Score-Based Diffusion meets Annealed Importance Sampling

Anonymous Author(s)
Affiliation
Address
email

Abstract

More than twenty years after its introduction, Annealed Importance Sampling (AIS) remains one of the most effective methods for marginal likelihood estimation. It relies on a sequence of distributions interpolating between a tractable initial distribution and the target distribution of interest which we simulate from approximately using a non-homogeneous Markov chain. To obtain an importance sampling estimate of the marginal likelihood, AIS introduces an extended target distribution to reweight the Markov chain proposal. While much effort has been devoted to improving the proposal distribution used by AIS, an underappreciated issue is that AIS uses a convenient but suboptimal extended target distribution. This can hinder its performance. We here leverage recent progress in score-based generative modeling (SGM) to approximate the optimal extended target distribution for AIS proposals corresponding to the discretization of Langevin and Hamiltonian dynamics. We demonstrate these novel, differentiable, AIS procedures on a number of synthetic benchmark distributions and variational auto-encoders.

1 Introduction

Evaluating the marginal likelihood, also known as evidence, is of key interest in Bayesian statistics as it allows not only model comparison but is also often used to select hyperparameters. A large variety of Monte Carlo methods have been proposed to address this problem, including path sampling [Gelman and Meng, 1998], AIS [Neal, 2001] and related Sequential Monte Carlo methods [Del Moral et al., 2006]. An appealing feature of AIS is that it provides an unbiased estimate of the marginal likelihood and can thus be used to define an evidence lower bound (ELBO) or mutual information bounds; see e.g. [Wu et al., 2020; Thin et al., 2021; Brekelmans et al., 2022].

AIS builds a proposal distribution using a Markov chain $(x_k)_{k=0}^K$ initialized at an easy-to-sample distribution followed by a sequence of Markov chain Monte Carlo (MCMC) transitions targeting typically annealed versions of the posterior. By proceeding this way, we obtain a proposal x_K whose distribution is expected to be a reasonable approximation to the target posterior. However, this distribution is intractable as it requires integrating the joint proposal distribution over previous states $(x_k)_{k=0}^{K-1}$. AIS bypasses this issue by instead using Importance Sampling (IS) on the whole path $(x_k)_{k=0}^K$ through the introduction of an artificial extended target distribution whose marginal at time K coincides with the posterior.

There has been much work devoted to improving AIS in machine learning and statistics but also in physics where it was introduced independently in [Jarzynski, 1997; Crooks, 1998]. A standard approach to improve AIS is to modify the intermediate distributions [Schmiedl and Seifert, 2007; Grosse et al., 2013; Masrani et al., 2021] and corresponding transition kernels of the proposal [Dai et al., 2020; Wu et al., 2020; Geffner and Domke, 2021; Thin et al., 2021; Zhang et al., 2021]. We here address a distinct issue: it was shown in [Del Moral et al., 2006] that the extended target

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

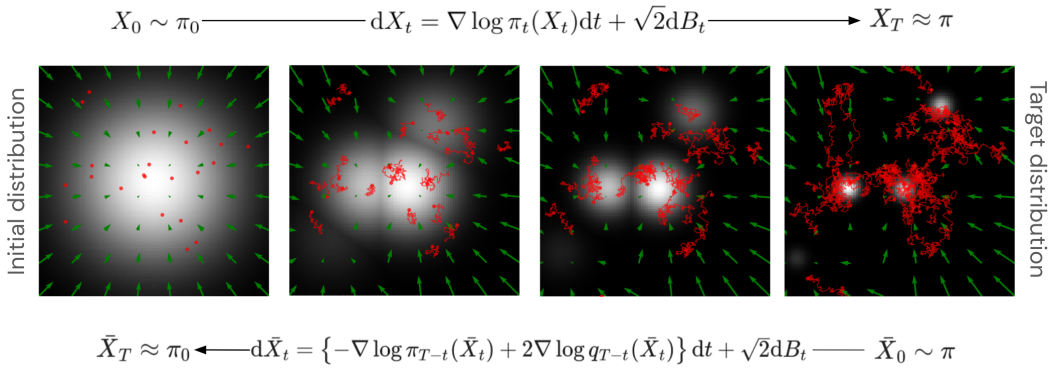

Figure 1: **Top**: Samples X_t from an AIS proposal (red) obtained by sampling initially from a Gaussian at $t = 0$ and diffusing through Langevin dynamics on intermediate targets π_t (white). The intermediate marginals of the proposal, q_t, approximated by the samples are such that $q_T \approx \pi$ for a reasonably fast mixing diffusion. **Bottom**: Computing importance weights. The optimal extended target used to compute the weights is the distribution obtained by exactly initializing $\bar{X}_T$ exactly from π and then following the reverse-time dynamics of the forward AIS proposal. This requires access to score vectors of the marginals q_t.

distribution used by AIS is suboptimal and that the optimal extended target minimizing the variance of the evidence estimate is defined through the time-reversal of the proposal. However, this result is difficult to exploit algorithmically as the time-reversal is intractable for useful proposals.

In this paper, we show how one can combine this result with recent advances in SGM to obtain improved, lower variance, AIS estimates. We concentrate on scenarios where we use unadjusted overdamped Langevin [Heng et al., 2020; Wu et al., 2020; Thin et al., 2021] and unadjusted Hamiltonian proposals with partial momentum refreshment (i.e. underdamped Langevin) [Dai et al., 2020; Wu et al., 2020; Geffner and Domke, 2021; Zhang et al., 2021] which correspond to time-discretized diffusion processes. The first benefit of using such proposals is that, by omitting Metropolis–Hastings steps, one obtains differentiable versions of the ELBO amenable to the reparameterization trick. The second benefit of these proposals is that their time-reversal can be approximated by adapting techniques developed for SGM [Ho et al., 2020; Song et al., 2021b; Dockhorn et al., 2022] to our setup. We derive a principled parameterization for an approximation of their time-reversal which we learn by maximizing the ELBO. As for SGM, this ELBO coincides with a denoising score matching loss [Hyvärinen, 2005; Vincent, 2011; Ho et al., 2020; Song et al., 2021b]. This provides novel, optimized and differentiable, AIS estimators which we refer to as Monte Carlo Diffusion (MCD).

We demonstrate the benefits of this approach on synthetic benchmark distributions and variational auto-encoders (VAEs) [Kingma and Welling, 2013]. All proofs can be found in the Appendix.

2 Annealed Importance Sampling

2.1 Setup and algorithm.

Consider a probability density π on $\mathbb{R}^d$ of the form

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

where $\gamma(x)$ can be evaluated pointwise and we want to approximate the intractable normalizing constant Z. In a Bayesian framework, $\gamma(x) = p(x)p(\mathcal{D}|x)$ is the joint density of parameter x and data $\mathcal{D}$, $\pi(x) = p(x|\mathcal{D})$ the corresponding posterior and $Z = p(\mathcal{D})$ the evidence.

To estimate Z, AIS introduces the intermediate distributions $(\pi_k)_{k=1}^K$ bridging smoothly from a tractable distribution π_0 to the target distribution $\pi_K = \pi$ of interest. One typically uses $\pi_k(x) \propto \gamma_k(x)$ with $\gamma_k(x) = \pi_0(x)^{1-\beta_k}\gamma_K(x)^{\beta_k}$ for $0 = \beta_0 < \beta_1 < \cdots < \beta_K = 1$ but other choices are possible [Grosse et al., 2013]. The IS proposal used by AIS is then obtained by running a Markov

chain $(x_k)_{k=0}^K$ such that $x_0 \sim \pi_0(\cdot)$, and then $x_k \sim F_k(\cdot|x_{k-1})$ for $k \geq 1$ where F_k is a MCMC kernel invariant w.r.t. π_k. The proposal is thus given by

$$Q(x_{0:K}) = \pi_0(x_0) \prod_{k=1}^K F_k(x_k|x_{k-1}). \tag{2}$$

Denote by q_k the intractable marginal distribution of x_k under Q satisfying $q_k(x_k) = \int q_{k-1}(x_{k-1}) F_k(x_k|x_{k-1}) \mathrm{d}x_{k-1}$ for $k \geq 1$ and $q_0 = \pi_0$. As q_K cannot be evaluated in complex scenarios, the marginal IS estimate $w_{\text{mar}}(x_K) = \gamma_K(x_K)/q_K(x_K)$ of Z is intractable.

One can bypass this issue by introducing an extended target

$$P(x_{0:K}) = \frac{\Gamma(x_{0:k})}{Z}, \qquad \Gamma(x_{0:K}) = \gamma_K(x_K) \prod_{k=0}^{K-1} B_k(x_k|x_{k+1}), \tag{3}$$

where $(B_k)_{k=0}^{K-1}$ are backward Markov transition kernels. For any selection of backward kernels such that the ratio Γ/Q is well-defined, we have

$$\mathbb{E}_Q[w(x_{0:K})] = Z, \quad \text{for } w(x_{0:K}) = \frac{\Gamma(x_{0:K})}{Q(x_{0:K})}, \tag{4}$$

i.e. $w(x_{0:K})$ is an unbiased estimate of Z for $x_{0:K} \sim Q$.

The AIS estimate of the evidence is a specific instance of the estimator (4) relying on the backward kernels $B_k^{\text{ais}}(x'|x) = \pi_{k+1}(x') F_{k+1}(x|x')/\pi_{k+1}(x)$. This yields the following expression for $\log w(x_{0:K})$:

$$\log w_{\text{ais}}(x_{0:K}) = \sum_{k=1}^K \log\big(\gamma_k(x_{k-1})/\gamma_{k-1}(x_{k-1})\big). \tag{5}$$

2.2 Limitations of AIS.

While designing P in (3) by using the backward Markov kernels $(B_k^{\text{ais}})_{k=0}^{K-1}$ is convenient, it is also suboptimal. For example, consider the ideal scenario where $F_k(x_k|x_{k-1}) = \pi_k(x_k)$. This scenario has been used many times in the literature to provide some guidelines on AIS, see e.g. Neal [2001]; Grosse et al. [2013]. In this case, $\text{var}_Q[\log w_{\text{ais}}(x_{0:K})] = \sum_{k=1}^K \text{var}_{\pi_{k-1}}[\log(\gamma_k(x_{k-1})/\gamma_{k-1}(x_{k-1}))] > 0$ while $\text{var}_{q_K}[w_{\text{mar}}(x_K)] = \text{var}_{\pi_K}[w_{\text{mar}}(x_K)] = 0$.

Another illustration of the suboptimality of AIS is to consider a scenario where the proposal is a homogeneous MCMC chain, i.e. $x_0 \sim \pi_0$ and $x_k \sim F(\cdot|x_{k-1})$ for F a π-invariant MCMC kernel; i.e. use $F_k = F$ and $\pi_k = \pi$ for $k = 1, ..., K$. If F is reasonably well-mixing, then $q_K \approx \pi$ for K large enough and the evidence estimate $w_{\text{mar}}(x_K) = \gamma_K(x_K)/q_K(x_K)$ should have small variance. However, it is easy to check that we have $w_{\text{ais}}(x_{0:K}) = \gamma(x_0)/\pi_0(x_0)$ for the exact same proposal; i.e. the AIS estimate does not depend on the MCMC samples $x_{1:K}$ and boils down to the IS estimate of Z using the proposal π_0.

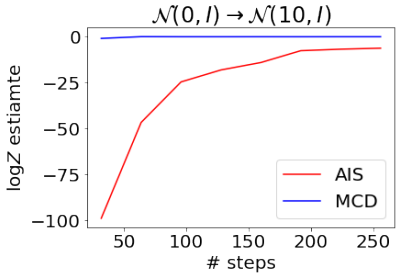

Figure 2: Comparing $\log Z$ estimates as a function of K using AIS and MCD. Both estimates use the same forward kernels but reweight samples in a different way using distinct backward kernels. Initial distribution π_0 is 20-dimensional $\mathcal{N}(0, I)$ and progressively shifts to the target $\pi = \pi_K = \mathcal{N}(10, I)$. The MCD estimate is much closer to the ground truth ($\log Z = 0$) than AIS.

These two examples illustrate that it would be preferable to use $w_{\text{mar}}(x_K)$ rather than $w_{\text{ais}}(x_{0:K})$. In Appendix A, we provide a detailed comparison of both estimates in a scenario where their variance can be computed analytically. We propose in the next section an unbiased estimate of the evidence (MCD) approximating $w_{\text{mar}}(x_K)$ based on a different choice of backward kernels. As illustrated in Figure 2, significant gains can be achieved.

3 Optimized Annealed Importance Sampling

We show here that the optimal extended target distribution P is defined through the time-reversal of the proposal Q. By exploiting a connection to SGM, we can approximate this reversal using score matching when the proposal is obtained through an unadjusted overdamped/underdamped Langevin algorithm.

3.1 Optimal Extended Target Distribution via Time Reversal

We summarize here Proposition 1 of Del Moral et al. [2006]; see also [Sohl-Dickstein et al., 2015; Bernton et al., 2019].

Proposition 1. *For a proposal Q of the form (2), the extended target P of the form (3) minimizing both the Kullback–Leibler divergence $D_{\mathrm{KL}}(Q||P)$ and the variance of the evidence estimate $w(x_{0:K}) = \Gamma(x_{0:K})/Q(x_{0:K})$ for $x_{0:K} \sim Q$ is given by $P_{\mathrm{opt}}(x_{0:K}) = \Gamma_{\mathrm{opt}}(x_{0:K})/Z$ where*

$$\Gamma_{\mathrm{opt}}(x_{0:K}) = \gamma_K(x_K) \prod_{k=0}^{K-1} B_k^{\mathrm{opt}}(x_k|x_{k+1}), \qquad B_k^{\mathrm{opt}}(x_k|x_{k+1}) = \frac{q_k(x_k)F_{k+1}(x_{k+1}|x_k)}{q_{k+1}(x_{k+1})}. \tag{6}$$

In particular, one has

$$w_{\mathrm{mar}}(x_K) = \frac{\gamma(x_K)}{q_K(x_K)} = \frac{\Gamma_{\mathrm{opt}}(x_{0:K})}{Q(x_{0:K})}, \quad \textit{and} \quad D_{\mathrm{KL}}(Q||P_{\mathrm{opt}}) = D_{\mathrm{KL}}(q_K||\pi_K). \tag{7}$$

This result follows simply from the chain rule and the law of total variance which yield

$$D_{\mathrm{KL}}(Q||P) = D_{\mathrm{KL}}(q_K||\pi_K) + \mathbb{E}_{q_K}\Big[D_{\mathrm{KL}}(Q(\cdot|x_K)||P(\cdot|x_K))\Big], \tag{8}$$

$$\mathrm{var}_Q[w(x_{0:K})] = \mathrm{var}_{q_K}[w_{\mathrm{mar}}(x_K)] + \mathbb{E}_{q_K}[\mathrm{var}_{Q(\cdot|x_K)}[w(x_{0:K})]]. \tag{9}$$

Both quantities are clearly minimized by selecting $P(x_{0:K-1}|x_K) = Q(x_{0:K-1}|x_K)$.

We emphasize that Proposition 1 applies to any forward kernels $(F_k)_{k=1}^K$ including MCMC kernels, unadjusted Langevin kernels or even deterministic maps[1]. It shows that P^{opt} is the distribution of a backward process initialized at π_K which then follows the time-reversed dynamics of the forward process Q. If we had $\pi_K = \pi$, then we would have $P^{\mathrm{opt}} = Q$ as then P^{opt} corresponds to the backward decomposition of Q.

3.2 Time reversal, Score matching and ELBO for unadjusted overdamped Langevin

We concentrate here on the case where $(F_k)_{k=1}^K$ correspond to a time-inhomogeneous unadjusted (overdamped) Langevin algorithm (ULA) as used in [Heng et al., 2020; Wu et al., 2020; Thin et al., 2021]; that is we consider $F_k(x_k|x_{k-1}) = \mathcal{N}(x_k; x_{k-1} + \delta\nabla \log \pi_k(x_{k-1}), 2\delta I)$ where $\delta > 0$ is a stepsize. Let $\delta := T/K$ then, as $K \to \infty$, the proposal Q converges to the path measure $\mathcal{Q}$ of the following inhomogeneous Langevin diffusion $(x_t)_{t\in[0,T]}$ defined by the stochastic differential equation (SDE)

$$\mathrm{d}x_t = \nabla \log \pi_t(x_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \qquad x_0 \sim \pi_0, \tag{10}$$

where $(B_t)_{t\in[0,T]}$ is standard multivariate Brownian motion and we are slightly abusing notation from now on as π_t for $t = t_k = k\delta$ corresponds to π_k in discrete-time. Many quantitative results measuring the discrepancy between the law of x_T and $\pi_T = \pi$ for such annealed diffusions have been obtained; see e.g. [Fournier and Tardif, 2021; Tang and Zhou, 2021]. It can be shown [Haussmann and Pardoux, 1986] that the time-reversed process $(\bar{x}_t) = (x_{T-t})_{t\in[0,T]}$ is also a diffusion given by

$$\mathrm{d}\bar{x}_t = \big\{ -\nabla \log \pi_{T-t}(\bar{x}_t) + 2\nabla \log q_{T-t}(\bar{x}_t)\big\}\mathrm{d}t + \sqrt{2}\mathrm{d}\bar{B}_t, \qquad \bar{x}_0 \sim q_T, \tag{11}$$

where $(\bar{B}_t)_{t\in[0,T]}$ is another multivariate Brownian motion. The continuous-time version of P_{opt} is the path measure $\mathcal{P}_{\mathrm{opt}}$ defined by the diffusion (11) but initialized at $\bar{x}_0 \sim \pi_T$ rather than q_T as noted in Bernton et al. [2019]; see Figure 1 for an illustration. This shows that approximating $(B_k^{\mathrm{opt}})_{k=0}^{K-1}$ requires approximating the so-called scores $(\nabla \log q_t(x))_{t\in[0,T]}$. This can be derived heuristically through the fact that a Taylor expansion yields the following approximation of the optimal backward kernels, $B_k^{\mathrm{opt}}(x_k|x_{k+1}) \approx \mathcal{N}(x_k; x_{k+1} - \delta\nabla \log \pi_{k+1}(x_{k+1}) + 2\delta\nabla \log q_{k+1}(x_{k+1}), 2\delta I)$, which indeed corresponds to a Euler discretization of (11); see e.g. [De Bortoli et al., 2021b, Section 2.1].

In SGM [Song et al., 2021b], one gradually adds noise to data using an Ornstein–Ulhenbeck diffusion to transform the complex data distribution into a Gaussian distribution and the generative model is obtained by approximating the time-reversal of this diffusion initialized by Gaussian noise. Practically,

[1] The case of deterministic maps corresponds to normalizing flow components, where the inverse flow is the optimal and only valid reversal, see e.g. [Arbel et al., 2021] which includes a detailed literature review.

the time-reversal approximation is obtained by estimating the scores of the noising diffusion using denoising score matching [Vincent, 2011]. While in our setup, the diffusion (10) instead goes from a simple distribution to a complex one (see Appendix D for a discussion), we can still adapt score matching ideas to our setup. We define a path measure $\mathcal{P}_\theta$ approximating $\mathcal{P}_{\text{opt}}$ using a neural network $s_\theta(T-t, \bar{x}_t)$ in place of $\nabla \log q_{T-t}(\bar{x}_t)$ in (11), i.e. we consider

$$\mathrm{d}\bar{x}_t = \big\{ -\nabla \log \pi_{T-t}(\bar{x}_t) + 2 s_\theta(T-t, \bar{x}_t) \big\} \mathrm{d}t + \sqrt{2} \mathrm{d}\bar{B}_t, \qquad \bar{x}_0 \sim \pi_T. \tag{12}$$

We would like to learn θ by minimizing $D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P}_\theta)$ over θ, i.e. equivalently we maximize a continuous-time ELBO. Note that it is neither feasible to minimize $D_{\mathrm{KL}}(\mathcal{P}^{\text{opt}}||\mathcal{P}_\theta)$ as one cannot sample from π_T nor it is desirable as the evidence estimate is computed using samples from $\mathcal{Q}$. Hence we want the scores to be well-approximated in regions of high-probability mass under $\mathcal{Q}$.

In practice, the diffusions corresponding to $\mathcal{Q}$ and $\mathcal{P}_\theta$ have to be discretized, so a more direct route adopted here is to simply take inspiration from (11) and to consider the parameterized backward kernels $B_k^\theta(x_k|x_{k+1}) = \mathcal{N}(x_k; x_{k+1} - \delta \nabla \log \pi_{k+1}(x_{k+1}) + 2\delta s_\theta(k+1, x_{k+1}), 2\delta I)$ to obtain a parameterized extended target P_θ and corresponding unnormalized target Γ_θ. We then learn θ by minimizing $D_{\mathrm{KL}}(Q||P_\theta)$ where

$$Q(x_{0:K}) = \pi_0(x_0) \textstyle\prod_{k=0}^{K-1} F_{k+1}(x_{k+1}|x_k), \quad P_\theta(x_{0:K}) = \pi(x_K) \textstyle\prod_{k=0}^{K-1} B_k^\theta(x_k|x_{k+1}).$$

This is obviously equivalent to maximizing the ELBO $\mathbb{E}_Q[\log w_\theta(x_{0:K})]$ where $w_\theta(x_{0:K}) = \Gamma_\theta(x_{0:K})/Q(x_{0:K})$. We note that it has previously been proposed to learn parameterized backward kernels for general AIS proposals [Salimans et al., 2015; Huang et al., 2018]. However, the parameterization adopted therein, $B_k^\theta(x_k|x_{k+1}) = \mathcal{N}(x_k; \mu_\theta(x_{k+1}), \Sigma_\theta(x_{k+1}))$, does not leverage the structure of the true reversal and performs poorly experimentally [Thin et al., 2021, Section 4.2].

As established in the next proposition, the continuous and discrete time approaches coincide for $\delta \ll 1$. Once θ is learned, we then obtain an unbiased estimate of Z through $w_\theta(x_{0:K})$ for $x_{0:K} \sim Q$.

Proposition 2. *Under regularity conditions, we have*

$$D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P}_\theta) = \mathbb{E}_{\mathcal{Q}}\Big[\int_0^T ||s_\theta(t, x_t) - \nabla \log q_t(x_t)||^2 \mathrm{d}t \Big] + C_1 \tag{13}$$

$$= \sum_{k=1}^K \int_{t_{k-1}}^{t_k} \mathbb{E}_{\mathcal{Q}} \left[||s_\theta(t, x_t) - \nabla \log q_{t|t_{k-1}}(x_t|x_{t_{k-1}})||^2 \right] \mathrm{d}t + C_2, \tag{14}$$

where $t_k = k\delta$, $K = T/\delta$, $q_{t|s}(x'|x)$ is the density of $x_t = x'$ given $x_s = x$ under $\mathcal{Q}$ and C_1, C_2 constants independent of θ. Let $\mathcal{L}(\theta) = \delta \sum_{k=1}^K \mathbb{E}_Q \left[||s_\theta(k, x_k) - \nabla \log F_k(x_k|x_{k-1})||^2 \right]$ denote a discrete-time approximation of this loss. We have $\nabla D_{\mathrm{KL}}(Q||P_\theta) = \nabla \mathcal{L}(\theta) + \epsilon(\theta)$ for some function ϵ satisfying $\lim_{K\to\infty} \epsilon(\theta) = 0$.

Equation (13) shows that $D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P}_\theta)$ corresponds to a score matching loss as for SGM [Song et al., 2021a]. It is possible to rewrite this loss as (14) so as to replace the intractable score term $\nabla \log q_t(x_t)$ by the easy to approximate gradients of the log-transitions $\nabla \log q_t(x_t|x_{t_{k-1}})$ [Vincent, 2011]. In practice, as mentioned above, we simply learn θ by minimizing the discrete-time KL discrepancy $D_{\mathrm{KL}}(Q||P_\theta)$. This formulation is also very convenient as we can additionally learn potential parameters ϕ of a Q_ϕ using the same criterion. Note from equation (8) that the KL divergence decomposes as firstly a term penalizing the difference in the approximating measure and the fixed target at the final time and secondly another term which can be reduced by optimization of both Q_ϕ and P_θ conditioned on x_T; see e.g. [Agakov and Barber, 2004].

Pseudo-code for our approach (with a comparison to the AIS algorithm proposed in Heng et al. [2020]; Wu et al. [2020]; Thin et al. [2021]) can be found in Algorithm 1.

3.3 Incorporating Hamiltonian dynamics via the underdamped Langevin equation

We now consider a proposal defined on an extended space which arises from the time-discretization of a time-inhomogeneous underdamped Langevin dynamics; see e.g. [Leimkuhler and Matthews, 2016, Chapter 6]. In this scenario, we first focus on continuous-time as the development of suitable numerical integrators is much more involved than for overdamped Langevin diffusions. We consider the diffusion $(x_t, p_t)_{t\in[0,T]}$ where $p_t \in \mathbb{R}^d$ is a momentum variable

$$dx_t = M^{-1} p_t \mathrm{d}t, \qquad dp_t = \nabla \log \pi_t(x_t) \mathrm{d}t - \gamma p_t \mathrm{d}t + \sqrt{2\gamma} M^{1/2} \mathrm{d}B_t \tag{15}$$

Algorithm 1 Unadjusted Langevin AIS/MCD – red instructions for AIS and blue for MCD

Require: Unnormalized target $\gamma(x)$, initial state proposal $\pi_0(x)$, number steps K, stepsize δ, annealing schedule $\{\beta_k\}_{k=0}^K$, score model $s_\theta(k, x)$
Sample $x_0 \sim \pi_0(x_0)$
Set $\log w = -\log \pi_0(x_0)$
for $k = 1$ to K **do**
  Define $\log \gamma_k(\cdot) = \beta_k \log \gamma(\cdot) + (1 - \beta_k) \log \pi_0(\cdot)$
  Define $F_k(x_k|x_{k-1}) = \mathcal{N}(x_k; x_{k-1} + \delta \nabla \log \gamma_k(x_{k-1}), 2\delta I)$
  Sample $x_k \sim F_k(x_k|x_{k-1})$
  Define $B_{k-1}(x_{k-1}|x_k) = F_k(x_{k-1}|x_k)$ ▷ AIS
  Define $B_{k-1}(x_{k-1}|x_k) = \mathcal{N}(x_{k-1}; x_k - \delta \nabla \log \gamma_k(x_k) + 2\delta s_\theta(k, x_k), 2\delta I)$ ▷ MCD
  Set $\log w = \log w + \log B_{k-1}(x_{k-1}|x_k) - \log F_k(x_k|x_{k-1})$
end for
Set $\log w = \log w + \log \gamma(x_K)$

initialized at $x_0 \sim \pi_0, p_0 \sim \mathcal{N}(0, M)$ defining the path measure $\mathcal{Q}$. Here M is a positive definite mass matrix, $\gamma > 0$ a friction coefficient and $(B_t)_{t\in[0,T]}$ a multivariate Brownian motion. If π_t was not time-varying, the invariant distribution of this diffusion would be given by $\bar{\pi}_t(x, p) = \pi_t(x)\mathcal{N}(p; 0, M)$. Intuitively, in the time varying case the SDE will have enough time to approximate each intermediate $\bar{\pi}_t(x, p)$ if we change the target sufficiently slowly. We can think of underdamped Langevin as a continuous-time version of Hamiltonian dynamics with continuous stochastic partial momentum refreshment [Horowitz, 1991].

From [Haussmann and Pardoux, 1986], the time-reversal of the diffusion (15) is also a diffusion process $(\bar{x}_t, \bar{p}_t)_{t\in[0,T]} = (x_{T-t}, p_{T-t})_{t\in[0,T]}$ given by $(\bar{x}_0, \bar{p}_0) \sim \eta_T$ and

$$\begin{aligned} \mathrm{d}\bar{x}_t &= -M^{-1}\bar{p}_t \mathrm{d}t, \\ \mathrm{d}\bar{p}_t &= -\nabla \log \pi_{T-t}(\bar{x}_t)\mathrm{d}t + \gamma \bar{p}_t \mathrm{d}t + 2\gamma M \nabla_{\bar{p}_t} \log \eta_{T-t}(\bar{x}_t, \bar{p}_t)\mathrm{d}t + \sqrt{2\gamma} M^{1/2} \mathrm{d}\bar{B}_t, \end{aligned} \tag{16}$$

where η_t denotes the density (x_t, p_t) under (15). In this case, the continuous-time version of P_{opt} is the path measure $\mathcal{P}_{\text{opt}}$ defined by the diffusion (16) but initialized at $\bar{x}_0 \sim \pi_T, \bar{p}_0 \sim \mathcal{N}(0, M)$ rather than η_T. We will approximate it by the path measure $\mathcal{P}_\theta$ using a neural network $s_\theta(T - t, \bar{x}_t, \bar{p}_t)$ in place of $\nabla \log \eta_{T-t}(\bar{x}_t, \bar{p}_t)$ in (16), i.e. we consider

$$\begin{aligned} \mathrm{d}\bar{x}_t &= -M^{-1}\bar{p}_t \mathrm{d}t, \\ \mathrm{d}\bar{p}_t &= -\nabla \log \pi_{T-t}(\bar{x}_t)\mathrm{d}t + \gamma \bar{p}_t \mathrm{d}t + 2\gamma M s_\theta(T - t, \bar{x}_t, \bar{p}_t)\mathrm{d}t + \sqrt{2\gamma} M^{1/2} \mathrm{d}\bar{B}_t. \end{aligned} \tag{17}$$

As for overdamped Langevin, we could also learn θ by minimizing $D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P}_\theta)$ over θ. This again corresponds to minimizing a score matching loss albeit of a form slightly different from (13).

Proposition 3. *Under regularity conditions, we have*

$$\begin{aligned} D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P}_\theta) &= \gamma \mathbb{E}_{\mathcal{Q}}\Big[\int_0^T ||s_\theta(t, x_t, p_t) - \nabla_{p_t} \log \eta_t(x_t, p_t)||^2 \mathrm{d}t\Big] + C_1 \\ &= \gamma \sum_{k=1}^K \int_{t_{k-1}}^{t_k} \mathbb{E}_{\mathcal{Q}}\left[||s_\theta(t, x_t, p_t) - \nabla_{p_t} \log \eta_{t|t_{k-1}}(x_t, p_t|x_{t_{k-1}}, p_{t_{k-1}})||_M^2\right] \mathrm{d}t + C_2, \end{aligned} \tag{18}$$

where $||x||_M := u^{\mathrm{T}} M u$, $t_k = k\delta$, $K = T/\delta$, C_1, C_2 *are constants independent of* θ *and* $\eta_{t|s}(x', p'|x, p)$ *is the density of* $(x_t, p_t) = (x', p')$ *given* $(x_s, p_s) = (x, p)$ *under* $\mathcal{Q}$.

While the continuous-time perspective shed light on how to parameterize an approximation to the time-reversal, this does not lead directly to an implementable discrete-time algorithm for underdamped Langevin. Contrary to overdamped Langevin, we cannot indeed simply use an Euler discretization of (15) defining $\mathcal{Q}$ and (17) defining $\mathcal{P}_\theta$ to obtain some discrete-time forward and backward kernels and then compute $w_\theta(x_{0:K}, p_{0:K}) = \Gamma_\theta(x_{0:K}, p_{0:K})/Q(x_{0:K}, p_{0:K})$. This is because this ratio is not well-defined due to the lack of noise on the position component in both (15) and (17).

The integrator we use for the forward equation (15), consists in alternating partial momentum refreshments and deterministic leapfrog steps (see e.g [Leimkuhler and Matthews, 2016; Neal, 2011]) giving

$$\tilde{p}_{t+\delta} \sim \mathcal{N}(hp_t, (1 - h^2)M), \qquad (x_{t+\delta}, p_{t+\delta}) = \Phi_t(x_t, \tilde{p}_{t+\delta}), \tag{19}$$

Algorithm 2 Unadjusted Hamiltonian AIS/MCD – red instructions for AIS and blue for MCD

Require: Unnormalized target $\gamma(x)$, initial state proposal $\pi_0(x)$, number steps K, stepsize η, annealing schedule $\{\beta_k\}_{k=0}^K$, damping coefficient h, mass matrix M, score model $s_\theta(k, x, p)$
Sample $x_0 \sim \pi_0(x_0)$, and $p_0 \sim \mathcal{N}(p_0; 0, M)$
Set $\log w = -\log \pi_0(x_0) - \log \mathcal{N}(p_0; 0, M)$
for $k = 1$ to K **do**
  Define $\log \gamma_k(\cdot) = \beta_k \log \gamma(\cdot) + (1 - \beta_k) \log \pi_0(\cdot)$
  Sample $\tilde{p}_k \sim \mathcal{N}(\tilde{p}_k; hp_{k-1}, (1 - h^2)M)$
  Set $\mu_q = p_{k-1}$
  Set $\mu_p = \tilde{p}_k$ ▷ UHA reversal mean
  Set $\mu_p = \tilde{p}_k - 2\log(h)[Ms_\theta(k, x_{k-1}, \tilde{p}_k) + \tilde{p}_k]$ ▷ MCD reversal mean
  Set $\log w = \log w + \log \mathcal{N}(p_{k-1}; h\mu_p, (1 - h)^2 M) - \log \mathcal{N}(\tilde{p}_k; h\mu_q, (1 - h)^2 M)$
  Run leapfrog integrator on γ_k and set $(x_k, p_k) = \Phi(x_{k-1}, \tilde{p}_k)$
end for
Set $\log w = \log w + \log \gamma(x_K) + \log \mathcal{N}(p_K; 0, M)$

with $h = \exp\{-\gamma\delta\}$ and Φ_t is the leapfrog integrator for π_t. The resulting forward sampler is similar to the one proposed by Geffner and Domke [2021], except we do not flip the momentum after the leapfrog step [2]. This integrator may be interpreted as a splitting method for equation (15); see e.g [Leimkuhler and Matthews, 2016, Chapter 7].

We need the integrator for the reversal to fulfill two criteria. First, by definition, as the time step $\delta \to 0$ it must recover the SDE (17). Second, the importance weight of the forward sampler to the reversal must be well defined. Since the leapfrog integrator is a diffeomorphism (or flow) the only possible way to get a well defined reversal for these steps is to take the inverse Φ_t^{-1}. Since the transformation is also volume preserving, the contribution from the deterministic forward and reverse terms will then exactly cancel in the importance weight. The required form of the reverse integrator is

$$(x_t, \tilde{p}_{t+\delta}) = \Phi_t^{-1}(x_{t+\delta}, p_{t+\delta}), \qquad p_t \sim \mathcal{N}(hf^\theta_{t+\delta}(x_t, \tilde{p}_{t+\delta}), (1 - h^2)M), \tag{20}$$

where $f^\theta_{t+\delta}(x_t, \tilde{p}_{t+\delta}) := \widetilde{p}_{t+\delta} + \delta 2\gamma[Ms_\theta(t, x_t, \widetilde{p}_{t+\delta}) + \widetilde{p}_{t+\delta}]$. In Appendix C we show that as $\delta \to 0$ this can indeed be interpreted as a valid split integrator for the reverse SDE (17). The crucial point of algorithmic difference from [Dai et al., 2020; Geffner and Domke, 2021; Zhang et al., 2021] arises from our necessary form for the mean of the reverse momentum refreshment. These works use $h\tilde{p}_{t+\delta}$ as the mean instead of $hf^\theta_{t+\delta}(x_t, \tilde{p}_{t+\delta})$. We again transition to discretized notation with $\delta := T/K$, and $k = 0, ..., K$. In this case, the log importance weight, corresponding to the log evidence estimate, satisfies

$$\log w_\theta(\mathbf{x}, \mathbf{p}) = \log \frac{\gamma(x_K)\mathcal{N}(p_K; 0, M)}{\pi_0(x_0)\mathcal{N}(p_0; 0, M)} + \sum_{k=1}^{K} \log \frac{\mathcal{N}(p_{k-1}; hf^\theta_k(x_{k-1}, \tilde{p}_k), (1 - h^2)M)}{\mathcal{N}(\tilde{p}_k; hp_{k-1}, (1 - h^2)M)}, \tag{21}$$

where $(\mathbf{x}, \mathbf{p})$ denote all the variables introduced by our integration scheme. We can show informally that minimizing $D_{\mathrm{KL}}(Q||P_\theta)$, i.e. maximizing the ELBO given $\mathbb{E}_Q[\log w_\theta(\mathbf{x}, \mathbf{p})]$, again corresponds to minimizing the score matching loss (18) when $\delta \ll 1$; see Appendix C. Pseudo-code for our approach can be found in Algorithm 2.

4 Experiments

We run a number of experiments estimating normalizing constants to validate our approach, MCD and compare to differentiable AIS with ULA [Wu et al., 2020; Thin et al., 2021] and Unadjusted Hamiltonian Annealing (UHA) [Geffner and Domke, 2021; Zhang et al., 2021]. We first investigate the performance of these approaches on static target distributions using the same, fixed initial distribution and annealing schedule. Finally, we explore the performance of the methods for VAEs. Here, being the most expensive of our experiments, we include runtime comparisons of our method compared to baselines. Additional results on a Normalizing Flow target can be found in Appendix F.2. Full experimental details, chosen hyper-parameters, and model architectures can be found in Appendix E.

[2] Geffner and Domke [2021] were not attempting to discretize an underdamped Langevin dynamics.

Our score model is parameterized by an MLP with residual connections that is conditioned on integration time t, and on the momentum term for the Hamiltonian case (see Algorithm 2). For an ablation on various network architectures we refer the reader to Appendiex F.3.

4.1 Static Targets

We estimate the normalizing constants of five simple distributions with known normalizing constants equal to $\log Z = 0$; $\mathcal{N}(10, I)$, $\mathcal{N}(0, 0.1I)$, a Gaussian mixture with 8 components whose means are drawn from $\mathcal{N}(3, I)$ where each component has variance 1, a standard Laplace distribution and a Student's T distribution with 3 degrees of freedom. We use $\mathcal{N}(0, I)$ as our initial distribution for all targets except for the Gaussian mixture and $\mathcal{N}(0, 0.1I)$ which both use $\mathcal{N}(0, 3^2 I)$. We run each method using $K \in \{64, 256\}$ steps and use a fixed, linear annealing schedule. For all methods, sampling step-sizes per-timestep are tuned to via gradient descent to maximize the ELBO and the diagonal mass matrix is learned for the Hamiltonian samplers. Gaussian Mixture and Student-T results can be found in Tables 1 and 2, respectively. Additional results can be found in Appendix F.1.

Sampler	ULA		UHA		ULA-MCD		UHA-MCD	
# steps	64	256	64	256	64	256	64	256
Dim-20	-0.47 ± 0.22	-0.13 ± 0.07	-0.03 ± 0.18	0.015 ± 0.03	0.01 ± 0.02	0.01 ± 0.01	0.01 ± 0.02	-0.01 ± 0.01
Dim-200	-85.62 ± 2.01	-21.98 ± 1.35	-8.20 ± 1.84	-1.26 ± 0.35	-0.25 ± 0.03	0.08 ± 0.122	0.20 ± 0.49	-0.05 ± 0.04
Dim-500	-304.06 ± 5.48	-83.50 ± 7.88	-44.45 ± 3.24	-8.60 ± 1.69	-2.61 ± 1.26	1.01 ± 0.99	-1.74 ± 1.25	-1.02 ± 0.03

Table 1: $\log Z$ estimates for a Gaussian mixture target. Averages and standard errors over 3 seeds.

On average and as expected, UHA outperforms ULA on most targets. Further our approximation to the optimal backward kernels can yield considerable improvements – enabling the ULA sampler to produce better results than UHA (using the standard AIS backward kernels). Further, we see that our approximation to the optimal backward kernels improves the performance of UHA as well.

We further emphasize that often ULA-MCD and UHA-MCD with 64 steps outperform or is on-par with ULA and UHA with 256 steps, a difference of factor 4 in terms of target gradient evaluations. As we show below in Table 3, the additional computational costs of fitting the score model of our method are only roughly twice that of the baselines.

Sampler	ULA		UHA		ULA-MCD		UHA-MCD	
# steps	64	256	64	256	64	256	64	256
Dim-20	-0.09 ± 0.02	-0.02 ± 0.01	-0.04 ± 0.02	-0.02 ± 0.01	-0.06 ± 0.02	-0.00 ± 0.01	-0.03 ± 0.04	-0.01 ± 0.01
Dim-200	-1.63 ± 0.23	-0.88 ± 0.15	-0.36 ± 0.34	-0.10 ± 0.10	-0.82 ± 0.28	-0.30 ± 0.38	-0.48 ± 0.20	-0.10 ± 0.07
Dim-500	-5.43 ± 0.78	-3.10 ± 0.07	-2.86 ± 0.36	-1.03 ± 0.13	-4.00 ± 0.51	-2.23 ± 0.18	-2.03 ± 0.18	0.06 ± 0.30

Table 2: $\log Z$ estimates for a Student-T target. Averages and standard errors over 3 seeds.

4.2 Application to Amortized Inference

Next, we explore the application of our method to amortized inference, in the context of VAEs [Kingma and Welling, 2013; Rezende et al., 2014]. These models are trained to infer latent representations using an inference neural network that consumes an input and produces parameters of an approximation to the true posterior distribution of the underlying generative model. In particular, this posterior distribution is different for each input. When applying AIS to VAE inference [Thin et al., 2021; Geffner and Domke, 2021; Zhang et al., 2021], the output of the inference network parameterizes the initial distribution of the annealing sequence to the true posterior. By training this end-to-end, we effectively learn the initial distribution for the diffusion process. Consequently, the diffusion marginals and their score vectors $\nabla \log q_t$ are different for every input, and we need to

condition our score model s_θ on the inputs to reflect that. We achieve this by projecting the last hidden layer of the inference network into a summary vector that is concatenated to the other conditioning inputs of s_θ.

We train a VAE on the binarized MNIST dataset [Salakhutdinov and Murray, 2008], re-using architectures proposed in [Geffner and Domke, 2021; Burda et al., 2016] (two-layer MLP encoder/decoder, Bernoulli likelihood). All generative models use the same architecture and hyper-parameters. We compare standard amortized variational inference with annealed ULA and UHA with standard AIS backward transition kernels, as well as ULA and UHA with our MCD transition kernels. We match the number of sampler steps between ULA/ULA-MCD and UHA/UHA-MCD to 64 and 32 respectively. ELBO and log-likelihood on the test set can be found in Table 3.

Sampler	VI	ULA	UHA	ULA-MCD	UHA-MCD
ELBO	-96.32 $\pm$ 0.40	-90.41 $\pm$ 0.17	-88.58 $\pm$ 0.51	-90.10 $\pm$ 0.10	-88.08 $\pm$ 0.07
Log-lik.	-69.35 $\pm$ 0.36	-62.43 $\pm$ 0.25	-61.05 $\pm$ 1.84	-61.83 $\pm$ 0.42	-58.58 $\pm$ 0.34
Iteration time	0.024s	0.055s	0.050s	0.101s	0.098s
Total time	3263.40s	4694.52s	5072.19s	11304.36s	11345.59s

Table 3: Test set performance for MNIST VAE. Averages and standard errors over 5 seeds. We additionally report runtimes for a single training iteration, and total experiment time.

We see that, as reported in prior works [Geffner and Domke, 2021; Thin et al., 2021; Zhang et al., 2021], Monte Carlo based inference methods provide a significant benefit over standard amortized variational inference. In addition, our learned backward kernels lead to improved performance of both ULA and UHA, with the difference being more distinct for the latter. We also note that UHA has significantly larger error bars than all other methods, but UHA-MCD does not appear to inherit this. We further provide a runtime comparison in Table 3, both for a single training iteration, and total experiment time (including evaluation). We find that fitting the score models of our method results in roughly twice as much runtime (recall we fixed the number of sampler steps).

5 Limitations

While we have demonstrated that our optimized backward kernels can lead to large improvements over standard AIS backward kernels, our proposed approach has a number of limitations. First, since our method relies upon unadjusted Langevin or Hamiltonian sampling, we inherit many of the issues with these approaches. Using unadjusted samplers enables a fully-differentiable ELBO estimate which can in theory be used to tune the many parameters of the sampler. However, these samplers require repeated gradient steps as an inner-loop which can lead to divergence dynamics and numerical instability, making optimization difficult. Next, our proposed training procedure has a notable increase in memory consumption as we store the entire sampling trajectory in memory to train our neural network. We further note that it is possible to compute our importance weights online using $\mathcal{O}(1)$ memory. We could utilize this with Monte Carlo sub-sampling of the timestep (as with SGM [Ho et al., 2020]) to derive a constant memory training procedure for our approximate time-reversal at the cost of increased variance but we leave exploring this to future work. Finally, we observed that, without mitigation, parameterized forward samplers Q_ϕ sometimes dropped modes during training. Since this is a known challenge of using reverse KL, this might be improved by using alternative optimization objectives [Masrani et al., 2019; Matthews et al., 2022].

6 Discussion

In this work we have explored AIS using unadjusted Langevin and Hamiltonian dynamics. We have demonstrated that the backward transition kernels typically used are suboptimal and we have presented the form of the optimal variance-minimizing backward kernels. We have further shown how an approximation to these kernels can be learned using score matching and that this objective converges to maximizing the ELBO in the limit of infinitesimally small time-discretization step-sizes. We have illustrated the benefit of using our proposed optimized backward kernels on a number of inference problems including fixed targets and amortized tasks with model learning.

References

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

Arbel, M., Matthews, A. G., and Doucet, A. (2021). Annealed flow transport Monte Carlo. In *International Conference on Machine Learning*.

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Babuschkin, I., Baumli, K., Bell, A., Bhupatiraju, S., Bruce, J., Buchlovsky, P., Budden, D., Cai, T., Clark, A., Danihelka, I., Fantacci, C., Godwin, J., Jones, C., Hennigan, T., Hessel, M., Kapturowski, S., Keck, T., Kemaev, I., King, M., Martens, L., Mikulik, V., Norman, T., Quan, J., Papamakarios, G., Ring, R., Ruiz, F., Sanchez, A., Schneider, R., Sezener, E., Spencer, S., Srinivasan, S., Stokowiec, W., and Viola, F. (2020). The DeepMind JAX Ecosystem.

Bernton, E., Heng, J., Doucet, A., and Jacob, P. E. (2019). Schrödinger bridge samplers. *arXiv preprint arXiv:1912.13170*.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. (2018). JAX: composable transformations of Python+NumPy programs.

Brekelmans, R., Huang, S., Ghassemi, M., Ver Steeg, G., Grosse, R. B., and Makhzani, A. (2022). Improving mutual information estimation with annealed and energy-based bounds. In *International Conference on Learning Representations*.

Burda, Y., Grosse, R. B., and Salakhutdinov, R. (2016). Importance weighted autoencoders. In *International Conference on Learning Representations*.

Crooks, G. E. (1998). Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems. *Journal of Statistical Physics*, 90(5-6):1481–1487.

Dai, C., Heng, J., Jacob, P. E., and Whiteley, N. (2020). An invitation to sequential Monte Carlo samplers. *arXiv preprint arXiv:2007.11936*.

De Bortoli, V., Doucet, A., Heng, J., and Thornton, J. (2021a). Simulating diffusion bridges with score matching. *arXiv preprint arXiv:2111.07243*.

De Bortoli, V., Thornton, J., Heng, J., and Doucet, A. (2021b). Diffusion Schrödinger bridge with applications to score-based generative modeling. In *Advances in Neural Information Processing Systems*.

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

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

Dockhorn, T., Vahdat, A., and Kreis, K. (2022). Score-based generative modeling with critically-damped Langevin diffusion. In *International Conference on Learning Representations*.

Fournier, N. and Tardif, C. (2021). On the simulated annealing in $\mathbb{R}^d$. *Journal of Functional Analysis*, 281(5):109086.

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

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185.

Grillo, G. (1994). Logarithmic sobolev inequalities and langevin algorithms in $\mathbf{R}^n$. *Stochastic Analysis and Applications*, 12(3):309–328.

Grosse, R. B., Maddison, C. J., and Salakhutdinov, R. R. (2013). Annealing between distributions by averaging moments. In *Advances in Neural Information Processing Systems*.

Haussmann, U. G. and Pardoux, E. (1986). Time reversal of diffusions. *The Annals of Probability*, 14(3):1188–1205.

Heng, J., Bishop, A. N., Deligiannidis, G., and Doucet, A. (2020). Controlled sequential Monte Carlo. *The Annals of Statistics*, 48(5):2904–2929.

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

Horowitz, A. M. (1991). A generalized guided Monte Carlo algorithm. *Physics Letters B*, 268(2):247–252.

Huang, C.-W., Tan, S., Lacoste, A., and Courville, A. C. (2018). Improving explorability in variational inference with annealed variational objectives. In *Advances in Neural Information Processing Systems*.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *The Journal of Machine Learning Research*, 6:695–709.

Jarzynski, C. (1997). Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14):2690–2963.

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

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

Klebaner, F. C. (2012). *Introduction to Stochastic Calculus with Applications*. Imperial College Press.

Leimkuhler, B. and Matthews, C. (2016). *Molecular Dynamics with Deterministic and Stochastic Numerical Methods*. Springer.

Masrani, V., Brekelmans, R., Bui, T., Nielsen, F., Galstyan, A., Ver Steeg, G., and Wood, F. (2021). q-paths: Generalizing the geometric annealing path using power means. In *Uncertainty in Artificial Intelligence*.

Masrani, V., Le, T. A., and Wood, F. (2019). The thermodynamic variational objective. In *Advances in Neural Information Processing Systems*.

Matthews, A. G. D. G., Arbel, M., Rezende, D. J., and Doucet, A. (2022). Continual repeated annealed flow transport Monte Carlo. In *International Conference on Machine Learning*.

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

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

Opper, M. (2019). Variational inference for stochastic differential equations. *Annalen der Physik*, 531:1800233.

Ramachandran, P., Zoph, B., and Le, Q. V. (2017). Searching for activation functions. *arXiv preprint arXiv:1710.05941*.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*.

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

Salakhutdinov, R. and Murray, I. (2008). On the quantitative analysis of deep belief networks. In *International Conference on Machine Learning*.

Salimans, T., Kingma, D., and Welling, M. (2015). Markov chain Monte Carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*.

Schmiedl, T. and Seifert, U. (2007). Optimal finite-time processes in stochastic thermodynamics. *Physical Review Letters*, 98(10):108301.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*.

Song, Y., Durkan, C., Murray, I., and Ermon, S. (2021a). Maximum likelihood training of score-based diffusion models. In *Advances in Neural Information Processing Systems*.

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

Sottinen, T. and Särkkä, S. (2008). Application of Girsanov theorem to particle filtering of discretely observed continuous-time non-linear systems. *Bayesian Analysis*, 3(3):555–584.

Tang, W. and Zhou, X. Y. (2021). Simulated annealing from continuum to discretization: a convergence analysis via the eyring–kramers law. *arXiv preprint arXiv:2102.02339*.

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

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

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

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

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Paper is methodological paper on approximate inference methods with only standard datasets.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See proofs of main results in Appendix.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Appendix for architectures, Algorithm 2 for pseudo-code. We used standard datasets from the literature. We did not publish code.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See experimental section and Appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Where appropriate and feasible.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]