}$	$11653.343_{\pm 1628.}$	$1071.777_{\pm 9.475}$
CRAFT	$0.091_{\pm 0.018}$	$255.046_{\pm 7.478}$	$337.094_{\pm 9.296}$	$100.987_{\pm 0.065}$	$\mathbf{0.822}_{\pm \mathbf{0.087}}$	$210.245_{\pm 6.098}$	$1.458_{\pm 0.406}$	$63.792_{\pm 3.329}$	$445.101_{\pm 8.273}$	$1156.718_{\pm 7.810}$
FAB	$\mathbf{0.001}_{\pm \mathbf{0.000}}$	$0.019_{\pm 0.003}$	$2.952_{\pm 0.247}$	$126.363_{\pm 1.789}$	$3.358_{\pm 1.062}$	$84.592_{\pm 22.64}$	$\mathbf{0.847}_{\pm \mathbf{0.076}}$	$63.910_{\pm 1.565}$	$350.544_{\pm 599.0}$	$3721.720_{\pm 4646.}$
MCD	$0.207_{\pm 0.039}$	N/A	$31.319_{\pm 1.793}$	$21.148_{\pm 1.478}$	$28.607_{\pm 1.275}$	$24.757_{\pm 0.841}$	$884.610_{\pm 9.674}$	$258.840_{\pm 3.047}$	$15122.090_{\pm 996.7}$	$1125.475_{\pm 5.198}$
LDVI	N/A	N/A	$8.159_{\pm 0.775}$	$15.477_{\pm 0.815}$	$4.360_{\pm 0.741}$	$5.472_{\pm 0.938}$	$537.763_{\pm 25.07}$	$265.674_{\pm 1.181}$	$12237.989_{\pm 381.8}$	$1087.592_{\pm 4.844}$
PIS	$0.918_{\pm 0.598}$	$0.436_{\pm 0.002}$	$7.122_{\pm 0.630}$	$3113.492_{\pm 1.978}$	$12.248_{\pm 0.326}$	$54.090_{\pm 0.151}$	$104.002_{\pm 0.847}$	$2149.224_{\pm 19.39}$	$1884.013_{\pm 10.20}$	$8785.873_{\pm 9.880}$
DIS	$0.113_{\pm 0.083}$	$25.544_{\pm 8.267}$	$87.709_{\pm 8.942}$	$369.352_{\pm 16.29}$	$10.448_{\pm 0.607}$	$87.897_{\pm 5.255}$	$569.837_{\pm 35.40}$	$1354.472_{\pm 181.1}$	$8807.430_{\pm 337.6}$	$17566.520_{\pm 256.6}$
DDS	$0.190_{\pm 0.077}$	$0.321_{\pm 0.052}$	$1.739_{\pm 0.442}$	$207.545_{\pm 1.163}$	$7.952_{\pm 0.299}$	$53.411_{\pm 0.024}$	$82.460_{\pm 5.480}$	$659.497_{\pm 9.786}$	$1579.602_{\pm 41.65}$	$2910.345_{\pm 71.25}$
GBS	$0.553_{\pm 0.273}$	$0.127_{\pm 0.008}$	$8.103_{\pm 1.696}$	$9.321_{\pm 0.776}$	$53.767_{\pm 0.732}$	$47.441_{\pm 0.098}$	$75.160_{\pm 2.321}$	$62.733_{\pm 1.168}$	$1495.194_{\pm 42.03}$	$527.580_{\pm 9.426}$
	ELBO ↑	EUBO ↓	ELBO ↑	EUBO ↓	ELBO ↑	EUBO ↓	ELBO ↑	EUBO ↓	ELBO ↑	EUBO ↓
MFVI	$-1.834_{\pm 0.009}$	$105.694_{\pm 0.002}$	$-3.690_{\pm 0.000}$	$164.114_{\pm 0.000}$	$-5.957_{\pm 0.007}$	$72.663_{\pm 0.005}$	$-14.004_{\pm 0.005}$	$210.713_{\pm 0.024}$	$-58.082_{\pm 0.009}$	$938.632_{\pm 0.055}$
GMMVI	$\mathbf{-0.011}_{\pm \mathbf{0.001}}$	$\mathbf{0.012}_{\pm \mathbf{0.001}}$	$\mathbf{-1.715}_{\pm \mathbf{0.119}}$	$240.459_{\pm 51.13}$	$-3.890_{\pm 0.122}$	$57.746_{\pm 1.928}$	$\mathbf{-7.135}_{\pm \mathbf{0.148}}$	$142.636_{\pm 9.701}$	$\mathbf{-18.478}_{\pm \mathbf{4.104}}$	$595.239_{\pm 120.4}$
SMC	$-0.242_{\pm 0.047}$	$4.690_{\pm 0.000}$	$-877.034_{\pm 10.23}$	$161.921_{\pm 0.000}$	$-4.634_{\pm 1.088}$	$81.325_{\pm 0.000}$	$-185.057_{\pm 0.257}$	$376.093_{\pm 0.000}$	$-12187.873_{\pm 134.6}$	$1532.904_{\pm 0.000}$
AFT	$-0.293_{\pm 0.088}$	$431.329_{\pm 143.1}$	$-927.160_{\pm 103.8}$	$117.630_{\pm 22.16}$	$-4.923_{\pm 1.546}$	$207.625_{\pm 24.14}$	$-64.442_{\pm 4.464}$	$214.486_{\pm 4.870}$	$-11828.529_{\pm 1608.}$	$1448.335_{\pm 11.08}$
CRAFT	$-0.027_{\pm 0.060}$	$263.474_{\pm 7.864}$	$-451.399_{\pm 7.561}$	$103.674_{\pm 0.069}$	$\mathbf{-0.295}_{\pm \mathbf{0.256}}$	$212.210_{\pm 6.160}$	$-11.154_{\pm 0.307}$	$89.518_{\pm 1.904}$	$-520.475_{\pm 5.531}$	$1578.114_{\pm 2.360}$
FAB	$\mathbf{-0.014}_{\pm \mathbf{0.003}}$	$\mathbf{0.012}_{\pm \mathbf{0.002}}$	$-299.916_{\pm 253.4}$	$93.560_{\pm 5.086}$	$-26.496_{\pm 1.875}$	$\mathbf{18.088}_{\pm \mathbf{2.503}}$	$-11.396_{\pm 0.153}$	$\mathbf{12.084}_{\pm \mathbf{0.171}}$	$-892.971_{\pm 1518.}$	$\mathbf{394.346}_{\pm \mathbf{263.6}}$
MCD	$-0.611_{\pm 0.005}$	N/A	$-185.021_{\pm 0.743}$	$\mathbf{43.670}_{\pm \mathbf{0.457}}$	$-69.358_{\pm 0.633}$	$47.834_{\pm 0.820}$	$-1457.646_{\pm 13.80}$	$293.191_{\pm 0.208}$	$-21196.583_{\pm 472.8}$	$1276.456_{\pm 1.033}$
LDVI	N/A	N/A	$-29.034_{\pm 0.591}$	$51.137_{\pm 0.177}$	$-28.471_{\pm 1.018}$	$20.887_{\pm 1.042}$	$-875.104_{\pm 43.59}$	$323.158_{\pm 0.142}$	$-16227.975_{\pm 738.0}$	$1185.331_{\pm 6.660}$
PIS	$-3.198_{\pm 0.042}$	$104.975_{\pm 0.002}$	$-16.881_{\pm 0.026}$	$3626.120_{\pm 1.360}$	$-29.261_{\pm 1.743}$	$88.192_{\pm 0.005}$	$-172.988_{\pm 0.630}$	$2748.938_{\pm 19.19}$	$-2988.210_{\pm 14.13}$	$11179.374_{\pm 11.72}$
DIS	$-1.021_{\pm 0.436}$	$40.892_{\pm 38.48}$	$-181.348_{\pm 15.47}$	$546.335_{\pm 30.86}$	$-36.704_{\pm 0.629}$	$193.270_{\pm 3.293}$	$-840.122_{\pm 18.66}$	$1745.719_{\pm 205.7}$	$-15337.229_{\pm 154.0}$	$20347.781_{\pm 318.7}$
DDS	$-0.597_{\pm 0.142}$	$148.841_{\pm 7.347}$	$-13.284_{\pm 0.460}$	$291.867_{\pm 0.047}$	$-31.681_{\pm 0.363}$	$86.014_{\pm 0.001}$	$-156.145_{\pm 6.063}$	$881.476_{\pm 22.83}$	$-2617.761_{\pm 46.78}$	$3925.231_{\pm 106.8}$
GBS	$-2.600_{\pm 0.078}$	$110.167_{\pm 0.000}$	$-35.771_{\pm 1.105}$	$67.819_{\pm 2.157}$	$-99.369_{\pm 0.158}$	$73.545_{\pm 0.107}$	$-154.186_{\pm 1.387}$	$106.777_{\pm 0.113}$	$-2198.997_{\pm 36.56}$	$705.996_{\pm 11.66}$

Table 3: Results for various sampling methods. Evaluation criteria include 2-Wasserstein distance ($\mathcal{W}_2$), maximum mean discrepancy (MMD), reverse and forward partition function error ($\Delta \log Z_r$, $\Delta \log Z_f$), and lower and upper evidence bounds (ELBO, EUBO). The best results are highlighted in bold. Arrows (↑, ↓) indicate whether higher or lower values are preferable, respectively. N/A denotes cases where reasonable results could not be obtained due to numerical issues.

7. Experiments

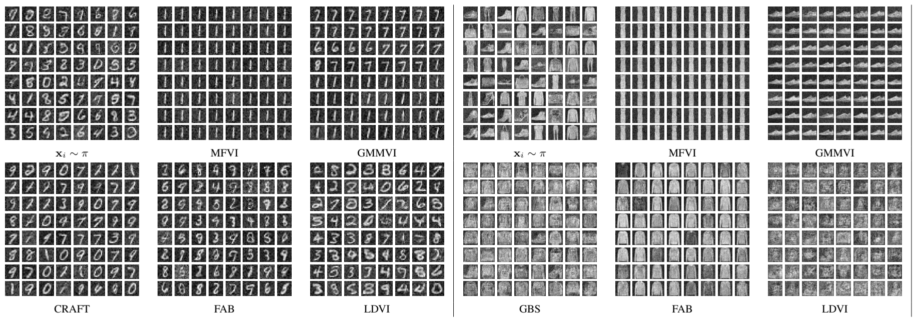
Here, we offer an overview of the evaluation protocol. Next, we present the results obtained for synthetic target densities, followed by those for real targets. We provide further results in Appendix E and ablation studies in Appendix F.

Evaluation Protocol. We compute all performance criteria 100 times during training, applying a running average with a length of 5 over these evaluations to obtain robust results within a single run. To ensure robustness across runs, we use four different random seeds and average the best results from each run. We use 2000 samples to compute the performance criteria and tune key hyperparameters such as the learning rate and variance of the initial proposal distribution π_0. We report the EMC values corresponding to the method's highest ELBO value to avoid high EMC values caused by a large initial support of the model.

7.1. Evaluation on Synthetic Target Densities

Funnel. We utilize the funnel distribution as a testing ground to assess whether sampling methods capture high curvatures in the target density. Our findings indicate that while most methods successfully capture the funnel-like structure, they struggle to generate samples at the neck and opening of the funnel, except for FAB and GMMVI (cf. Figure 4). This observation is further supported by quantitative analysis, revealing that both FAB and GMMVI achieve the best performance in terms of reverse and forward estimation of $\log Z$ and evidence bounds as shown in Table 3.

Digits and Fashion. For a comprehensive assessment of sampling methods, we conduct both qualitative and quantitative analyses on two high-dimensional target densities. For the qualitative analysis, model samples are interpreted as images and shown in Table 4. For the quantitative analysis, we report various performance criteria values, with results presented in Table 3. Additionally, we report EMC values in Table 4 to quantify mode collapse.


$\mathbf{x}_i \sim \pi$ MFVI GMMVI $\mathbf{x}_i \sim \pi$ MFVI GMMVI

CRAFT FAB LDVI GBS FAB LDVI

EMC ↑	14×14 **Digits**	28×28 **Fashion**
MFVI	0.000 ± 0.000	0.000 ± 0.000
GMMVI	0.164 ± 0.081	0.217 ± 0.167
SMC	0.873 ± 0.000	0.000 ± 0.000
AFT	0.727 ± 0.000	0.011 ± 0.000
CRAFT	0.772 ± 0.070	0.016 ± 0.027
FAB	0.915 ± 0.007	0.349 ± 0.137
MCD	0.851 ± 0.010	0.619 ± 0.001
LDVI	$\mathbf{0.951 \pm 0.002}$	0.608 ± 0.005
PIS	0.816 ± 0.011	0.620 ± 0.004
DIS	0.818 ± 0.009	0.612 ± 0.008
DDS	0.816 ± 0.012	0.621 ± 0.008
GBS	0.796 ± 0.005	$\mathbf{0.621 \pm 0.006}$

Table 4: Sample visualizations for *Digits* (left) and *Fashion* (middle) using various methods, as indicated by the subcaptions. '$\mathbf{x}_i \sim \pi$' refers to samples from the target density. Visualizations for the remaining methods are provided in Figure 5. Corresponding EMC values are reported on the right.

For *Digits*, most methods are able to find the majority of modes and produce high-quality samples, as visually evident from the sample visualizations and EMC values in Table 4. However, many methods, particularly diffusion-based ones, struggle to obtain reasonable estimations of $\log Z$. They also perform poorly in terms of lower and upper evidence bounds, as shown in Table 3. For *Fashion*, we observe that methods either suffer from mode collapse or produce low-quality samples. Interestingly, the methods experiencing mode collapse achieve the lowest estimation error of $\log Z$ in both reverse and forward estimations.

Mixture Models. We employ *MoG* and *MoS* to investigate mode collapse across different dimensions, specifically considering $d \in 2, 50, 200$. For $d = 2$, all methods except MFVI demonstrate the capability to generate samples from all modes, as indicated by EMC ≈ 1. This is further supported by visualizations in Figure 2. According to EMC, all methods except diffusion-based ones exhibit mode collapse for $d = 50$ and $d = 200$.

We also report additional evaluation criteria for *MoG* and *MoS*, including 2-Wasserstein distance, maximum mean discrepancy, reverse and forward partition function error, lower and upper evidence bounds, and reverse and forward effective sample size in Appendix E Table 9.

7.2. Evaluation on Real Target Densities

For real-world target densities, we do not have access to the ground truth normalizer Z or samples from π. Consequently, we present the ELBO values in Table 5. Surprisingly, we find that GMMVI performs well across all tasks, often outperforming more complex variational Monte Carlo methods. However, it is noteworthy that GMMVI encounters scalability challenges in very high-dimensional problems, such as LGCP. Another method, FAB, consistently performs well across a majority of tasks.

8. Discussion and Conclusion

Here, we list several general observations **O1)-O6)** and observations tied to specific methods **M1)-M6)** that are based on the experiments from Section 7 and Appendix E and the Ablation studies in Appendix F.

O1) Mode collapse gets worse in high dimensions. We observe that several methods, that do not suffer from mode collapse in low-dimensional problems encounter significant mode collapse when applied to higher-dimensional ones (cf. Fig 2).

O2) ELBO and reverse $\log Z$ estimates are not well suited for evaluating a model's capability to avoid mode collapse. This observation is evident, for instance, in Table 4, where MFVI achieves relatively good ELBO and $\log Z$ estimates despite suffering from mode collapse.

O3) While the EUBO helps to quantify mode collapse, comparing different method categories is challenging due to the additional looseness introduced by latent variables in the extended EUBO. This is evident on the *Fashion* dataset, where MFVI and GMMVI achieve a lower EUBO compared to most other methods, despite suffering from mode collapse.

O4) Despite being influenced by subjective design choices like the kernel or cost function, the 2-Wasserstein distance and Maximum Mean Discrepancy (MMD) generally show consistent performance across different sampling methods, as demonstrated in Table 3. Additionally, the quantitative results frequently align with the qualitative outcomes. For instance, this alignment is evident from GMMVI samples on *Funnel* or the GBS samples on the *Fashion*.

O5) For multimodal target distributions, both forward and reverse ESS tend to exhibit a 'binary' pattern, frequently taking values of 0 or 1. Forward ESS, in particular, often tends to be predominantly zero for higher dimensional problems, further complicating the evaluation of mode collapse severity. In contrast, EUBO and ELBO offer a more continuous and informative perspective in assessing model performance

ELBO ↑	CREDIT	SEEDS	CANCER	BROWNIAN	IONOSPHERE	SONAR	LGCP
MFVI	−524.859±0.035	−76.733±0.012	−29.407±0.557	−3.872±0.012	−123.419±0.040	−137.672±0.043	383.18±0.059
GMMVI	**−504.487±0.001**	**−73.415±0.002**	**121.442±5.591**	**1.092±0.006**	−111.832±0.007	−108.726±0.007	OOM
SMC	−580.936±15.915	−74.699±0.100	−67.959±4.345	−1.874±0.622	−114.751±0.238	−111.355±1.177	393.907±5.727
AFT	−584.766±13.979	−74.269±0.090	−15.515±5.100	N/A	−113.272±0.647	−110.671±1.240	394.271±6.432
CRAFT	−573.387±17.59	−73.793±0.015	19.283±0.523	0.886±0.053	−112.386±0.182	−115.618±1.316	**495.291±0.384**
FAB	**−504.496±0.001**	**−73.418±0.002**	39.922±8.200	1.031±0.010	**−111.678±0.003**	**−108.593±0.008**	402.212±0.941
MCD	N/A	−73.652±0.003	N/A	0.643±0.012	−111.942±0.006	−109.534±0.055	444.313±0.452
LDVI	N/A	−73.530±0.003	N/A	0.772±0.016	−111.788±0.003	−108.841±0.006	161.839±1.436
PIS	−846.568±2.417	−88.919±2.051	39.542±5.302	N/A	−125.030±0.688	−142.868±3.289	479.542±0.403
DDS	−514.736±1.223	−75.206±0.209	19.997±0.690	0.561±0.228	−114.191±0.105	−121.222±5.985	N/A
GBS	−508.108±0.145	−88.778±0.109	−23.495±0.737	N/A	−133.777±0.152	−153.094±0.500	N/A

Table 5: ELBO values for various target densities. The best results are highlighted in bold. N/A denotes cases where reasonable results could not be obtained due to numerical issues. OOM refers to problems caused by memory constraints.

(cf. Appendix E, Table 9).

O6) No single method exhibits superiority across all situations. Generally, GMMVI and FAB demonstrate good ELBO values across a diverse set of tasks, although both tend to suffer from mode collapse in high dimensions. In contrast, diffusion-based methods such as MCD and LDVI exhibit resilience against mode collapse but frequently fall short of achieving satisfactory ELBO values.

M1) Resampling causes mode collapse in high dimensions (cf. Ablation F.3). SIS methods, in particular, experience severe mode collapse in high dimensions, as illustrated in Figure 2. Notably, eliminating the resampling step in Sequential Monte Carlo (SMC) proves effective in mitigating this issue, but results in worse ELBO values.

M2) There exists an exploration-exploitation trade-off when setting the support of the proposal distribution π_0 in Variational Monte Carlo (cf. Ablation F.4). Opting for a small initial support of π_0 results in tight ELBO values but can limit coverage to only a few modes. Conversely, employing a sufficiently large initial support helps prevent mode collapse but introduces additional looseness in the ELBO.

M3) Learning the proposal distribution π_0 in Variational Monte Carlo methods often leads to mode collapse, especially in high dimensions. Training the base distribution end-to-end by maximizing the extended ELBO or pre-training the base distribution, for example, using methods like MFVI, results in mode collapse, as indicated in Ablation F.5 and Ablation F.8. Despite the occurrence of mode collapse, these strategies yield higher ELBOs, emphasizing the inherent exploration-exploitation trade-off discussed in **M2)**.

M4) Variational Monte Carlo methods heavily benefit from using a large number of steps T. This is shown in Ablation F.2, where increasing the annealing steps for SIS methods and discretization steps for diffusion-based methods leads to tighter evidence bounds. However, increasing T results in prolonged computational runtimes and demands substantial memory resources.

M5) GMMVI exhibits high sample efficiency (cf. Table 10). Arenz et al. (2022) employ a replay buffer to enhance the sample efficiency of GMMVI, leading to orders of magnitude fewer target evaluations required for convergence. Consequently, GMMVI may be the preferable choice when target evaluations are time-consuming.

M6) Langevin diffusion methods demonstrate low sample efficiency, as highlighted in Table 10. These methods require evaluating the target at each intermediate discretization step due to the score function being part of the SDE, and they typically need many iterations to converge. Other diffusion-based methods that do not require target evaluations at every step, such as DDS, often perform poorly and suffer from mode collapse (cf. Ablation F.7). To address this, Zhang & Chen (2021) proposed incorporating the score function into the network architecture, resulting again in poor sample efficiency.

9. Conclusion

In this work, we assessed the latest sampling methods using a standardized set of tasks. Our exploration encompassed various performance criteria, with a specific focus on quantifying mode collapse. Through a comprehensive evaluation, we illuminated the strengths and weaknesses of state-of-the-art sampling methods, thereby offering a valuable reference for future developments.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgments and Disclosure of Funding

We thank Julius Berner, Lorenz Richter, and Vincent Stimper for many useful discussions. D.B. acknowledges support by funding from the pilot program Core Informatics of the Helmholtz Association (HGF) and the state of Baden-Württemberg through bwHPC, as well as the HoreKa supercomputer funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the German Federal Ministry of Education and Research.

References

Agakov, F. V. and Barber, D. An auxiliary variational method. In *Advances in Neural Information Processing Systems*, 2004.

Akhound-Sadegh, T., Rector-Brooks, J., Bose, A. J., Mittal, S., Lemos, P., Liu, C.-H., Sendera, M., Ravanbakhsh, S., Gidel, G., Bengio, Y., et al. Iterated denoising energy matching for sampling from boltzmann densities. *arXiv preprint arXiv:2402.06121*, 2024.

Anderson, B. D. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Arbel, M., Matthews, A., and Doucet, A. Annealed flow transport monte carlo. In *International Conference on Machine Learning*, pp. 318–330. PMLR, 2021.

Arenz, O., Neumann, G., and Zhong, M. Efficient gradient-free variational inference using policy search. In *International conference on machine learning*, pp. 234–243. PMLR, 2018.

Arenz, O., Dahlinger, P., Ye, Z., Volpp, M., and Neumann, G. A unified perspective on natural gradient variational inference with gaussian mixture models. *arXiv preprint arXiv:2209.11533*, 2022.

Aronszajn, N. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Berner, J., Richter, L., and Ullrich, K. An optimal control perspective on diffusion-based generative modeling. *arXiv preprint arXiv:2211.01364*, 2022.

Bishop, C. Pattern recognition and machine learning. *Springer google schola*, 2:531–537, 2006.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Chib, S. and Greenberg, E. Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4): 327–335, 1995.

Cover, T. M. *Elements of information theory*. John Wiley & Sons, 1999.

Crowder, M. J. Beta-binomial anova for proportions. *Applied statistics*, pp. 34–37, 1978.

Dai Pra, P. A stochastic control approach to reciprocal diffusion processes. *Applied mathematics and Optimization*, 23(1):313–329, 1991.

Del Moral, P., Doucet, A., and Jasra, A. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

Dinh, L., Krueger, D., and Bengio, Y. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Douc, R. and Cappé, O. Comparison of resampling schemes for particle filtering. In *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005.*, pp. 64–69. Ieee, 2005.

Doucet, A., Grathwohl, W., Matthews, A. G., and Strathmann, H. Score-based diffusion meets annealed importance sampling. *Advances in Neural Information Processing Systems*, 35:21482–21494, 2022a.

Doucet, A., Grathwohl, W., Matthews, A. G. d. G., and Strathmann, H. Score-based diffusion meets annealed importance sampling. In *Advances in Neural Information Processing Systems*, 2022b.

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

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. Neural spline flows. *Advances in neural information processing systems*, 32, 2019.

Frenkel, D. and Smit, B. *Understanding molecular simulation: from algorithms to applications*. Elsevier, 2023.

Geffner, T. and Domke, J. MCMC variational inference via uncorrected Hamiltonian annealing. In *Advances in Neural Information Processing Systems*, 2021.

Geffner, T. and Domke, J. Langevin diffusion variational inference. *arXiv preprint arXiv:2208.07743*, 2022.

Gorman, R. P. and Sejnowski, T. J. Analysis of hidden units in a layered network trained to classify sonar targets. *Neural networks*, 1(1):75–89, 1988.

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.

Hammersley, J. *Monte carlo methods*. Springer Science & Business Media, 2013.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jankowiak, M. and Phan, D. Surrogate likelihoods for variational annealed importance sampling. In *International Conference on Machine Learning*, pp. 9881–9901. PMLR, 2022.

Ji, C. and Shen, H. Stochastic variational inference via upper bound. *arXiv preprint arXiv:1912.00650*, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29, 2016.

Lahlou, S., Deleu, T., Lemos, P., Zhang, D., Volokhova, A., Hernández-Garcıa, A., Ezzine, L. N., Bengio, Y., and Malkin, N. A theory of continuous generative flow networks. In *International Conference on Machine Learning*, pp. 18269–18300. PMLR, 2023.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, pp. 2278–2324, 1998. doi: 10.1109/5.726791.

Léonard, C. A survey of the schr\" odinger problem and some of its connections with optimal transport. *arXiv preprint arXiv:1308.0215*, 2013.

Li, Y. and Turner, R. E. Rényi divergence variational inference. *Advances in neural information processing systems*, 29, 2016.

Liu, J. S. and Liu, J. S. *Monte Carlo strategies in scientific computing*, volume 75. Springer, 2001.

Malkin, N., Jain, M., Bengio, E., Sun, C., and Bengio, Y. Trajectory balance: Improved credit assignment in gflownets. *Advances in Neural Information Processing Systems*, 35:5955–5967, 2022a.

Malkin, N., Lahlou, S., Deleu, T., Ji, X., Hu, E., Everett, K., Zhang, D., and Bengio, Y. Gflownets and variational inference. *arXiv preprint arXiv:2210.00580*, 2022b.

Matthews, A., Arbel, M., Rezende, D. J., and Doucet, A. Continual repeated annealed flow transport monte carlo. In *International Conference on Machine Learning*, pp. 15196–15219. PMLR, 2022.

Midgley, L. I., Stimper, V., Simm, G. N., Schölkopf, B., and Hernández-Lobato, J. M. Flow annealed importance sampling bootstrap. *arXiv preprint arXiv:2208.01893*, 2022.

Midgley, L. I., Stimper, V., Antorán, J., Mathieu, E., Schölkopf, B., and Hernández-Lobato, J. M. Se (3) equivariant augmented coupling flows. *arXiv preprint arXiv:2308.10364*, 2023.

Mittal, S., Bracher, N. L., Lajoie, G., Jaini, P., and Brubaker, M. A. Exploring exchangeable dataset amortization for bayesian posterior inference. In *ICML 2023 Workshop on Structured Probabilistic Inference* {\&} *Generative Modeling*, 2023.

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. Log gaussian cox processes. *Scandinavian journal of statistics*, 25(3):451–482, 1998.

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

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

Nishihara, R., Murray, I., and Adams, R. P. Parallel mcmc with generalized elliptical slice sampling. *The Journal of Machine Learning Research*, 15(1):2087–2112, 2014.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22(1):2617–2680, 2021.

Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Richter, L., Boustati, A., Nüsken, N., Ruiz, F., and Akyildiz, O. D. Vargrad: a low-variance gradient estimator for variational inference. *Advances in Neural Information Processing Systems*, 33:13481–13492, 2020.

Richter, L., Berner, J., and Liu, G.-H. Improved sampling via learned diffusions. *arXiv preprint arXiv:2307.01198*, 2023.

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

Särkkä, S. and Solin, A. *Applied stochastic differential equations*, volume 10. Cambridge University Press, 2019.

Sendera, M., Kim, M., Mittal, S., Lemos, P., Scimeca, L., Rector-Brooks, J., Adam, A., Bengio, Y., and Malkin, N. On diffusion models for amortized inference: Benchmarking and improving stochastic control and sampling. *arXiv preprint arXiv:2402.05098*, 2024.

Shapiro, A. Monte carlo sampling methods. *Handbooks in operations research and management science*, 10:353–425, 2003.

Sigillito, V. G., Wing, S. P., Hutton, L. V., and Baker, K. B. Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest*, 10(3):262–266, 1989.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Sountsov, P., Radul, A., and contributors. Inference gym, 2020. URL https://pypi.org/project/inference_gym.

Stoltz, G., Rousset, M., et al. *Free energy computations: A mathematical perspective*. World Scientific, 2010.

Thin, A., Kotelevskii, N., Durmus, A., Moulines, E., Panov, M., and Doucet, A. Monte Carlo variational autoencoders. In *International Conference on Machine Learning*, 2021.

Tzen, B. and Raginsky, M. Neural stochastic differential equations: Deep latent gaussian models in the diffusion limit. *arXiv preprint arXiv:1905.09883*, 2019.

Vargas, F., Grathwohl, W., and Doucet, A. Denoising diffusion samplers. *arXiv preprint arXiv:2302.13834*, 2023a.

Vargas, F., Ovsianas, A., Fernandes, D., Girolami, M., Lawrence, N. D., and Nüsken, N. Bayesian learning via neural schrödinger–föllmer flows. *Statistics and Computing*, 33(1):3, 2023b.

Vargas, F., Padhy, S., Blessing, D., and Nüsken, N. Transport meets variational inference: Controlled monte carlo diffusions. In *The Twelfth International Conference on Learning Representations*, 2024.

Wainwright, M. J. and Jordan, M. I. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, November 2008.

Wan, N., Li, D., and Hovakimyan, N. F-divergence variational inference. *Advances in neural information processing systems*, 33:17370–17379, 2020.

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

Wu, H., Köhler, J., and Noé, F. Stochastic normalizing flows. In *Advances in Neural Information Processing Systems*, 2020b.

Zhang, D., Chen, R. T. Q., Liu, C.-H., Courville, A., and Bengio, Y. Diffusion generative flow samplers: Improving learning signals through partial trajectory optimization. *arXiv preprint arXiv:2310.02679*, 2023.

Zhang, G., Hsu, K., Li, J., Finn, C., and Grosse, R. Differentiable annealed importance sampling and the perils of gradient noise. In *Advances in Neural Information Processing Systems*, 2021.

Zhang, Q. and Chen, Y. Path integral sampler: a stochastic control approach for sampling. *arXiv preprint arXiv:2111.15141*, 2021.

A. Performance Criteria Details

Here, we provide further details on the computation of the various performance criteria introduced in the main manuscript.

A.1. Density-Ratio-Based Criteria

Forward and Reverse Importance-Weighted Estimation of Z. Using the definition of the normalization constant, the importance-weighted reverse estimate of Z is given by

$$Z_r := \int \gamma(\mathbf{x})\mathrm{d}\mathbf{x} = \int \frac{q^\theta(\mathbf{x})}{q^\theta(\mathbf{x})}\gamma(\mathbf{x}) = \mathbb{E}_{q^\theta}\left[\frac{\gamma(\mathbf{x})}{q^\theta(\mathbf{x})}\right] \approx \frac{1}{N_{q^\theta}} \sum_{\mathbf{x}_i \sim q^\theta} \frac{\gamma(\mathbf{x}_i)}{q^\theta(\mathbf{x}_i)} \tag{14}$$

where N_{q^θ} denotes the number of samples from q^θ used for the Monte Carlo estimate of the expectation. Using the identity $Z^{-1} = \pi(\mathbf{x})/\gamma(\mathbf{x})$, we obtain the forward estimation of Z as

$$Z^{-1} = \int Z^{-1} q^\theta(\mathbf{x})\mathrm{d}\mathbf{x} = \mathbb{E}_\pi\left[\frac{q^\theta(\mathbf{x})}{\gamma(\mathbf{x})}\right], \text{ and thus, } Z_f := 1/\mathbb{E}_\pi\left[\frac{q^\theta(\mathbf{x})}{\gamma(\mathbf{x})}\right] \approx 1/\Big(\frac{1}{N_\pi}\sum_{\mathbf{x}_i\sim\pi}\frac{q^\theta(\mathbf{x}_i)}{\gamma(\mathbf{x}_i)}\Big), \tag{15}$$

where N_π denotes the number of samples from π used for the Monte Carlo estimate of the expectation.

Forward and Reverse Effective Sample Size. The (reverse) effective sample size (ESS), or equivalently, reverse ESS (Shapiro, 2003) is defined as

$$\text{ESS}_r := 1/\mathbb{E}_{q^\theta}\left[\Big(\frac{\pi(\mathbf{x})}{q^\theta(\mathbf{x})}\Big)^2\right] = Z_r^2/\mathbb{E}_{q^\theta}\left[\Big(\frac{\gamma(\mathbf{x})}{q^\theta(\mathbf{x})}\Big)^2\right] = \Big(\mathbb{E}_{q^\theta}\left[\frac{\gamma(\mathbf{x})}{q^\theta(\mathbf{x})}\right]\Big)^2/\mathbb{E}_{q^\theta}\left[\Big(\frac{\gamma(\mathbf{x})}{q^\theta(\mathbf{x})}\Big)^2\right], \tag{16}$$

where Z is approximated using the reverse estimate as defined in Eq. 14. Using the definition of the ESS, it is straightforward to see that

$$\text{ESS}_f := 1/\mathbb{E}_{q^\theta}\left[\Big(\frac{\pi(\mathbf{x})}{q^\theta(\mathbf{x})}\Big)^2\right] = 1/\mathbb{E}_\pi\left[\frac{\pi(\mathbf{x})}{q^\theta(\mathbf{x})}\right] = Z_f/\mathbb{E}_\pi\left[\frac{\gamma(\mathbf{x})}{q^\theta(\mathbf{x})}\right] = \mathbb{E}_\pi\left[\frac{q^\theta(\mathbf{x})}{\gamma(\mathbf{x})}\right]^{-1}/\ \mathbb{E}_\pi\left[\frac{\gamma(\mathbf{x})}{q^\theta(\mathbf{x})}\right], \tag{17}$$

where Z is approximated using the forward estimate as defined in Eq. 15.

A.2. Integral Probability Metrics

Maximum Mean Discrepancy. The Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) is a kernel-based measure of distance between two distributions. The MMD quantifies the dissimilarity between these distributions by comparing their mean embeddings in a reproducing kernel Hilbert space (RKHS) (Aronszajn, 1950) with kernel k. In our setting, we are interested in computing the MMD between a model q^θ and target distribution π. Formally, if $\mathcal{H}_k$ is the RKHS associated with kernel function k, the MMD between q^θ and π is the integral probability metric defined by:

$$\text{MMD}_k(q^\theta,\pi) = \sup_{f\in\mathcal{H}_k:\|f\|_{\mathcal{H}_k}\le 1} \big(\mathbb{E}_{\mathbf{x}\sim q^\theta}[f(\mathbf{x})] - \mathbb{E}_{\mathbf{y}\sim\pi}[f(\mathbf{y})]\big), \tag{18}$$

with $\text{MMD}_k(q^\theta,\pi) \ge 0$ and $\text{MMD}_k(q^\theta,\pi) = 0$ if and only if $q^\theta = \pi$. The minimum variance unbiased estimate of MMD_k between two sample sets $\mathbf{X} \sim q^\theta$ and $\mathbf{Y} \sim \pi$ with sizes n and m respectively is given by

$$\text{MMD}_k(q^\theta,\pi) \approx \sqrt{\frac{1}{n(n-1)}\sum_{i,j}^{n} k(\mathbf{x}_i,\mathbf{x}_j) + \frac{1}{m(m-1)}\sum_{i,j}^{m} k(\mathbf{y}_i,\mathbf{y}_j) - \frac{2}{nm}\sum_{i}^{n}\sum_{j}^{m} k(\mathbf{x}_i,\mathbf{y}_j)}, \tag{19}$$

In our experiments, we took a squared exponential kernel given by $k(\mathbf{x},\mathbf{y}) = \exp\big(-\|\mathbf{x}-\mathbf{y}\|_2^2/\alpha\big)$, where the bandwidth α is determined using the median heuristic (Gretton et al., 2012). The code for computing the MMD was built upon `https://github.com/antoninschrab/mmdfuse-paper`.

Entropic Optimal Transport Distance. The 2-Wasserstein distance is given by

$$W_2(q^\theta,\pi) = \inf\left\{\int_{\mathbb{R}^d\times\mathbb{R}^d} c(\mathbf{x},\mathbf{y})\xi(\mathbf{x},\mathbf{y})\mathrm{d}x\mathrm{d}y : \int_{\mathbb{R}^d}\xi(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{y} = q^\theta(\mathbf{x}),\ \int_{\mathbb{R}^d}\xi(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{x} = \pi(\mathbf{y})\right\}^{1/2}, \tag{20}$$

with cost c, chosen as $c(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|^2$ in our experiments. To obtain a tractable objective, an entropy regularized version has been proposed (Peyré et al., 2019), that is,

$$W_{2,\varepsilon}(q^\theta,\pi) = \inf\left\{ \int_{\mathbb{R}^d\times\mathbb{R}^d} c(\mathbf{x},\mathbf{y})\xi(\mathbf{x},\mathbf{y})\mathrm{d}x\mathrm{d}y - \varepsilon\mathcal{H}(\xi) : \int_{\mathbb{R}^d} \xi(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{y} = q^\theta(\mathbf{x}),\ \int_{\mathbb{R}^d} \xi(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{x} = \pi(\mathbf{y}) \right\}^{1/2}. \quad (21)$$

with entropy $\mathcal{H}(\xi) = -\int_{\mathbb{R}^d\times\mathbb{R}^d} \xi(\mathbf{x},\mathbf{y})\log\xi(\mathbf{x},\mathbf{y})\mathrm{d}x\mathrm{d}y$. We chose $\varepsilon = 10^{-3}$ for all experiments. The code was taken from `https://github.com/ott-jax/ott`.

A.3. Extending the Entropic Mode Coverage

If the true mode probabilities $p^*(\xi|\mathbf{x})$ are not uniformly distributed, EMC=1 does not correspond to the optimal value. In that case, we propose the expected Jensen-Shannon divergence, that is,

$$\text{EJS} := \mathbb{E}_{q^\theta(\mathbf{x})} D_{\text{JS}}(p(\xi|\mathbf{x})\|p^*(\xi|\mathbf{x})), \quad (22)$$

with

$$D_{\text{JS}}(p(\xi|\mathbf{x})\|p^*(\xi|\mathbf{x})) = \frac{1}{2}D_{\text{KL}}\left(p(\xi|\mathbf{x})\|\frac{p^*(\xi|\mathbf{x})+p(\xi|\mathbf{x})}{2}\right) + \frac{1}{2}D_{\text{KL}}\left(p^*(\xi|\mathbf{x})\|\frac{p^*(\xi|\mathbf{x})+p(\xi|\mathbf{x})}{2}\right), \quad (23)$$

as an alternative heuristic to quantify mode collapse. Similar to EMC, EJS is bounded and is straightforward to interpret: When employing the base 2 logarithm, EJS remains bounded, i.e., $0 \leq \text{EJS}. \leq 1$. Moreover $\text{EJS} = 0$ implies that the model matches the potentially unbalanced true mode probabilities, while $\text{EJS} = 1$ indicates that p and p^* possess no overlapping probability mass.

B. Details on Unnormalized Importance Weights / Density Ratios

Here, we provide further details on how the unnormalized importance weights / density ratios are computed for different methods.

Tractable Density Methods. For models with tractable density $q^\theta(\mathbf{x})$ the marginal (unnormalized) importance weights can trivially computed using

$$w = \frac{\gamma(\mathbf{x})}{q^\theta(\mathbf{x})}.$$

Diffusion-based Methods. For diffusion-based methods, the extended importance weights can then be constructed as

$$\frac{p^\theta(\mathbf{x}_{0:T})}{q^\theta(\mathbf{x}_{0:T})} = \frac{\pi(\mathbf{x}_T)\prod_{t=1}^T B_{t-1}^\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{\pi_0(\mathbf{x}_0)\prod_{t=0}^{T-1} F_{t+1}^\theta(\mathbf{x}_{t+1}|\mathbf{x}_t)}. \quad (24)$$

The different choices of forward and backward transition kernels $F_{t+1}^\theta, B_{t-1}^\theta$ are listed in Table 6. Some methods such as DDS (Vargas et al., 2023a), PIS (Zhang & Chen, 2021) and GFN (Zhang et al., 2023) introduce a reference process p^{ref} with

$$p^{\text{ref}}(\mathbf{x}_{0:T}) = p_0^{\text{ref}}(\mathbf{x}_0)\prod_{t=0}^{T-1} F_{t+1}^{\text{ref}}(\mathbf{x}_{t+1}|\mathbf{x}_t) = p_T^{\text{ref}}(\mathbf{x}_T)\prod_{t=1}^{T} B_{t-1}^\theta(\mathbf{x}_{t-1}|\mathbf{x}_t). \quad (25)$$

This allows for rewriting Eq. 24 as

$$\frac{p^\theta(\mathbf{x}_{0:T})}{q^\theta(\mathbf{x}_{0:T})} = \frac{p^\theta(\mathbf{x}_{0:T})}{p^{\text{ref}}(\mathbf{x}_{0:T})}\cdot\frac{p^{\text{ref}}(\mathbf{x}_{0:T})}{q^\theta(\mathbf{x}_{0:T})} = \frac{\pi(\mathbf{x}_T)}{p_T^{\text{ref}}(\mathbf{x}_T)}\cdot\frac{p^{\text{ref}}(\mathbf{x}_{0:T})}{q^\theta(\mathbf{x}_{0:T})}, \quad (26)$$

potentially resulting in more tractable density ratios compared to Eq. 24. For concrete examples see e.g. (Zhang et al., 2023). A continuous-time analogous of the reference process is detailed in (Vargas et al., 2024). Moreover, in continuous-time, the importance weights correspond to a Radon–Nikodym derivative. For the sake of simplicity, we only consider the discrete-time setting in this work. We refer the reader to (Vargas et al., 2024; Richter et al., 2023) for further details.

Method	$\pi_0(\mathbf{x}_0)$	$F^\theta_{t+1}(\mathbf{x}_{t+1}\mid\mathbf{x}_t)$	$B^\theta_{t-1}(\mathbf{x}_{t-1}\mid\mathbf{x}_t)$
DDS	$\mathcal{N}(\mathbf{x}_0\mid 0, \sigma_0^2\mathbf{I})$	$\mathcal{N}(\mathbf{x}_{t+1}\mid(\sqrt{1-\beta_t}\mathbf{x}_t + \mathbf{s}^\theta(\mathbf{x}_t, t))\Delta_t, \beta_t\sigma_0^2\Delta_t)$	$\mathcal{N}(\mathbf{x}_{t-1}\mid\sqrt{1-\beta_t}\mathbf{x}_t\Delta_t, \beta_t\sigma_0^2\Delta_t)$
DIS	$\mathcal{N}(\mathbf{x}_0\mid 0, \sigma_0^2\mathbf{I})$	$\mathcal{N}(\mathbf{x}_{t+1}\mid\mathbf{x}_t + (\beta_t\mathbf{x}_t + \mathbf{s}^\theta(\mathbf{x}_t, t))\Delta_t, 2\beta_t\sigma_0^2\Delta_t)$	$\mathcal{N}(\mathbf{x}_{t-1}\mid(\mathbf{x}_t - \beta_t\mathbf{x}_t)\Delta_t, 2\beta_t\sigma_0^2\Delta_t)$
PIS/GFN	δ_0	$\mathcal{N}(\mathbf{x}_{t+1}\mid\mathbf{x}_t + \mathbf{s}^\theta(\mathbf{x}_t, t)\Delta_t, \sigma_t^2\Delta_t)$	$\mathcal{N}(\mathbf{x}_{t-1}\mid\frac{t-\Delta_t}{t}\mathbf{x}_t, \frac{t-\Delta_t}{t}\sigma_t^2\Delta_t)$
ULA	arbitrary*	$\mathcal{N}(\mathbf{x}_{t+1}\mid\mathbf{x}_t + \nabla_{\mathbf{x}_t}\sigma_t^2\log\pi_t(\mathbf{x}_t)\Delta_t, \sigma_t^2\Delta_t)$	$\mathcal{N}(\mathbf{x}_{t-1}\mid\mathbf{x}_t + \nabla_{\mathbf{x}_t}\sigma_t^2\log\pi_t(\mathbf{x}_t)\Delta_t, \sigma_t^2\Delta_t)$
MCD	arbitrary*	$\mathcal{N}(\mathbf{x}_{t+1}\mid\mathbf{x}_t + \nabla_{\mathbf{x}_t}\sigma_t^2\log\pi_t(\mathbf{x}_t)\Delta_t, \sigma_t^2\Delta_t)$	$\mathcal{N}(\mathbf{x}_{t-1}\mid\mathbf{x}_t + (\nabla_{\mathbf{x}_t}\sigma_t^2\log\pi_t(\mathbf{x}_t) + \mathbf{s}^\theta(\mathbf{x}_t, t))\Delta_t, \sigma_t^2\Delta_t)$
CMCD	arbitrary*	$\mathcal{N}(\mathbf{x}_{t+1}\mid\mathbf{x}_t + (\nabla_{\mathbf{x}_t}\sigma_t^2\log\pi_t(\mathbf{x}_t) + \mathbf{s}^\theta(\mathbf{x}_t, t))\Delta_t, \sigma_t^2\Delta_t)$	$\mathcal{N}(\mathbf{x}_{t-1}\mid\mathbf{x}_t + (\nabla_{\mathbf{x}_t}\sigma_t^2\log\pi_t(\mathbf{x}_t) - \mathbf{s}^\theta(\mathbf{x}_t, t))\Delta_t, \sigma_t^2\Delta_t)$
GBS	arbitrary*	$\mathcal{N}(\mathbf{x}_{t+1}\mid\mathbf{x}_t + (\nabla_{\mathbf{x}_t}\sigma_t^2\mathbf{f}^\theta(\mathbf{x}_t, t))\Delta_t, \sigma_t^2\Delta_t)$	$\mathcal{N}(\mathbf{x}_{t-1}\mid\mathbf{x}_t + (\nabla_{\mathbf{x}_t}\sigma_t^2\mathbf{b}^\theta(\mathbf{x}_t, t))\Delta_t, \sigma_t^2\Delta_t)$

Table 6: Characterization of diffusion-based sampling methods. Here, $\mathbf{s}^\theta, \mathbf{f}^\theta, \mathbf{b}^\theta : \mathbb{R}^d \times [0, T] \to \mathbb{R}^d$ denotes a parameterized function approximator. * In our experiments, we choose $\pi_0(\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_0\mid 0, \sigma_0^2\Delta_t)$.

Method	$B^\theta_{t-1}(\mathbf{x}_{t-1}\mid\mathbf{x}_t)$	$G_t(\mathbf{x}_{t-1}, \mathbf{x}_t)$	$F^\theta_t(\mathbf{x}_t\mid\mathbf{x}_{t-1})$	$G_t(\mathbf{x}_{t-1}, \mathbf{x}_t)$
Optimal	$\pi_{t-1}(\mathbf{x}_{t-1})F^\theta_t(\mathbf{x}_t\mid\mathbf{x}_{t-1})/\pi_t(\mathbf{x}_t)$	Z_t/Z_{t-1}	$\pi_t(\mathbf{x}_t)B^\theta_{t-1}(\mathbf{x}_{t-1}\mid\mathbf{x}_t)/\pi_{t-1}(\mathbf{x}_{t-1})$	Z_t/Z_{t-1}
AIS/SMC/FAB	$\pi_t(\mathbf{x}_{t-1})F^\theta_t(\mathbf{x}_t\mid\mathbf{x}_{t-1})/\pi_t(\mathbf{x}_t)$	$\gamma_t(\mathbf{x}_{t-1})/\gamma_{t-1}(\mathbf{x}_{t-1})$	$\pi_{t-1}(\mathbf{x}_t)B^\theta_{t-1}(\mathbf{x}_{t-1}\mid\mathbf{x}_t)/\pi_{t-1}(\mathbf{x}_{t-1})$	$\gamma_t(\mathbf{x}_t)/\gamma_{t-1}(\mathbf{x}_t)$
AFT/CRAFT	$\delta_{T^\theta_t(x_t)}(x_{t-1})$	$\gamma_t(T^\theta_t(\mathbf{x}_{t-1}))\lvert\det\nabla T^\theta_t(\mathbf{x}_{t-1})\rvert/\gamma_{t-1}(\mathbf{x}_{t-1})$	$\delta_{(T^\theta_{t-1})^{-1}(x_{t-1})}(x_t)$	$\gamma_t(\mathbf{x}_t)/\gamma_{t-1}((T^\theta_{t-1})^{-1}(\mathbf{x}_t))\lvert\det\nabla(T^\theta_{t-1})^{-1}(\mathbf{x}_t)\rvert$

Table 7: Characterization of Sequential Importance Methods methods: The middle column shows the backward kernels B^θ_{t-1} and the corresponding G_t when transporting samples from the prior π_0 to the target π_T to compute reverse criteria. The right-most column shows the forward kernels F^θ_t and the corresponding G_t when transporting samples from the target π_T back to the prior π_T to compute forward criteria.

Sequential Importance Sampling Methods. Sequential importance sampling methods express the importance weights in terms of incremental importance sampling weights, i.e.,

$$\overline{w} = \prod_{t=1}^{T} G_t(x_{t-1}, x_t) \qquad \text{with} \qquad G_t(\mathbf{x}_{t-1}, \mathbf{x}_t) = \frac{\gamma_t(\mathbf{x}_t)B^\theta_{t-1}(\mathbf{x}_{t-1}|\mathbf{x}_t)}{\gamma_{t-1}(\mathbf{x}_{t-1})F^\theta_t(\mathbf{x}_t|\mathbf{x}_{t-1})}.$$

For given forward transitions F^θ_t, the optimal backward transitions B^θ_{t-1} ensure that $\overline{w} = w$. As the optimal transitions are typically not available, SMC uses the AIS approximation (Neal, 2001). Moreover flow transport methods (Wu et al., 2020a; Arbel et al., 2021; Matthews et al., 2022) use a flow as a deterministic map T^θ to approximate the incremental IS weights. In Table 7, we list different $F^\theta_t, B^\theta_{t-1}$ and their corresponding incremental importance sampling weights.

C. Benchmark Target Details

Here, we introduce the target densities considered in this benchmark more formally.

C.1. Bayesian Logistic Regression

We used four binary classification problems in our benchmark, which have also been used in various other work to compare different state-of-the-art methods in variational inference and Markov chain Monte Carlo. We assess the performance of a Bayesian logistic model with:

$$\begin{aligned} &\mathbf{x} \sim \mathcal{N}\left(0, \sigma_w^2 I\right), \\ &y_i \sim \text{Bernoulli}(\text{sigmoid}(\mathbf{x}^\top u_i)) \end{aligned}$$

on two standardized datasets $\{(u_i, y_i)\}_i$, namely **Ionosphere** ($d = 35$) with 351 data points and **Sonar** ($d = 61$) with 208 data points.

The **German Credit** dataset consists of ($d = 25$) features and 1000 data points, while the **Breast Cancer** dataset has ($d = 31$) dimensions with 569 data points, which we standardize and apply linear logistic regression.

C.2. Random Effect Regression

The **Seeds** ($d = 26$) target is a random effect regression model trained on the *seeds* dataset:

$$\begin{aligned}
&\tau \sim \text{Gamma}(0.01, 0.01) \\
&a_0, a_1, a_2, a_{12} \sim \mathcal{N}(0, 10) \\
&b_i \sim \mathcal{N}\left(0, \frac{1}{\sqrt{\tau}}\right), \quad i = 1, \dots, 21 \\
&\text{logits}_i = a_0 + a_1 x_i + a_2 y_i + a_{12} x_i y_i + b_1, \quad i = 1, \dots, 21 \\
&r_i \sim \text{Binomial}\left(\text{logits}_i, N_i\right), \quad i = 1, \dots, 21.
\end{aligned}$$

The goal is to do inference over the variables $\tau, a_0, a_1, a_2, a_{12}$ and b_i for $i = 1, \dots, 21$, given observed values for x_i, y_i and N_i.

C.3. Time Series Models

The **Brownian** ($d = 32$) model corresponds to the time discretization of a Brownian motion:

$$\begin{aligned}
\alpha_{\text{inn}} &\sim \text{LogNormal}(0, 2), \\
\alpha_{\text{obs}} &\sim \text{LogNormal}(0, 2), \\
x_1 &\sim \mathcal{N}(0, \alpha_{\text{inn}}), \\
x_i &\sim \mathcal{N}(x_{i-1}, \alpha_{\text{inn}}), \quad i = 2, \dots 20, \\
y_i &\sim \mathcal{N}(x_i, \alpha_{\text{obs}}), \quad i = 1, \dots 30.
\end{aligned}$$

inference is performed over the variables $\alpha_{\text{inn}}, \alpha_{\text{obs}}$ and $\{x_i\}_{i=1}^{30}$ given the observations $\{y_i\}_{i=1}^{10} \cup \{y_i\}_{i=20}^{30}$.

C.4. Spatial Statistics

The **Log Gaussian Cox process** ($d = 1600$) is a popular high-dimensional task in spatial statistics (Møller et al., 1998) which models the position of pine saplings. Using a $d = M \times M = 1600$ grid, we obtain the unnormalized target density by

$$\mathcal{N}(\mathbf{x}; \mu, K) \prod_{i \in [1:M]^2} \exp\left(x_i y_i - a \exp\left(x_i\right)\right).$$

C.5. Synthetic Targets

We evaluate on three different mixture models which all follow the structure, that is,

$$\begin{aligned}
\pi(\mathbf{x}) = &\sum_{k=1}^{K} w_k \pi_k(\mathbf{x}), \\
&\sum_{k=1}^{K} w_k = 1,
\end{aligned}$$

where K denotes the number of components.

The **MoG** ($d = \mathbb{N}$) distribution, taken from (Midgley et al., 2022), consists of $K = 40$ mixture components with

$$\begin{aligned}
&\pi_k(\mathbf{x}) = \mathcal{N}(\mu_k, I) \\
&\mu_k \sim \mathcal{U}(-40, 40) \\
&w_k = 1/K,
\end{aligned}$$

where $\mathcal{U}(l, u)$ refers to a uniform distribution on $[l, u]$.

The **MoS** ($d = \mathbb{N}$) comprises 10 Student's t-distributions t_2, where the 2 refers to the degree of freedom. Generally, Student's t-distributions have heavier tails compared to Gaussian distributions, making them sharper and more challenging to approximate.

$$\pi_k(\mathbf{x}) = t_2 + \mu_k,$$
$$\mu_k \sim \mathcal{U}(-10, 10),$$
$$w_k = 1/K,$$

where μ_k refers to the translation of the individual components.

The **Funnel** ($d = 10$) target introduced in (Neal, 2003) is a challenging funnel-shaped distribution given by

$$\pi(\mathbf{x}) = \mathcal{N}(x_1; 0, \sigma_f^2)\mathcal{N}(x_{2:10}; 0, \exp(x_1)I),$$

with $\sigma_f^2 = 9$.

Lastly, we follow Doucet et al. (2022a) and use NICE (Dinh et al., 2014) to train a normalizing flow on a 14×14 and 28×28 variant of MNIST (**DIGITS**) and on the 28×28 Fashion MNIST dataset (**Fashion**).

D. Algorithms and Parameter Choices

Here, we discuss the parameter choices of all methods. Most of these choices are based on recommendations of the authors. For some choices, we run ablation studies to find suitable values.

Gaussian Mean Field Variational Inference (MFVI). We updated the mean and the diagonal covariance matrix using the Adam optimizer (Kingma & Ba, 2014) for $100k$ iterations with a batch size of 2000. We ensured non-negativeness of the variance by using a $\log$ transformation. The mean is initialized at 0 for all experiments. The initial covariance/scale and the learning rate are set according to Table 8.

Gaussian Mixture Model Variational Inference (GMMVI). For GMMVI, we ported the tensorflow implementation of `https://github.com/OlegArenz/gmmvi` to Jax and integrated it into our framework. We use the specifications (Arenz et al., 2022) described as SAMTRUX. We make use of their adaptive component initializer and start using ten components. The initial variance of the components is set according to Table 8.

Sequential Monte Carlo (SMC). For the Sequential Monte Carlo (SMC) approach, we leveraged the codebase available at `https://github.com/google-deepmind/annealed_flow_transport`. We used 2000 particles and 128 annealing steps (temperatures) T. We used resampling with a threshold of 0.3. We used one Hamiltonian Monte Carlo (HMC) step per temperature with 10 leapfrog steps. We tuned the stepsize of HMC according to Table 8 where we used different stepsizes depending on the annealing parameter β_t. We additionally tune the scale of the initial proposal distribution π_0 as shown in Table 8.

Continual Repeated Annealed Flow Transport (CRAFT/AFT). As AFT and CRAFT build on Sequential Monte Carlo (SMC), we employed the same SMC specifications detailed above. Notably, we found that employing simpler flows in conjunction with a greater number of temperatures yielded superior and more robust performance compared to the use of more sophisticated flows such as RealNVP or Neural Spline Flows. Consequently, we opted for 128 temperatures, utilizing diagonal affine flows as the transport map. Specifically for AFT, we determined that 400 iterations per temperature were sufficient to achieve converged training results. For CRAFT and SNF, we found that a total of 3000 iterations provided satisfactory convergence during training. For all methods, we use 2000 particles for training and testing and tune the learning rate and the scale of the initial proposal distribution π_0 as shown in Table 8.

Flow Annealed Importance Sampling Bootstrap (FAB). We built our implementation off of `https://github.com/lollcat/fab-jax`. We adjusted the parameters of FAB in accordance with the author's most important hyperparameter suggestions and to ensure that SMC performs reasonably well. To achieve this, we set the number of temperatures to 128 and used HMC as MCMC kernel. For the flow architecture we use RealNVP (Dinh et al., 2016) where the conditioner is given by an 8-layer MLP. Furthermore, we utilized FAB's replay buffer to speed up computations. The learning rate and base distribution scale are adjusted for target specificity, following the specifications outlined in Table 8. We used a batch size of 2048 and trained FAB for 3000 iterations which proved sufficient for achieving a satisfactory convergence.

Denoising Diffusion Sampler (DDS) and Path Integral Sampler (PIS). We use the implementation of `https://github.com/franciscovargas/denoising_diffusion_samplers` to integrate the Diffusion and Path

Integral Sampler into our framework. We set the parameters of the SDEs according to the authors, i.e., (Zhang & Chen, 2021) and (Vargas et al., 2023a) and use 128 timesteps and a batch size of 2000 if not otherwise stated. Both methods use the network proposed in (Zhang & Chen, 2021) which uses a sinusoidal position embedding for the timestep and uses the gradient of the log target density as an additional term. As proposed, we use a two-layer neural network with 64 hidden units. For DDS we use a cosine scheduler (Vargas et al., 2023a) and for PIS a uniform time scheduler (Zhang & Chen, 2021). Both methods were trained using $40k$ iterations.

Monte Carlo Diffusions (MCD) and Langevin Diffusion Variational Inference (LDVI). We build our implementation of Langevin Diffusion methods on `https://github.com/tomsons22/LDVI`. For experiments where performance is solely measured in terms of ELBO, due to the lack of samples from π or access to Z, we train all parameters of the SDE by maximizing the EUBO as suggested by (Geffner & Domke, 2021). For multimodal target densities, we fix the proposal distribution and the magnitude of the timestep. We found that this stabilizes training and yields better results (cf. Ablation 15). We use the network architecture proposed by (Zhang & Chen, 2021) with two hidden layers with 64 hidden units each. We discretize the SDEs using 128 timesteps and a batchsize of 2000 if not otherwise stated. All methods were trained using $40k$ iterations.

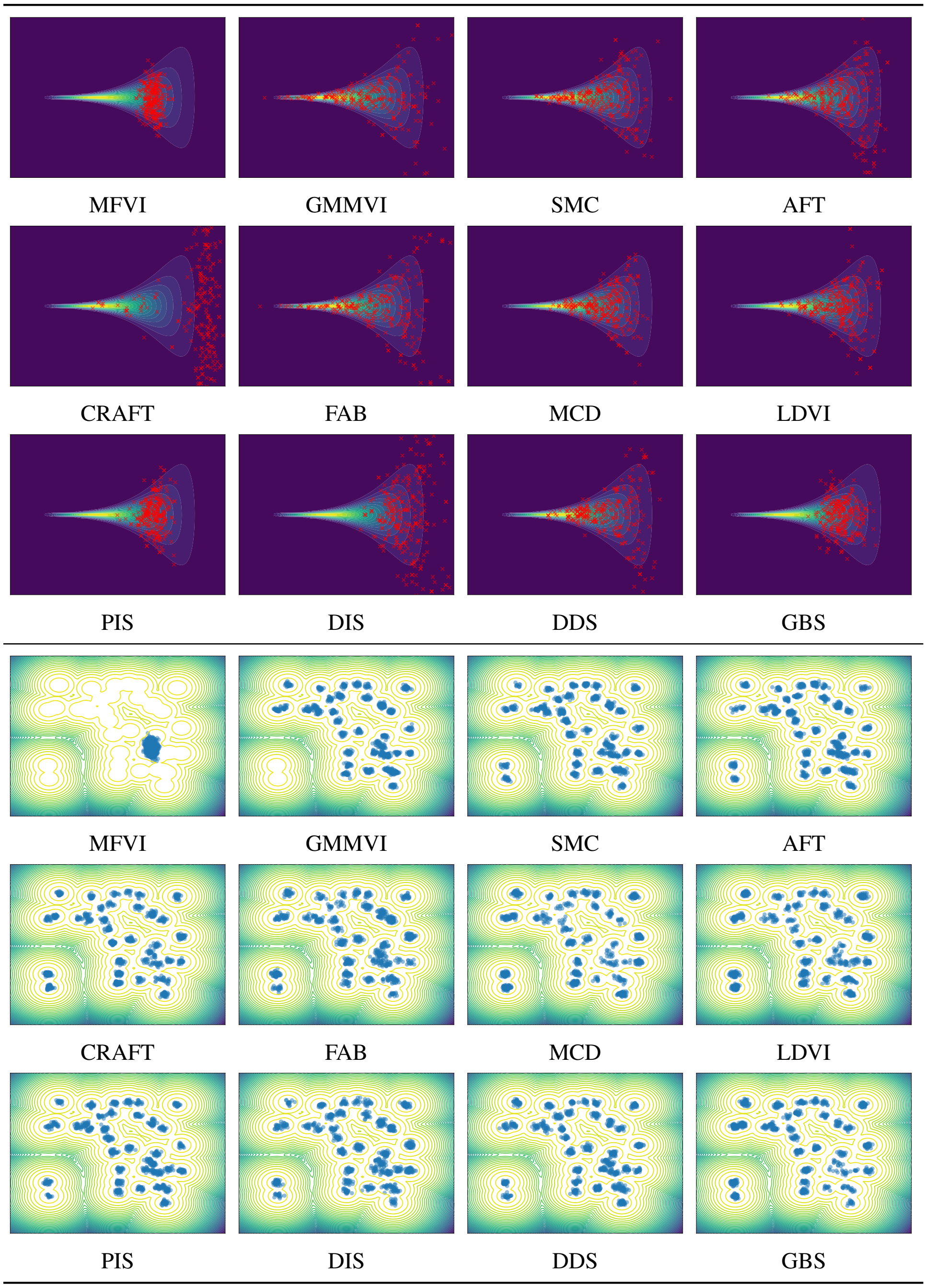
Time-Reversed Diffusion Sampler (DIS) and General Bridge Samples (GBS). We base the implementation of DIS and GBS on `https://github.com/juliusberner/sde_sampler` and implemented them in Jax. The remaining parameters follow the description of DDS and PIS above.

Methods / Parameters	Grid	MoG	MoS	Funnel	Digits/Fashion	Credit	Cancer	Brownian	Sonar	Seeds	Ionosphere	LGCP
MFVI												
Initial Scale	{0.1, 1, 10}	-	-	1	0.1	0.1	10	1	0.1	1	1	0.1
Learning Rate	$\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$	10^{-2}	10^{-3}	10^{-2}	10^{-4}	10^{-3}	10^{-2}	10^{-3}	10^{-3}	10^{-4}	10^{-4}	10^{-3}
GMMVI												
Initial Scale	{0.1, 1, 10}	-	-	0.1	10	0.1	1	0.1	1	0.1	10	NA
SMC												
Initial Scale	{0.1, 1, 10}	-	-	1	1	0.1	1	1	1	1	1	1
HMC stepsize ($\beta \leq 0.5$)	{0.001, 0.01, 0.05, 0.1, 0.2}	0.2	0.2	0.001	0.2	0.1	0.05	0.001	0.05	0.2	0.2	0.01
HMC stepsize ($\beta > 0.5$)	{0.001, 0.01, 0.05, 0.1, 0.2}	0.001	0.2	0.1	0.2	0.1	0.01	0.05	0.001	0.05	0.2	0.2
AFT												
Initial Scale	{0.1, 1, 10}	-	-	1	-	0.1	0.1	NA	1	1	1	1
Learning Rate	$\{10^{-3}, 10^{-4}, 10^{-5}\}$	10^{-3}	10^{-4}	10^{-3}	10^{-3}	10^{-3}	10^{-3}	NA	10^{-3}	10^{-4}	10^{-3}	10^{-3}
CRAFT												
Initial Scale	{0.1, 1, 10}	-	-	1	-	0.1	1	1	1	0.1	0.1	1
Learning Rate	$\{10^{-3}, 10^{-4}, 10^{-5}\}$	10^{-3}	10^{-4}	10^{-3}	10^{-4}	10^{-4}	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}
FAB												
Initial Scale	{0.1, 1, 10}	-	-	1	-	0.1	0.1	1	0.1	0.1	1	0.1
Learning Rate	$\{10^{-3}, 10^{-4}, 10^{-5}\}$	10^{-5}	-	10^{-4}	10^{-3}	10^{-4}	10^{-3}	10^{-3}	10^{-3}	10^{-4}	10^{-3}	10^{-3}
DDS/DIS												
Initial Scale	{0.1, 1, 10}	-	-	1	-	1	1	0.1	1	0.1	1	0.1
Learning Rate	$\{10^{-3}, 10^{-4}, 10^{-5}\}$	10^{-4}	10^{-4}	10^{-3}	10^{-3}	10^{-3}	10^{-4}	10^{-4}	10^{-3}	10^{-4}	10^{-4}	10^{-4}
PIS												
Learning Rate	$\{10^{-3}, 10^{-4}, 10^{-5}\}$	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-4}	10^{-4}	10^{-4}	10^{-3}	10^{-3}	10^{-4}	NA
LDVI												
Initial Scale	{0.1, 1, 10}	-	-	1	-	0.1	0.1	1	1	0.1	0.1	0.1
Learning Rate	$\{10^{-3}, 10^{-4}, 10^{-5}\}$	10^{-3}	10^{-3}	10^{-4}	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}
MCD												
Initial Scale	{0.1, 1, 10}	-	-	1	-	0.1	0.1	1	0.1	0.1	1	1
Learning Rate	$\{10^{-3}, 10^{-4}, 10^{-5}\}$	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-4}	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}

Table 8: Hyperparameter selection for all different sampling algorithms. The ‘Grid’ column indicates the values over which we performed a grid search. The values in the column which are marked with experiment names indicate which values were chosen for the reported results.

E. Further Experimental results

We additionally provide sample visualizations for *Funnel* and *MoG* in Figure 4, and *Digits* and *Fashion* in Figure 5. We also report additional evaluation criteria for *MoG* and *MoS*, including 2-Wasserstein distance, maximum mean discrepancy, reverse and forward partition function error, lower and upper evidence bounds, and reverse and forward effective sample size in Table 9. Lastly, we provide insights into the models efficiency by providing values for the number of target queries and wallclock time needed, for obtaining the best ELBO value. These results are shown in Table 10.



MFVI GMMVI SMC AFT

CRAFT FAB MCD LDVI

PIS DIS DDS GBS

MFVI GMMVI SMC AFT

CRAFT FAB MCD LDVI

PIS DIS DDS GBS

Figure 4: Visualization of samples drawn from different sampling methods for Funnel (top) and MoG (bottom).

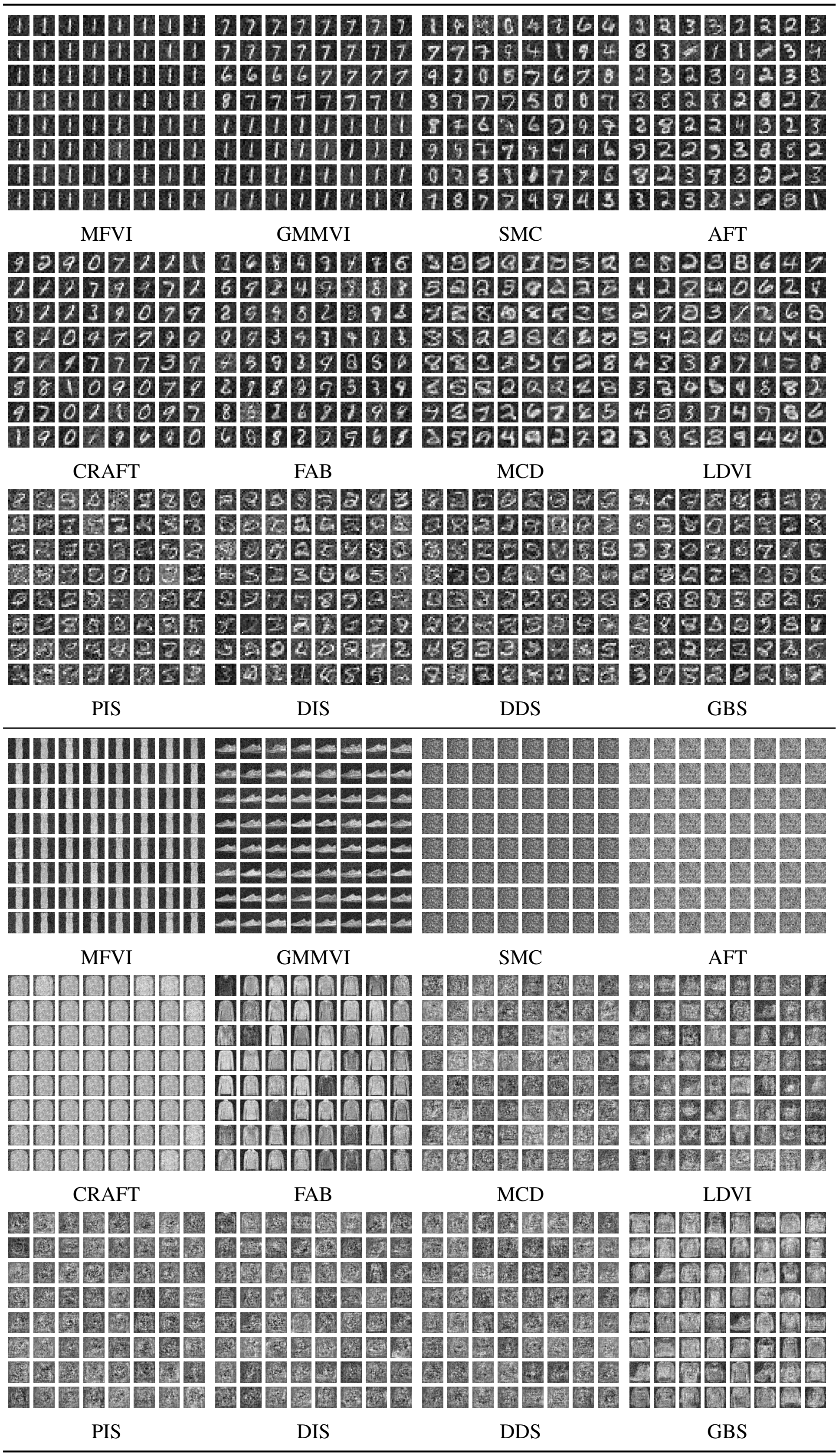


Figure 5: Visualization of samples drawn from different sampling methods for Digits (top) and Fashion (bottom).

	MoG						MoS					
	$d = 2$	$d = 50$	$d = 200$	$d = 2$	$d = 50$	$d = 200$	$d = 2$	$d = 50$	$d = 200$	$d = 2$	$d = 50$	$d = 200$
	$\mathcal{W}_2$ ↓			MMD ↓			$\mathcal{W}_2$ ↓			MMD ↓		
MFVI	506.967±7.385	36158.898±8.765	148945.539±14.42	0.251±0.002	0.209±0.000	0.211±0.000	24.688±0.225	2282.540±1.959	12956.415±6.530	0.162±0.001	0.187±0.001	0.195±0.000
GMMVI	76.474±20.60	31983.344±1065.	140166.746±3020.	0.052±0.010	0.202±0.013	0.214±0.012	2.851±0.128	1249.010±297.3	10402.243±870.9	0.036±0.000	0.133±0.018	0.211±0.026
SMC	32.387±9.219	46351.236±4.795	176586.789±3.638	0.047±0.004	0.631±0.000	0.611±0.000	34.963±2.833	3297.640±1372.	17612.889±2423.	0.069±0.003	0.431±0.161	0.509±0.113
AFT	21.571±6.374	44914.194±1154.	184075.172±4347.	0.040±0.003	0.622±0.009	0.622±0.008	41.299±11.27	2648.410±301.3	20207.756±998.6	0.077±0.011	0.395±0.082	0.611±0.019
CRAFT	24.554±4.216	42953.544±389.9	177039.500±329.3	0.041±0.003	0.600±0.003	0.609±0.002	10.108±0.186	1806.321±117.4	14411.712±305.9	0.048±0.000	0.233±0.021	0.425±0.024
FAB	57.111±24.53	9567.319±626.1	58832.370±1092.	0.047±0.007	0.073±0.005	0.099±0.001	8.868±1.673	**1193.455±152.3**	**7490.803±433.9**	**0.035±0.003**	**0.093±0.014**	**0.102±0.012**
MCD	211.657±3.504	4892.591±71.26	**30977.775±276.6**	0.136±0.001	0.043±0.000	**0.054±0.000**	102.002±0.338	6406.902±20.87	32034.058±40.86	0.215±0.001	0.256±0.000	0.257±0.000
LDVI	178.241±3.129	4931.898±87.43	31019.831±278.6	0.118±0.003	0.043±0.000	**0.054±0.000**	38.758±4.940	2899.472±102.9	17435.914±299.8	0.084±0.008	0.181±0.003	0.183±0.002
PIS	**10.398±1.599**	10405.749±69.41	92623.455±1219.	**0.031±0.001**	0.082±0.000	0.168±0.003	**2.476±0.236**	2078.751±41.51	32415.244±63.11	**0.033±0.001**	0.205±0.008	0.258±0.001
DIS	65.162±35.72	**3044.733±464.7**	31573.015±702.4	0.071±0.017	**0.034±0.003**	**0.055±0.001**	3.486±0.214	2200.590±18.73	13059.766±72.12	0.037±0.002	0.155±0.001	0.152±0.001
DDS	16.217±3.202	5435.177±172.2	38576.259±392.9	0.035±0.002	0.045±0.001	0.065±0.001	3.641±0.224	2145.188±3.960	24187.186±256.4	**0.034±0.003**	0.124±0.001	0.219±0.003
GBS	140.138±39.76	5027.819±103.7	31970.248±1177.	0.108±0.021	0.043±0.000	**0.055±0.001**	2.572±0.099	5708.871±20.91	22914.911±300.4	**0.034±0.001**	0.232±0.000	0.203±0.001
	$\boldsymbol{\Delta} \log \mathbf{Z}_r$ ↓			$\boldsymbol{\Delta} \log \mathbf{Z}_f$ ↓			$\boldsymbol{\Delta} \log \mathbf{Z}_r$ ↓			$\boldsymbol{\Delta} \log \mathbf{Z}_f$ ↓		
MFVI	0.084±0.066	3.658±0.040	3.676±0.0130	0.150±0.002	0.185±0.002	0.176±0.0050	0.018±0.003	3.009±0.291	8.048±0.758	0.114±0.000	**0.048±0.002**	5.982±0.019
GMMVI	0.044±0.011	**1.715±0.119**	**1.709±0.0580**	**0.003±0.002**	**0.048±0.007**	**0.028±0.0270**	**0.000±0.000**	1.282±0.221	**7.126±0.377**	**0.000±0.000**	0.084±0.055	**5.708±0.478**
SMC	0.069±0.010	690.721±11.21	6326.621±51.428	2.728±0.000	161.796±0.000	661.945±0.0000	0.016±0.009	3.880±1.105	49.846±7.638	1.262±0.000	80.992±0.000	338.745±0.000
AFT	0.023±0.015	765.624±108.0	5567.272±277.52	1.157±0.038	110.955±18.37	420.932±12.987	0.024±0.014	4.081±1.579	47.121±6.693	0.639±0.095	205.297±23.91	12765.117±2877.
CRAFT	0.008±0.001	337.094±9.296	2504.363±64.970	0.901±0.007	100.987±0.065	415.277±0.4550	0.004±0.001	**0.822±0.087**	19.738±0.342	0.333±0.013	210.245±6.098	12516.502±631.8
FAB	**0.007±0.003**	2.952±0.247	3.331±0.2290	1.193±0.125	126.363±1.789	545.226±5.6200	0.005±0.001	3.358±1.062	43.419±4.690	0.268±0.093	84.592±22.64	13514.417±101.9
MCD	0.010±0.002	31.319±1.793	2354.020±60.855	0.009±0.005	21.148±1.478	305.656±2.6620	0.010±0.002	28.607±1.275	210.536±1.393	0.068±0.010	24.757±0.841	147.321±1.272
LDVI	0.038±0.015	8.159±0.775	647.953±7.2120	0.031±0.008	15.477±0.815	282.699±4.1050	0.004±0.001	4.360±0.741	103.224±2.118	0.017±0.003	5.472±0.938	83.029±0.819
CMCD	0.026±0.011	51.218±2.809	306.127±24.673	0.030±0.008	79.227±3.758	440.341±4.2520	0.004±0.001	10.533±0.404	167.654±1.564	0.005±0.002	12.835±0.275	148.676±2.851
PIS	0.267±0.006	7.122±0.630	40.699±0.5430	0.094±0.038	3113.492±1.978	16071.743±3.0460	0.275±0.016	12.248±0.326	209.981±2.573	0.342±0.001	54.090±0.151	304.178±0.329
DIS	0.058±0.030	87.709±8.942	11646.394±15938.	1.390±0.458	369.352±16.29	14376.906±17877.	0.049±0.005	10.448±0.607	658.634±4.952	3.212±0.028	87.897±5.255	433.741±10.78
DDS	0.012±0.005	1.739±0.442	27.506±2.5840	1.698±0.029	207.545±1.163	1052.805±1.7320	0.005±0.001	7.952±0.299	155.502±3.594	0.315±0.000	53.411±0.024	291.566±0.102
GBS	**0.007±0.001**	8.103±1.696	87.971±14.656	0.008±0.001	9.321±0.776	72.634±12.301	0.002±0.000	53.767±0.732	157.791±2.947	0.010±0.002	47.441±0.098	101.874±2.214
	ELBO ↑			**EUBO** ↓			**ELBO** ↑			**EUBO** ↓		
MFVI	−3.011±0.002	−3.690±0.000	−3.695±0.0010	3.089±0.000	164.114±0.000	666.954±0.0000	−1.038±0.007	−5.957±0.007	−16.969±0.011	1.218±0.001	72.663±0.005	324.202±0.044
GMMVI	**−0.045±0.011**	**−1.715±0.119**	**−1.709±0.0580**	3.619±1.308	240.459±51.13	645.405±6.3090	**−0.001±0.000**	−3.890±0.122	**−15.649±0.173**	**0.002±0.001**	57.746±1.928	268.513±17.65
SMC	−2.095±0.009	−877.034±10.23	−6816.697±44.195	2.734±0.000	161.921±0.000	662.404±0.0000	−0.010±0.016	−4.634±1.088	−52.535±7.564	1.272±0.000	81.325±0.000	340.984±0.000
AFT	−1.778±0.090	−927.16±103.8	−6053.823±260.99	1.248±0.045	117.63±22.16	439.434±16.788	−0.041±0.031	−4.923±1.546	−50.328±6.627	0.67±0.100	207.625±24.14	12801.561±2892.
CRAFT	−0.666±0.026	−451.399±7.561	−2836.471±57.695	0.976±0.007	103.674±0.069	425.500±0.7070	**−0.002±0.002**	**−0.339±0.180**	−22.687±0.358	0.346±0.013	212.210±6.160	12553.883±645.7
FAB	−19.932±12.80	−299.916±253.4	−63.212±56.191	0.865±0.113	93.560±5.086	386.884±12.161	−0.257±0.075	75.735±175.8	−98.558±7.688	0.162±0.055	**18.088±2.503**	227.514±0.320
MCD	−0.651±0.014	−185.021±0.743	−4017.832±20.356	0.652±0.008	**43.670±0.457**	358.687±2.1120	−1.215±0.005	−69.358±0.633	−308.728±0.450	0.734±0.002	47.834±0.820	208.626±0.525
LDVI	−0.986±0.136	−29.034±0.591	−956.576±6.2700	1.072±0.242	51.137±0.177	375.527±3.1100	−0.311±0.034	−28.471±1.018	−173.716±2.629	0.198±0.008	20.887±1.042	**132.711±1.817**
PIS	−0.585±0.016	−16.881±0.026	−65.700±0.2010	7.344±0.004	3626.120±1.360	16979.347±4.4700	−0.387±0.009	−29.261±1.743	−306.678±0.548	1.868±0.000	88.192±0.005	363.435±0.030
DIS	−1.850±0.359	−181.348±15.47	−14142.693±17807.	6.653±0.357	546.335±30.86	15792.004±18966.	−0.157±0.023	−36.704±0.629	−819.959±6.264	4.778±0.038	193.270±3.293	658.575±7.820
DDS	−0.527±0.022	−13.284±0.460	−60.642±2.3330	4.176±0.000	291.867±0.047	1224.926±2.4850	−0.110±0.007	−31.681±0.363	−244.188±3.504	1.783±0.000	86.014±0.001	351.204±0.005
GBS	−0.473±0.061	−35.771±1.105	−161.259±20.704	**0.485±0.047**	67.819±2.157	**204.498±48.539**	−0.064±0.004	−99.369±0.158	−258.263±2.639	0.064±0.004	73.545±0.107	147.412±1.504
	ESS$_r$ ↑			**ESS**$_f$ ↑			**ESS**$_r$ ↑			**ESS**$_f$ ↑		
MFVI	0.077±0.016	0.997±0.000	0.988±0.001	0.286±0.000	0.000±0.000	0.000±0.000	0.180±0.007	**0.031±0.007**	**0.006±0.001**	0.163±0.000	0.000±0.000	0.000±0.000
GMMVI	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	0.000±0.000	0.000±0.000	0.000±0.000	**0.997±0.000**	0.027±0.004	**0.006±0.001**	**0.997±0.000**	0.000±0.000	0.000±0.000
MCD	0.311±0.013	0.001±0.000	0.000±0.000	0.289±0.010	0.000±0.000	0.000±0.000	0.332±0.004	0.001±0.000	0.000±0.000	0.352±0.004	0.000±0.000	0.000±0.000
LDVI	0.207±0.044	0.002±0.000	0.000±0.000	0.269±0.046	0.000±0.000	0.000±0.000	0.742±0.006	0.002±0.000	0.000±0.000	0.761±0.004	0.000±0.000	0.000±0.000
PIS	0.529±0.012	0.006±0.001	0.002±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.840±0.004	0.003±0.000	0.000±0.000	0.042±0.000	0.000±0.000	0.000±0.000
DIS	0.078±0.025	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.580±0.003	0.002±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000
DDS	0.338±0.003	0.003±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.780±0.010	0.002±0.000	0.000±0.000	0.032±0.000	0.000±0.000	0.000±0.000
GBS	0.405±0.029	0.002±0.000	0.000±0.000	**0.380±0.027**	0.000±0.000	0.000±0.000	0.879±0.002	0.001±0.000	0.000±0.000	0.721±0.161	0.000±0.000	0.000±0.000

Table 9: Results for various sampling methods for MoG and MoS with varying dimensions d. Evaluation criteria include 2-Wasserstein distance ($\mathcal{W}_2$), maximum mean discrepancy (MMD), reverse and forward partition function error ($\Delta \log Z_r$, $\Delta \log Z_f$), lower and upper evidence bounds (ELBO, EUBO), reverse and forward effective sample size (ESS$_r$, ESS$_f$). The best results are highlighted in bold. Arrows (↑, ↓) indicate whether higher or lower values are preferable, respectively.

	NFE ↓		
Method	$d = 2$	$d = 50$	$d = 200$
MFVI	6.5×10^6	2.3×10^6	1.9×10^6
GMMVI	1.4×10^5	5.9×10^5	7.9×10^5
SMC	2.8×10^6	2.8×10^6	2.8×10^6
AFT	2.0×10^5	2.0×10^5	2.0×10^5
CRAFT	4.5×10^9	4.4×10^9	4.5×10^9
FAB	1.5×10^7	3.4×10^7	3.4×10^7
DDS	6.0×10^8	4.8×10^8	3.6×10^8
MCD	1.3×10^9	1.3×10^9	1.2×10^9
LDVI	1.3×10^9	1.3×10^9	1.3×10^9

Table 10: Number of function evaluations (NFE), that is number of times a sampling method queries $\gamma(\mathbf{x})$ until achieving the highest ELBO value for varying dimensions d on MoG.

F. Ablation Studies

F.1. Ablation Study: Batchsize and Number of Particles

Experimental Setup. We test the influence of different batchsizes/number of particles on ELBO and EMC on the MoG experiment for various methods. We use the parameters detailed in Appendix D.

Discussion. The results for the ablation study for the batchsize can be found in Table 11. We find that increasing batchsizes do not yield significant performance increases for simple methods such as MFVI. For more complex methods such as MCD or DDS, larger batchsizes tend to yield consistently better ELBO values across varying dimensionalities of the target density. In contrast, EMC values are unaffected by larger batchsizes (cf. MCD $d = 200$).

The results for the number of particles can be found in Table 12. Surprisingly, ELBO values do often not improve beyond 512 particles, despite particle interactions through resampling (Del Moral et al., 2006). Moreover, similar to the batch size, EMC does not change significantly when using a larger number of particles.

			ELBO ↑			**EMC** ↑	
METHOD	BATCHSIZE	$d = 2$	$d = 50$	$d = 200$	$d = 2$	$d = 50$	$d = 200$
MFVI	64	$\mathbf{-3.011}_{\pm \mathbf{0.003}}$	$-3.707_{\pm 0.002}$	$-3.746_{\pm 0.001}$	$\mathbf{0.383}_{\pm \mathbf{0.002}}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
	128	$\mathbf{-3.012}_{\pm \mathbf{0.004}}$	$\mathbf{-3.698}_{\pm \mathbf{0.002}}$	$-3.731_{\pm 0.002}$	$\mathbf{0.382}_{\pm \mathbf{0.003}}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
	512	$\mathbf{-3.011}_{\pm \mathbf{0.004}}$	$\mathbf{-3.694}_{\pm \mathbf{0.0}}$	$\mathbf{-3.706}_{\pm \mathbf{0.0}}$	$\mathbf{0.382}_{\pm \mathbf{0.002}}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
	1024	$\mathbf{-3.012}_{\pm \mathbf{0.003}}$	$\mathbf{-3.692}_{\pm \mathbf{0.001}}$	$\mathbf{-3.701}_{\pm \mathbf{0.002}}$	$\mathbf{0.382}_{\pm \mathbf{0.002}}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
	2048	$\mathbf{-3.012}_{\pm \mathbf{0.003}}$	$\mathbf{-3.691}_{\pm \mathbf{0.0}}$	$\mathbf{-3.697}_{\pm \mathbf{0.001}}$	$\mathbf{0.383}_{\pm \mathbf{0.002}}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
MCD	64	$-3.017_{\pm 0.2}$	$-942.74_{\pm 8.447}$	$-4699.422_{\pm 269.44}$	$\mathbf{0.796}_{\pm \mathbf{0.003}}$	$\mathbf{0.994}_{\pm \mathbf{0.001}}$	$\mathbf{0.989}_{\pm \mathbf{0.0}}$
	128	$-2.685_{\pm 0.168}$	$-889.472_{\pm 7.41}$	$-4145.279_{\pm 179.564}$	$\mathbf{0.798}_{\pm \mathbf{0.001}}$	$\mathbf{0.994}_{\pm \mathbf{0.001}}$	$\mathbf{0.988}_{\pm \mathbf{0.0}}$
	512	$-2.409_{\pm 0.05}$	$-876.718_{\pm 6.132}$	$\mathbf{-3442.883}_{\pm \mathbf{260.824}}$	$\mathbf{0.796}_{\pm \mathbf{0.002}}$	$\mathbf{0.994}_{\pm \mathbf{0.0}}$	$\mathbf{0.988}_{\pm \mathbf{0.0}}$
	1024	$-2.277_{\pm 0.131}$	$-844.588_{\pm 10.761}$	OOM	$\mathbf{0.797}_{\pm \mathbf{0.002}}$	$\mathbf{0.994}_{\pm \mathbf{0.0}}$	OOM
	2048	$\mathbf{-2.257}_{\pm \mathbf{0.075}}$	$\mathbf{-823.443}_{\pm \mathbf{18.151}}$	OOM	$\mathbf{0.796}_{\pm \mathbf{0.002}}$	$\mathbf{0.994}_{\pm \mathbf{0.0}}$	OOM
DDS	64	$-0.807_{\pm 0.036}$	$-16.83_{\pm 0.404}$	$-67.053_{\pm 0.993}$	$0.973_{\pm 0.002}$	$\mathbf{0.992}_{\pm \mathbf{0.0}}$	$\mathbf{0.984}_{\pm \mathbf{0.001}}$
	128	$-0.716_{\pm 0.009}$	$-16.092_{\pm 0.247}$	$-65.232_{\pm 0.4}$	$0.978_{\pm 0.002}$	$\mathbf{0.991}_{\pm \mathbf{0.001}}$	$\mathbf{0.983}_{\pm \mathbf{0.001}}$
	512	$-0.611_{\pm 0.022}$	$-15.61_{\pm 0.206}$	$-63.135_{\pm 0.348}$	$\mathbf{0.984}_{\pm \mathbf{0.001}}$	$\mathbf{0.992}_{\pm \mathbf{0.001}}$	$\mathbf{0.983}_{\pm \mathbf{0.001}}$
	1024	$-0.593_{\pm 0.011}$	$-15.414_{\pm 0.161}$	$-62.086_{\pm 0.258}$	$\mathbf{0.985}_{\pm \mathbf{0.001}}$	$\mathbf{0.992}_{\pm \mathbf{0.0}}$	$\mathbf{0.982}_{\pm \mathbf{0.002}}$
	2048	$\mathbf{-0.556}_{\pm \mathbf{0.009}}$	$\mathbf{-15.313}_{\pm \mathbf{0.165}}$	$\mathbf{-61.576}_{\pm \mathbf{0.384}}$	$\mathbf{0.988}_{\pm \mathbf{0.001}}$	$\mathbf{0.992}_{\pm \mathbf{0.001}}$	$\mathbf{0.979}_{\pm \mathbf{0.001}}$

Table 11: ELBO and EMC values for varying batch sizes for different methods, and dimensions of the MoG target density. Best values are marked with bold font. Here, OOM refers to 'out of memory'.

			ELBO ↑			**EMC** ↑	
METHOD	PARTICLES	$d = 2$	$d = 50$	$d = 200$	$d = 2$	$d = 50$	$d = 200$
SMC	64	$-9.267_{\pm 0.217}$	$-2622.073_{\pm 21.637}$	$-17904.276_{\pm 25.557}$	$0.824_{\pm 0.018}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
	128	$-9.08_{\pm 0.041}$	$-2647.733_{\pm 31.935}$	$-16999.909_{\pm 14.959}$	$0.879_{\pm 0.035}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
	512	$-8.823_{\pm 0.033}$	$\mathbf{-1911.449}_{\pm \mathbf{8.527}}$	$-16867.03_{\pm 44.663}$	$0.941_{\pm 0.004}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
	1024	$\mathbf{-8.595}_{\pm \mathbf{0.035}}$	$-2323.482_{\pm 13.52}$	$-15565.314_{\pm 78.958}$	$\mathbf{0.971}_{\pm \mathbf{0.004}}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
	2048	$-10.317_{\pm 0.028}$	$-2041.686_{\pm 20.993}$	$\mathbf{-15032.371}_{\pm \mathbf{46.049}}$	$\mathbf{0.965}_{\pm \mathbf{0.002}}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
CRAFT	64	$-3.666_{\pm 0.048}$	$-793.354_{\pm 19.752}$	$-4646.891_{\pm 77.062}$	$\mathbf{0.986}_{\pm \mathbf{0.001}}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
	128	$-3.604_{\pm 0.039}$	$-790.385_{\pm 23.036}$	$-4656.227_{\pm 80.153}$	$\mathbf{0.986}_{\pm \mathbf{0.001}}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
	512	$-3.6_{\pm 0.061}$	$-784.881_{\pm 14.364}$	$\mathbf{-4624.869}_{\pm \mathbf{63.618}}$	$\mathbf{0.987}_{\pm \mathbf{0.0}}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
	1024	$\mathbf{-3.552}_{\pm \mathbf{0.041}}$	$-785.251_{\pm 16.847}$	$-4632.063_{\pm 68.715}$	$\mathbf{0.986}_{\pm \mathbf{0.002}}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
	2048	$\mathbf{-3.553}_{\pm \mathbf{0.05}}$	$\mathbf{-782.068}_{\pm \mathbf{13.855}}$	$-4625.769_{\pm 55.853}$	$\mathbf{0.987}_{\pm \mathbf{0.001}}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$

Table 12: ELBO and EMC values for varying number of particles and dimensions of the MoG target density.

F.2. Ablation Study: Number of Temperatures / Timesteps T

Experimental Setup. We test the influence of different number of temperatures/timesteps T for methods of sequential nature such as sequential importance sampling or SDE based methods. We use batch sizes of 512. The remaining parameters are set according to Appendix D.

Discussion. The results are illustrated in Figure 6. We can see that using larger values of T tend to improves ELBO and EUBO values across all methods.

F.3. Ablation Study: Sequential Monte Carlo Design Choices

Experimental Setup. As Sequential Monte Carlo is the basis for many sampling methods such as SNF (Wu et al., 2020a), AFT (Arbel et al., 2021), CRAFT (Matthews et al., 2022), or FAB (Midgley et al., 2022) we perform a thorough ablation of its design choices. In particular, we ablate the influence of the MCMC kernel and whether or not resampling is used. We tested Metropolis-Hastings (MH) and Hamiltonian Monte Carlo (HMC) MCMC kernels where we used the same number of function evaluations and hand-tuned the stepsizes such obtained a rejection rate ≈ 0.65. The results are shown in Table 13.

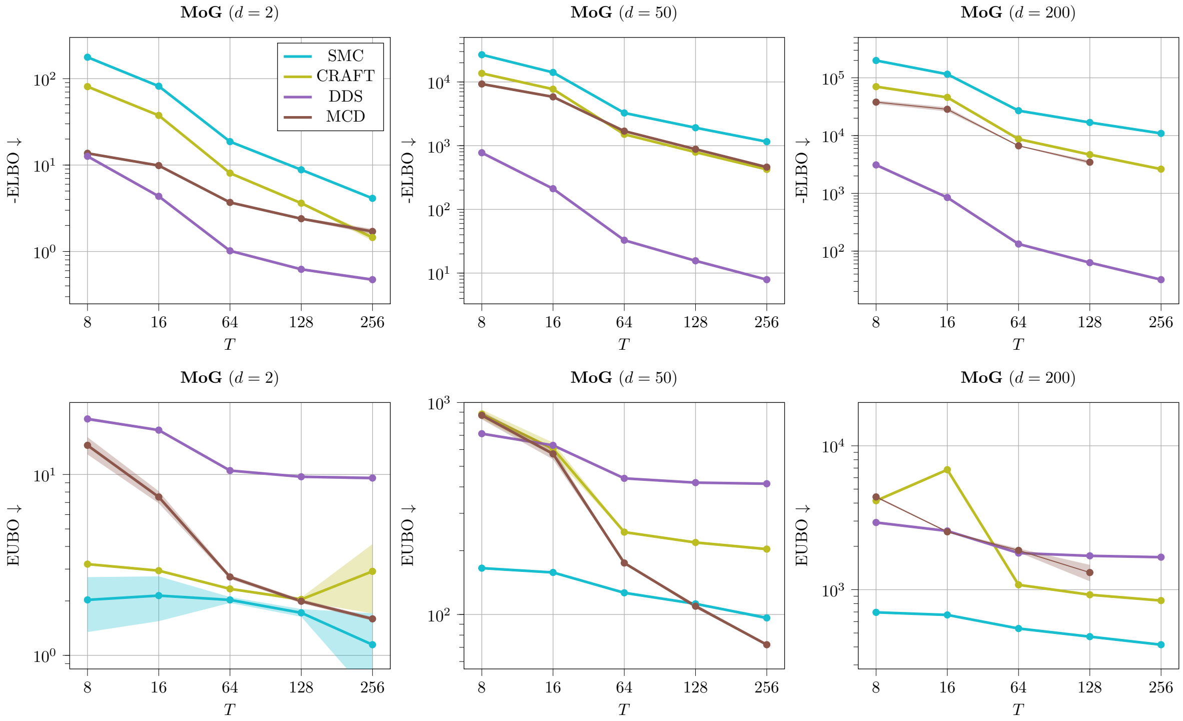


Figure 6: Negative ELBO and EUBO values for varying temperatures/timesteps T for different dimensions of the MoG target density. Best values are marked with bold font. Missing values for $T = 256$ are caused by out-of-memory problems.

Discussion. HMC outperforms MH across all dimensions with respect to both, ELBO and EMC values. Surprisingly, not using resampling avoids mode collapse entirely as indicated by EMC ≈ 1.

MCMC Kernel	Re-sampling	ELBO ↑			EMC ↑		
		$d = 2$	$d = 50$	$d = 200$	$d = 2$	$d = 50$	$d = 200$
-	✗	$-9.473_{\pm 0.000}$	$-32034.303_{\pm 0.000}$	$-292642.344_{\pm 0.000}$	$0.785_{\pm 0.000}$	$\mathbf{0.987}_{\pm \mathbf{0.000}}$	$\mathbf{0.988}_{\pm \mathbf{0.000}}$
-	✓	$-9.28_{\pm 0.2044}$	$-27534.303_{\pm 72.32}$	$-288123.325_{\pm 108.010}$	$0.618_{\pm 0.191}$	$0_{\pm 0}$	$0_{\pm 0}$
MH	✗	$-9.166_{\pm 0.138}$	$-26686.496_{\pm 412.669}$	$-275404.367_{\pm 1375.306}$	$0.785_{\pm 0.003}$	$\mathbf{0.987}_{\pm \mathbf{0.000}}$	$\mathbf{0.988}_{\pm \mathbf{0.000}}$
MH	✓	$-9.064_{\pm 0.034}$	$-22411.798_{\pm 69.874}$	$-251904.734_{\pm 422.895}$	$0.864_{\pm 0.021}$	$0_{\pm 0}$	$0_{\pm 0}$
HMC	✗	$-\mathbf{8.736}_{\pm \mathbf{0.031}}$	$-2272.619_{\pm 96.639}$	$-18270.795_{\pm 91.703}$	$0.798_{\pm 0.006}$	$\mathbf{0.986}_{\pm \mathbf{0.000}}$	$\mathbf{0.988}_{\pm \mathbf{0.000}}$
HMC	✓	$-8.850_{\pm 0.110}$	$-\mathbf{1931.168}_{\pm \mathbf{18.844}}$	$-\mathbf{16952.94}_{\pm \mathbf{49.119}}$	$\mathbf{0.940}_{\pm \mathbf{0.006}}$	$0_{\pm 0}$	$0_{\pm 0}$

Table 13: Ablation study for Sequential Monte Carlo (Del Moral et al., 2006). ELBO and EUBO values for different MCMC kernels and whether or not resampling is used. Here, MH refers to Metropolis-Hastings and HMC to Hamiltonian Monte Carlo (Bishop, 2006). Results are reported for different dimensions of the MoG target density.

F.4. Ablation Study: Initial Model Support

Experimental Setup. We test the influence of the initial model support for different methods of sequential nature. In particular, we vary the scale σ_0^2 of the initial proposal/base distribution $\pi_0(\mathbf{x}) = \mathcal{N}(0, \sigma_0^2 \mathbf{I})$. To that end, we report ELBO and EUBO values on the MoG experiment for varying dimensions. We use the parameters detailed in Appendix D. The results are shown in Table 14.

Discussion. The results of the ablation study investigating varied initial standard deviations for parameterizing the base distribution can be found in Table 14. We observe that, in terms of the ELBO, most methods exhibit poor performance with a higher initial scale, particularly in higher dimensions. Conversely, EMC values tend to get 0 for small initial scales and 1 for large initial scales.

METHOD	INITIAL SCALE	ELBO ↑			EMC ↑		
		$d = 2$	$d = 50$	$d = 200$	$d = 2$	$d = 50$	$d = 200$
SMC	1	$-3.717_{\pm 0.056}$	$-1800.882_{\pm 38.348}$	$-12181.939_{\pm 110.871}$	$0.002_{\pm 0.002}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
	10	$\mathbf{-0.175}_{\pm \mathbf{0.408}}$	$\mathbf{-313.926}_{\pm \mathbf{4.847}}$	$\mathbf{-2008.238}_{\pm \mathbf{10.954}}$	$0.722_{\pm 0.022}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
	30	$-0.666_{\pm 0.081}$	$-674.1_{\pm 12.747}$	$-5014.372_{\pm 27.284}$	$\mathbf{0.957}_{\pm \mathbf{0.007}}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
	60	$-8.823_{\pm 0.033}$	$-1911.449_{\pm 8.527}$	$-16867.03_{\pm 44.663}$	$0.941_{\pm 0.004}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
CRAFT	1	$-2.675_{\pm 0.236}$	$\mathbf{-11.333}_{\pm \mathbf{0.644}}$	$\mathbf{-83.301}_{\pm \mathbf{1.267}}$	$0.143_{\pm 0.033}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
	10	$-0.633_{\pm 0.538}$	$-136.414_{\pm 2.482}$	$-1090.374_{\pm 22.117}$	$0.657_{\pm 0.233}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
	30	$\mathbf{-0.229}_{\pm \mathbf{0.018}}$	$-350.247_{\pm 11.605}$	$-2482.919_{\pm 12.176}$	$0.974_{\pm 0.004}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
	60	$-3.563_{\pm 0.057}$	$-784.881_{\pm 14.364}$	$-4624.869_{\pm 63.618}$	$\mathbf{0.987}_{\pm \mathbf{0.001}}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
MCD	1	$-3.676_{\pm 0.001}$	$\mathbf{-3.292}_{\pm \mathbf{0.011}}$	$\mathbf{-4.281}_{\pm \mathbf{0.039}}$	$0.005_{\pm 0.0}$	$0.187_{\pm 0.0}$	$0.005_{\pm 0.001}$
	10	$-1.653_{\pm 0.032}$	$-87.5_{\pm 0.519}$	$-144.237_{\pm 4.133}$	$0.613_{\pm 0.003}$	$0.658_{\pm 0.004}$	$0.647_{\pm 0.002}$
	30	$\mathbf{-1.138}_{\pm \mathbf{0.064}}$	$-441.73_{\pm 2.245}$	$-1265.551_{\pm 6.991}$	$\mathbf{0.94}_{\pm \mathbf{0.001}}$	$0.961_{\pm 0.0}$	$0.942_{\pm 0.002}$
	60	$-2.384_{\pm 0.059}$	$-878.12_{\pm 8.598}$	$-3458.28_{\pm 248.958}$	$0.798_{\pm 0.003}$	$\mathbf{0.994}_{\pm \mathbf{0.001}}$	$\mathbf{0.988}_{\pm \mathbf{0.0}}$
DDS	1	$-3.622_{\pm 0.012}$	$\mathbf{-6.053}_{\pm \mathbf{0.624}}$	$-49.0_{\pm 10.277}$	$0.0_{\pm 0.0}$	$0.187_{\pm 0.0}$	$0.0_{\pm 0.0}$
	10	$-0.737_{\pm 0.024}$	$-6.954_{\pm 0.146}$	$\mathbf{-20.149}_{\pm \mathbf{0.075}}$	$0.85_{\pm 0.001}$	$0.26_{\pm 0.031}$	$0.348_{\pm 0.011}$
	30	$\mathbf{-0.408}_{\pm \mathbf{0.01}}$	$-10.604_{\pm 0.165}$	$-42.396_{\pm 0.105}$	$\mathbf{0.989}_{\pm \mathbf{0.001}}$	$0.941_{\pm 0.003}$	$0.841_{\pm 0.011}$
	60	$-0.612_{\pm 0.019}$	$-15.598_{\pm 0.106}$	$-63.101_{\pm 0.253}$	$\mathbf{0.984}_{\pm \mathbf{0.001}}$	$\mathbf{0.992}_{\pm \mathbf{0.001}}$	$\mathbf{0.983}_{\pm \mathbf{0.002}}$

Table 14: ELBO and EMC values for varying initial scales, and dimensions of the MoG target density.

F.5. Ablation Study: Langevin Methods

Experimental Setup. The augemented ELBO allows for end-to-end training of several parameters that otherwise need careful tuning. (Geffner & Domke, 2022) showed that learning the mean and variance of the proposal distribution π_0, the time discretization stepsize Δ_t and annealing schedule $(\beta_t)_{t=1}^T$ by maximizing the extended ELBO. Here, we test the influence of training vs. fixing these paramters for MCD (Doucet et al., 2022b) on the MoG target for varying dimensions. The results are shown in Table 15. The fixed parameters are chosen according to Table D.

Discussion. We observe that learning more parameters tend to yield higher ELBO values. However, especially learning the parameters of the proposal π_0 results in low EMC values.

TRAINABLE			ELBO ↑			EMC ↑		
σ_t	β_t	π_0	$d = 2$	$d = 50$	$d = 200$	$d = 2$	$d = 50$	$d = 200$
✗	✗	✗	$-3.519_{\pm 0.154}$	$-2513.292_{\pm 26.017}$	$-13575.6_{\pm 414.217}$	$0.799_{\pm 0.004}$	$\mathbf{0.994}_{\pm \mathbf{0.0}}$	$\mathbf{0.988}_{\pm \mathbf{0.001}}$
✓	✗	✗	$-2.441_{\pm 0.079}$	$-1141.639_{\pm 18.651}$	$-6574.401_{\pm 114.962}$	$0.819_{\pm 0.004}$	$\mathbf{0.994}_{\pm \mathbf{0.0}}$	$\mathbf{0.988}_{\pm \mathbf{0.001}}$
✗	✓	✗	$-2.384_{\pm 0.059}$	$-878.12_{\pm 8.598}$	$-3458.28_{\pm 248.958}$	$0.798_{\pm 0.003}$	$\mathbf{0.994}_{\pm \mathbf{0.001}}$	$\mathbf{0.988}_{\pm \mathbf{0.0}}$
✓	✓	✗	$-1.51_{\pm 0.035}$	$-173.002_{\pm 1.548}$	$-825.303_{\pm 44.797}$	$0.828_{\pm 0.003}$	$\mathbf{0.993}_{\pm \mathbf{0.0}}$	$\mathbf{0.989}_{\pm \mathbf{0.001}}$
✗	✗	✓	$-1.621_{\pm 0.216}$	$-38.022_{\pm 41.035}$	$-43.416_{\pm 7.242}$	$0.927_{\pm 0.015}$	$0.276_{\pm 0.132}$	$0.236_{\pm 0.053}$
✓	✗	✓	$-1.235_{\pm 0.072}$	$-29.238_{\pm 20.912}$	$\mathbf{-33.686}_{\pm \mathbf{4.508}}$	$\mathbf{0.95}_{\pm \mathbf{0.005}}$	$0.309_{\pm 0.109}$	$0.393_{\pm 0.014}$
✗	✓	✓	$-1.137_{\pm 0.118}$	$\mathbf{-8.323}_{\pm \mathbf{1.718}}$	$-103.968_{\pm 68.449}$	$0.936_{\pm 0.011}$	$0.19_{\pm 0.004}$	$0.341_{\pm 0.122}$
✓	✓	✓	$\mathbf{-1.05}_{\pm \mathbf{0.099}}$	$-10.526_{\pm 2.256}$	$-36.254_{\pm 15.018}$	$0.913_{\pm 0.017}$	$0.381_{\pm 0.075}$	$0.435_{\pm 0.062}$

Table 15: ELBO and EMC values of MCD for learning the mean and variance of the proposal distribution π_0, the diffusion coefficient σ_t and annealing schedule $(\beta_t)_{t=1}^T$ by maximizing the extended ELBO for varying dimensions d on the MoG target.

F.6. Ablation Study: Transport Flow Type

Experimental Setup. We test different flow types as transport maps for CRAFT using a different number of temperatures T. In particular, we consider diagonal affine flows, inverse autoregressive flows (Kingma et al., 2016) and neural spline flows (Durkan et al., 2019) where we set the spline bounds to match the support of the MoG target. The results are visualized in Figure 7.

Discussion. We found that diagonal affine paired with larger number of temperatures results in a better, more robust performance compared to using more sophisticated flow types. Moreover, the latter often result in out-of-memory problems on high dimensional problems.

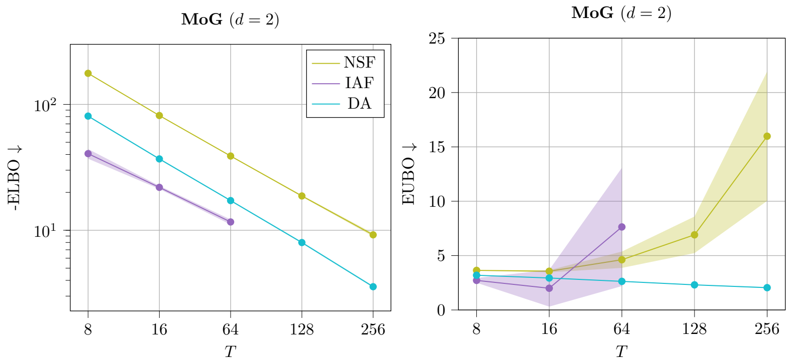


Figure 7: ELBO and EUBO values for CRAFT for different flow types and number of temperatures T on the two-dimensional MoG target. In particular, diagonal affine flows, inverse autoregressive flows (IAF) (Kingma et al., 2016) and neural spline flows (NSF) (Durkan et al., 2019). For larger T, IAF becomes numerically unstable.

F.7. Ablation Study: Gradient Guidance

Experimental Setup. (Zhang & Chen, 2021) proposed to use a network of the form $f^\theta(\mathbf{x}, t) = f_1^\theta(\mathbf{x}, t) + f_2^\theta(t)\nabla \log \gamma(\mathbf{x})$ and initialize such that $f_1^\theta(\mathbf{x}, t) = 0$. They showed that this gradient guidance helps with mode collapse and yields overall better results. (Vargas et al., 2023a; Berner et al., 2022; Richter et al., 2023) adopted the approach and reported similar results. Here, we test the network architecture with and without gradient guidance $f_2^\theta(t)\nabla \log \gamma(\mathbf{x})$ on the MoG target for a varying number of dimensions for the diffusion sampler.

Discussion. The results of this examination can be found in Table 16 and indicate that both the ELBO and EMC significantly deteriorate without gradient guidance, and this degradation increases with higher dimensions. This aligns with the findings from (Zhang & Chen, 2021; Vargas et al., 2023a; Berner et al., 2022; Richter et al., 2023).

Gradient Guidance	ELBO ↑			EMC ↑		
	$d = 2$	$d = 50$	$d = 200$	$d = 2$	$d = 50$	$d = 200$
✗	$-3.105_{\pm 0.27}$	$-543.099_{\pm 13.612}$	$-247920.463_{\pm 4258.605}$	$0.453_{\pm 0.011}$	$0.0_{\pm 0.0}$	$0.243_{\pm 0.421}$
✓	$\mathbf{-0.612}_{\pm 0.019}$	$\mathbf{-15.598}_{\pm 0.106}$	$\mathbf{-63.101}_{\pm 0.253}$	$\mathbf{0.984}_{\pm 0.001}$	$\mathbf{0.992}_{\pm 0.001}$	$\mathbf{0.983}_{\pm 0.002}$

Table 16: ELBO and EMC values with and without gradient guidance $f_2^\theta(t)\nabla \log \gamma(\mathbf{x})$ as part of the network architecture for the denoising diffusion sampler (DDS) on the MoG target for varying dimension d.

F.8. Ablation Study: Pre-training the Proposal/Base-Distribution π_0

Experimental Setup. We test the impact of pre-training the mean and covariance matrix of the Gaussian proposal/base distribution π_0 using MFVI on the MoG target for varying dimensions. The results are shown in Table 17.

Discussion. Pretraining the the mean and covariance matrix of the Gaussian proposal/base distribution π_0 yields significantly higher ELBO values at the cost of EMC values close to 0.

Method	Pretrained π_0	ELBO ↑			EMC ↑		
		$d = 2$	$d = 50$	$d = 200$	$d = 2$	$d = 50$	$d = 200$
CRAFT	✗	$\mathbf{-3.563}_{\pm 0.057}$	$-784.881_{\pm 14.364}$	$-4624.869_{\pm 63.618}$	$\mathbf{0.987}_{\pm 0.001}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
	✓	$-3.676_{\pm 0.007}$	$\mathbf{-3.501}_{\pm 0.087}$	$\mathbf{-3.699}_{\pm 0.135}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
MCD	✗	$\mathbf{-2.384}_{\pm 0.059}$	$-878.12_{\pm 8.598}$	$-3458.28_{\pm 248.958}$	$\mathbf{0.798}_{\pm 0.003}$	$\mathbf{0.994}_{\pm 0.001}$	$\mathbf{0.988}_{\pm 0.0}$
	✓	$-3.689_{\pm 0.0}$	$\mathbf{-3.746}_{\pm 0.003}$	$\mathbf{-3.938}_{\pm 0.003}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$

Table 17: ELBO and EMC values for pre-trained/fixed Gaussian proposal/base distribution π_0 on the MoG target with varying dimensions d.