Adaptive Annealed Importance Sampling with Constant Rate Progress

Shirin Goshtasbpour [1 2] **Victor Cohen** [2] **Fernando Perez-Cruz** [1 2]

Abstract

Annealed Importance Sampling (AIS) synthesizes weighted samples from an intractable distribution given its unnormalized density function. This algorithm relies on a sequence of interpolating distributions bridging the target to an initial tractable distribution such as the well-known geometric mean path of unnormalized distributions which is assumed to be suboptimal in general. In this paper, we prove that the geometric annealing corresponds to the distribution path that minimizes the KL divergence between the current particle distribution and the desired target when the feasible change in the particle distribution is constrained. Following this observation, we derive the constant rate discretization schedule for this annealing sequence, which adjusts the schedule to the difficulty of moving samples between the initial and the target distributions. We further extend our results to f-divergences and present the respective dynamics of annealing sequences based on which we propose the Constant Rate AIS (CR-AIS) algorithm and its efficient implementation for α-divergences. We empirically show that CR-AIS performs well on multiple benchmark distributions while avoiding the computationally expensive tuning loop in existing Adaptive AIS.

1. Introduction

Annealed Importance Sampling (AIS) (Neal, 2001) is one of the most popular sampling methods to estimate intractable expectations given an unnormalized density of a distribution. Together with its other variants such as thermodynamic integration (Ogata, 1989; Gelman & Meng, 1998) and Sequential Monte Carlo (SMC) (Del Moral et al., 2006) this algorithm has vast applications, such as marginal likelihood estimation (Salakhutdinov & Murray, 2008; Grosse et al., 2015; 2016), moment estimation (Johansen et al., 2015; Jasra et al., 2011), generative model evaluation (Wu et al., 2017) and more recently it has been incorporated in variational inference and training of deep generative networks (Maddison et al., 2017; Naesseth et al., 2018; Wu et al., 2020; Thin et al., 2021; Masrani et al., 2019).

To perform annealing, this algorithm uses a sequence of bridging distributions between proposal distribution and the target which is chosen in advance. Gelman & Meng (1998) have demonstrated that the optimal path with lowest variance for thermodynamic integral estimator depends on the Hellinger distance of the distributions and it is intractable in complex setups. Instead, the geometric mean path has been utilized for years (Neal, 2001; 1996). As alternatives, moment-averaging have been proposed for exponential family distributions (Grosse et al., 2013) and further generalized to power mean for arbitrary endpoint distributions (Masrani et al., 2021; Brekelmans et al., 2020). These heuristic annealing paths achieve viable estimation results even though they are considered to be suboptimal.

In this work, we analyze a version of AIS algorithm where we apply infinitesimal changes to the initial density along the annealing distribution path to get to the target distribution. We take a greedy approach and modify the particle distribution in the direction that optimally reduces the remaining estimation bias at every instance. The remaining bias is equivalent to the inverse KL-divergence between current particle distribution and the target distribution under common assumptions (Grosse et al., 2013) and we prove that in this setup the optimal greedy strategy is achieved using the geometric mean path. Extending our analysis to the larger class of f-divergences, we are able to derive an Ordinary Differential Equation (ODE) for the optimal greedy annealing dynamics. In the subclass of α-divergences, this ODE has a closed form solution and we show that power mean annealing is a solution to this equation.

While other variational representations of geometric and power mean paths have been provided in the previous work (Grosse et al., 2013; Masrani et al., 2021), to the best of our knowledge, we are the first to show the relation of these annealing sequences with the functional derivatives of various probability divergences. Using this framework,

[*]Equal contribution [1]Computer Science Department, ETH Zurich, Zurich, Switzerland [2]Swiss Data Science Center, Zurich, Switzerland. Correspondence to: Shirin Goshtasbpour <shirin.goshtasbpour@inf.ethz.ch>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

we are able to derive the constant rate schedule along the steepest descent annealing path.

Our greedy strategy is similar to existing Adaptive AIS in that we look ahead and measure the impact of each annealing step base on an objective function. However, in a typical Adaptive AIS algorithm, the schedule is adjusted in each step to keep the reduction of the Effective Sample Size (ESS) (Kong, 1992; Neal, 2001) or the Conditional ESS (CESS) (Johansen et al., 2015) at a constant rate using iterative search algorithms. Instead, we derive the annealing distribution path and its corresponding schedule using the same objective derivative. Therefore, the constant rate schedule is tightly connected to the bridging distributions and is able to account for the difficulty of synthesis along each annealing sequence.

Finally, we design the Constant Rate AIS (CR-AIS) algorithm to approximate the constant rate schedule of the variational objectives. We present multiple considerations for its practical implementation. CR-AIS does not rely on searching algorithms and excessive target density function evaluations as in adaptive versions of AIS and uses the information from the derivative of the objective to choose the bridging distributions. Using this algorithm we empirically verify our findings on high dimensional targets and illustrate how CR-AIS is able to trade-off computation complexity with estimation accuracy while improving adaptivity.

2. Annealed Importance Sampling

Formally, suppose P and Q_0 are two probability distributions on $\mathbb{R}^d$ with density functions π and q_0, respectively. We assume evaluation of q_0 is tractable while $\pi = \tilde{\pi}/Z_\pi$ is only known up to the normalization constant $Z_\pi = \int_{\mathbb{R}^d} \tilde{\pi}(z)dz$. To sample from π, AIS uses a sequence of annealing distributions defined by the density path $\gamma : [0,1] \times \mathbb{R}^d \to \mathbb{R}_+$ where $\gamma(t,\cdot) \in \mathcal{P}$ for all $t \in [0,1]$ starting from $\gamma(0,\cdot) = q_0(\cdot)$ and reaching to $\gamma(1,\cdot) = \pi(\cdot)$ and $\mathcal{P}$ is the family of normalized density functions. This path is discretized with the schedule $0 = t_0 < ... < t_M = 1$. Common choices for schedule are linear discretization with $t_i = i/M$, exponential with $t_i = 1 - \epsilon^i$ and sigmoidal with $t_i = \sigma(c(i/M - 0.5))$ for hyperparameters $\epsilon < 1$, $0 < c$ and $\sigma(x) = 1/(1 + e^{-x})$.

A Markov process $\overrightarrow{q}(z_{0:M}) = q_0(z_0) \prod_{i\in[M]} \overrightarrow{q}_i(z_i|z_{i-1})$ is used for sampling such that particles sampled from the initial distribution, $z_0 \sim q_0$, gradually move following each transition probability $\overrightarrow{q}_i$ to have a marginal distribution close to $\gamma(t_i,\cdot)$. An auxiliary backward Markov chain $\overleftarrow{q}_i(z_{i-1}|z_i)$ allows us to compute the unnormalized importance weights corresponding to each particle trajectory $z_{0:M}$,

$$w(z_{0:M}) = \frac{\tilde{\pi}(z_M) \prod_i \overleftarrow{q}_i(z_{i-1}|z_i)}{q_0(z_0) \prod_i \overrightarrow{q}_i(z_i|z_{i-1})}.$$

It is common to define the transition probabilities $\overrightarrow{q}_i$ to be reversible with respect to $\gamma(t_i,\cdot)$ (e.g. a Markov chain with Metropolis-Hastings corrected transition kernels) and select $\overleftarrow{q}_i$ as its reversal (Neal, 2001). Let us denote the unnormalized density path with $\tilde{\gamma}(t,\cdot) = Z_{\gamma_t} \cdot \gamma(t,\cdot)$, where $Z_{\gamma_t} = \int_{\mathbb{R}^d} \tilde{\gamma}(t,z)dz$. Therefore, with reversible transitions, $\overleftarrow{q}_i(z_{i-1}|z_i) = \overrightarrow{q}_i(z_i|z_{i-1})\tilde{\gamma}(t_i,z_{i-1})/\tilde{\gamma}(t_i,z_i)$, we can rewrite the importance weights as $w(z_{0:M}) = \prod_i \tilde{\gamma}(t_i,z_{i-1})/\tilde{\gamma}(t_{i-1},z_{i-1})$. The unbiased Monte Carlo (MC) estimator of the partition function from N sampled particle instances $z^{(j)}_{0:M} \sim q_0 \prod_i \overrightarrow{q}_i$ for $1 \le j \le N$ is

$$Z_\pi \approx \hat{Z}_\pi = \frac{1}{N} \sum_j w(z^{(j)}_{0:M}).$$

Although, to avoid numerical underflow, log space computations are preferred and $\log Z_\pi$ is bounded from below by (Grosse et al., 2015; Domke & Sheldon, 2018)

$$\begin{aligned} \log Z_\pi \ge & \mathbb{E}_{\overrightarrow{q}}[\log w(z_{0:M})] \\ \approx & \frac{1}{N} \sum_j \log w(z^{(j)}_{0:M}). \end{aligned} \tag{1}$$

It is possible to estimate the expectation of a test function $h : \mathbb{R}^d \to \mathbb{R}$ under the target distribution via self-normalized weighted average of $h(z^{(j)}_M)$ or equivalently by resampling the particles according to a multinomial distribution where $z^{(j)}_M$ is sampled proportional to $w(z^{(j)}_{0:M})$.

3. Adaptive Annealing Dynamics from Divergence Derivative

Given the particle trajectory, $z_{0:M} \sim \overrightarrow{q}$, we denote the density of the marginal distribution of z_i with $q_{t_i}(z_i)$. It is common to analyze the AIS algorithm in perfect transition regime (i.e. when $q_{t_i}(z) = \gamma(t_i,z)$ q_{t_i}-a.s.) with reversible transition kernels (Grosse et al., 2013; Kiwaki, 2015). This is not unrealistic if $t_i - t_{i-1}$ is sufficiently small and the consecutive annealing distributions are close to each other. We assume the same conditions apply in our paper. Under this setup, Grosse et al. (2013) decomposed the bias of the estimator in Equation (1) as the sum of KL divergences between consecutive annealing distributions (Grosse et al., 2013),

$$\log Z_\pi - \mathbb{E}_{\overrightarrow{q}}[\log w(z_{0:M})] = \sum_{i=1}^{M} \mathrm{D}_{\mathrm{KL}}(q_{t_{i-1}}||q_{t_i}), \tag{2}$$

where $\mathrm{D}_{\mathrm{KL}}(q||p) = \int q(z)\log(q(z)/p(z))dz$. In fact, as $M \to \infty$ the asymptotic bias decreases as

$$M\sum_{i=1}^{M}\mathrm{D}_{\mathrm{KL}}\left(q_{t_{i-1}}||q_{t_i}\right) \to \frac{1}{2}\int_0^1 \mathrm{Var}_{q_t}[\frac{d}{dt}\log q_t]dt. \quad (3)$$

The end-to-end asymptotic bias in Equation (3) was used to compare the efficiency of moment-averaging and geometric mean paths in (Grosse et al., 2013) and optimized with respect to the single dimensional schedule function for a given density path $\tilde{\gamma}$ in (Kiwaki, 2015). Their method does not scale to higher dimensional function space to optimize the annealing density path. Instead, we propose an adaptive approach to maximize the reduction in bias with every infinitesimal transition and we use the derivative of KL-divergence at the current particle distribution to find the optimal change in the annealing density path. By doing so, we are able to extend the optimization space from the space of discretization schedules to the space of unnormalized annealing density paths. In the following, we explain the details of our method. Proof of the results are provided in Appendix A.

3.1. Inverse KL Divergence Dynamics

Let $\tilde{q}_t(z) = \tilde{\gamma}(t,z)$ be the unnormalized marginal density at instant t and $J_{\mathrm{KL}}[\phi_t]$ define the functional of $\phi_t(z) = \log\tilde{q}_t(z)$ corresponding to the inverse KL divergence,

$$\begin{aligned} J_{\mathrm{KL}}[\phi_t] =& \mathrm{D}_{\mathrm{KL}}(q_t||\pi) \quad (4)\\ =& \frac{\int \tilde{q}_t(z)\log\frac{\tilde{q}_t(z)}{\tilde{\pi}(z)}dz}{\int \tilde{q}_t(z)dz} + \log Z_\pi - \log\int \tilde{q}_t(z)dz. \end{aligned}$$

In our greedy strategy, at step i, we fix all the previous annealing distributions up to q_{t_i} and consider none of the subsequent ones other than $q_1 = \pi$. We choose the next distribution $q_{t_{i+1}}$ as an infinitesimal modification of q_{t_i} which minimizes the updated sampling bias. We can derive the updated bias recursively from Equation (2) starting from $b_0 = \mathrm{D}_{\mathrm{KL}}(q_0||\pi)$ and repeating

$$\begin{aligned} b_i =& b_{i-1} + \mathrm{D}_{\mathrm{KL}}(q_{t_i}||q_{t_{i+1}}) \\ &+ \mathrm{D}_{\mathrm{KL}}(q_{t_{i+1}}||\pi) - \mathrm{D}_{\mathrm{KL}}(q_{t_i}||\pi), \quad (5) \end{aligned}$$

for $i < M$. To find $q_{t_{i+1}}$ we take the directional functional derivative of b_i and minimize it in a compact space to find the steepest descent direction that leads to the optimal annealing. As the first and last terms in Equation (5) are constant with respect to $q_{t_{i+1}}$ and the derivative of the second term is zero, the directional derivative is equivalent to the directional derivative of $J_{\mathrm{KL}}[\phi_{t_i}]$. First, we derive the directional derivative of J_{KL} in the following Lemma. Then, in Proposition 3.2, we show that the geometric path,

$$\log\tilde{q}_t^{\mathrm{geom}}(z) := (1-t)\log q_0(z) + t\log\tilde{\pi}(z), \quad (6)$$

corresponds to the path following this direction and is optimal in this sense.

To take the derivative of J_{KL}, we transform the negative energy ϕ_t by a small perturbation in direction of the smooth function $\eta : \mathbb{R}^d \to \mathbb{R}$ with a small step size $\epsilon > 0$.

Lemma 3.1. *Assume $\tilde{q}_t(z)$ and $\tilde{\pi}(z)$ are positive unnormalized density functions and let $\phi_{t+\epsilon}(z) = \phi_t(z) + \epsilon\eta(z)$ for $\phi_t(z) = \log\tilde{q}_t(z)$. Then we have,*

$$\frac{d}{d\epsilon}J_{\mathrm{KL}}[\phi_{t+\epsilon}]\Big|_{\epsilon=0} = \mathrm{Cov}_{q_t}\left[\eta(z), \log\frac{\tilde{q}_t(z)}{\tilde{\pi}(z)}\right], \quad (7)$$

where $\mathrm{Cov}_q[\cdot,\cdot]$ is the covariance under distribution of q and we use the definition of Gâteaux differential for the derivative,

$$\frac{d}{d\epsilon}J_{\mathrm{KL}}[\phi_{t+\epsilon}]\Big|_{\epsilon=0} = \lim_{\epsilon\to 0^+}\frac{J_{\mathrm{KL}}[\phi_t(z)+\epsilon\eta(z)] - J_{\mathrm{KL}}[\phi_t]}{\epsilon}.$$

To identify the optimal perturbation direction, $\eta^*_{q_t,\pi}$, we minimize $\mathrm{Cov}_{q_t}\left[\eta(z), \log\frac{\tilde{q}_t(z)}{\tilde{\pi}(z)}\right]$ with respect to η in the space of smooth functions with bounded variance. Using a bound on the variance of the perturbation as opposed to its norm is explained by the fact that the expectation of the perturbation only impacts the normalization factor of the bridging density function and does not affect the performance of AIS. Therefore, by constraining the perturbations to have bounded variance, we are accounting for the equivalency of annealing paths with different time dependent scaling.

The optimal perturbation $\eta^*_{q_t,\pi}$ is the steepest descent direction of the inverse KL divergence which we can use to derive the annealing dynamics via

$$\frac{d}{dt}\log\tilde{q}(z) = \eta^*_{q_t,\pi}(z) + b(t). \quad (8)$$

Note that $b(t) : \mathbb{R} \to \mathbb{R}$ in Equation (8) is an arbitrary log scale function of t which can be absorbed by $\eta^*_{q_t,\pi}(z)$ for each t without loss of generality. In the following, we show that with initial distribution density $\tilde{q}_0$ and following the infinitesimal perturbations in direction of $\eta^*_{q_t,\pi}$ we recover an arbitrarily scaled geometric mean path. In addition to the dynamics of the optimal greedy annealing path, we derive a discretization schedule in the following proposition which ensures a steady decrease in J_{KL} as AIS algorithm proceeds.

Proposition 3.2. *Assume the same conditions as in Lemma 3.1. Additionally, consider the set of smooth perturbation directions with bounded variance*

$$\mathcal{M}_{q_t,\pi} := \{\eta \in \mathcal{C}^1 : \mathrm{Var}_{q_t}[\eta(z)] \le c^{KL}_{q_t,\pi}\},$$

for $B \ge 0$ and $c^{KL}_{q_t,\pi} = B/\mathrm{Var}_{q_t}[\log(\tilde{\pi}(z)/\tilde{q}_t(z))]$. Then the steepest descent direction that minimizes the derivative

in Equation (7) in $\mathcal{M}_{q_t,\pi}$ is

$$\eta^*_{q_t,\pi}(z) = \frac{c^{KL}_{q_t,\pi}}{\sqrt{B}} \log \frac{\tilde{\pi}(z)}{\tilde{q}_t(z)} + b, \tag{9}$$

for arbitrary $b \in \mathbb{R}$. A solution to the Ordinary Differential Equation (ODE) $\frac{d}{dt}\phi_t(z) = \eta^*_{q_t,\pi}(z)$ with initial condition $\phi_0(z) = \log \tilde{q}_0(z)$ is the scaled geometric mean path $\log \tilde{q}^{geom}_{1-\beta(t)}(z)$ which for $\beta(t)$ set as

$$\beta^{KL}(t) \coloneqq e^{-\int_0^t c^{KL}_{q_r,\pi} dr/\sqrt{B}}, \tag{10}$$

decreases the inverse KL divergence in Equation (4) with constant rate.

Note that the annealing process described by Proposition 3.2 encompasses the duration $0 < t$ as opposed to $t \in [0, 1]$ in the original AIS algorithm. This is due to taking infinitesimal annealing steps along the derivatives of J_{KL}. In particular, the annealing schedule in Proposition 3.2 is defined with $\tau(t) = 1-\beta(t)$ and for $\beta(t) = \beta^{\mathrm{KL}}(t)$ as $t \to \infty$, $\tau(t) \to 1^-$ and $q_t(z) \to \pi(z)$ for all $z \in \mathbb{R}^d$, although the normalization factor of $\tilde{q}_t$ may grow to infinity. Additionally, due to the derivative of $\tau(t)$,

$$\dot{\tau}(t) \propto (1 - \tau(t))/\mathrm{Var}_{q_t}[\log(\tilde{\pi}/\tilde{q}_t)],$$

annealing is slower when the particle distribution is further away from the target (i.e. in the beginning of the annealing process). The constant rate schedule enforces a balanced division of the sampling difficulty per annealing step in comparison to the heuristics such as linear. In other words, to converge to the target, in each iteration, the particle distribution is altered by doing an equal amount of work which is measured in terms of the reduction of the KL-divergence.

Due to the logarithmic term in the derivative, annealing is slower when the particle distribution extends beyond the target distribution support, while being less sensitive to unexplored modes in the target distribution with a damped sensitivity as t grows. However, in Section 5.1 we show this not to be an issue as using this schedule results in coverage of all of the modes of multimodal targets in the experiments.

3.2. Extension to f-Divergences

In this section, we extend our method to f-divergences to explore other dynamics that are optimal with respect to alternative step-wise objectives. The f-divergence between two distributions with unnormalized densities $\tilde{\pi}(z)$ and $\tilde{q}_t(z)$ is defined as

$$J_f[\phi_t] = \mathrm{D}_f(\pi||q_t) = \frac{\int \tilde{q}_t(z) f\left(u_t(z)\right) dz}{\int \tilde{q}_t(z) dz},$$

where $f : \mathbb{R} \to \mathbb{R}$ is convex, lower-semicontinuous function with $f(1) = 0$ and

$$u_t(z) = \frac{\tilde{\pi}(z)}{\tilde{q}_t(z)} / \mathbb{E}_{q_t}\left[\frac{\tilde{\pi}(z)}{\tilde{q}_t(z)}\right].$$

Following a similar approach to the previous section, we can find the optimal perturbation of ϕ_t and use Equation (8) to obtain the annealing unnormalized density dynamics along the steepest descent direction. In Lemma 3.3 we derive the steepest descent direction of f-divergence.

Lemma 3.3. *Assume $\tilde{q}_t(z)$ and $\tilde{\pi}(z)$ are positive unnormalized density functions and $f : \mathbb{R} \to \mathbb{R}$ be convex and differentiable. Let $\phi_{t+\varepsilon(z)} = \phi_t(z) + \epsilon\eta(z)$ for $\phi_t(z) = \log \tilde{q}_t(z)$. Then we have,*

$$\frac{d}{d\epsilon} J_f[\phi_{t+\epsilon}]\Big|_{\epsilon=0} = \mathrm{Cov}_{q_t}\left[\eta(z), -g(u_t(z))\right],$$

where $g(u) = u\dot{f}(u) - f(u)$ and $\dot{f}(u) = df(u)/du$. Moreover, consider the set of smooth perturbation directions with bounded variance $\mathcal{M}^f_{q_t,\pi} \coloneqq \left\{\eta \in \mathcal{C}^1 : \mathrm{Var}_{q_t}[\eta(z)] \le c^f_{q_t,\pi}\right\}$ for $B \ge 0$ and

$$c^f_{q_t,\pi} = B/\mathrm{Var}_{q_t}\left[g(u_t(z))\right].$$

Then the steepest descent direction that minimizes this derivative in $\mathcal{M}^f_{q_t,\pi}$ is

$$\eta^*_{q_t,\pi}(z) = \frac{c^f_{q_t,\pi}}{\sqrt{B}} g(u_t(z)) + b,$$

for arbitrary $b \in \mathbb{R}$.

Unfortunately, the solution of ODE in Equation (8) with the optimal perturbation direction does not have a closed form for general f functions. However, we can get a set of solutions for the specific case of α-divergences in the following Proposition, which correspond to the power mean annealing path previously proposed in (Brekelmans et al., 2020)[1],

$$\tilde{q}^{\alpha\text{-pow}}_t(z) = \left(t\tilde{\pi}(z)^\alpha + (1-t)\tilde{q}_0(z)^\alpha\right)^{\frac{1}{\alpha}}. \tag{11}$$

Proposition 3.4. *Assume the same conditions as in Lemma 3.3 and $f(u) = (u^\alpha - 1 - \alpha(u-1))/\alpha(\alpha-1)$ for $\alpha \notin \{0, 1\}$ or $f(u) = u \log u$ for $\alpha = 1$. Then α-power mean path $\log \tilde{q}^{\alpha\text{-pow}}_{1-\beta(t)}$ is a solution to the ODE $\frac{d}{dt}\phi_t(z) = \eta^*_{q_t,\pi}(z)$ with initial condition $\phi_0(z) = \log \tilde{q}_0(z)$ and setting $\beta(t)$ to*

$$\beta^\alpha(t) \coloneqq e^{-\int_0^t (c^f_{q_r,\pi} Z^\alpha_{q_r}/\sqrt{B} Z^\alpha_\pi) dr}, \tag{12}$$

results in constant rate decrease in f-divergence.

Here, the annealing speed is inversely related to $\mathrm{Var}_{q_t}[g(\pi/q_t)]$. With $\alpha = 1$ the annealing dynamics correspond to the arithmetic mean path (moment-averaging path in the exponential family (Grosse et al., 2013)). We have listed a few of the popular choices of α-divergences with their respective f and g functions in Table 1.

[1] For illustrations of bridging distributions with different α values we refer the readers to the works of Brekelmans et al. (2020) and Masrani et al. (2021)

Table 1. List of f-divergences

NAME	$f(u)$	$g(u)$	α
KL DIVERGENCE	$u \log u$	u	1^*
INVERSE KL DIVERGENCE	$-\log u$	$\log u - 1$	0^*
PEARSON χ^2	$(u-1)^2$	$u^2 - 1$	2
NEYMAN χ^2	$\frac{(u-1)^2}{u}$	$2 - \frac{2}{u}$	-1
SQUARED HELLINGER	$(\sqrt{u}-1)^2$	$\sqrt{u} - 1$	$1/2$
α-DIV ($\alpha \notin \{0,1\}$)	$\frac{u^\alpha - \alpha u}{\alpha(\alpha-1)} + \frac{1}{\alpha}$	$\frac{1}{\alpha}(u^\alpha + 1)$	α

4. Constant Rate AIS

Using Proposition 3.4, we can design an adaptive AIS algorithm with constant rate decrease in α-divergence at each annealing iteration. In Algorithm 1, we provide a pseudocode for Constant Rate AIS (CR-AIS) where we alternate between updating the particle location with standard AIS steps and tuning the schedule. Evaluation of the constant rate schedule $\tau(t) = 1 - \beta^\alpha(t)$ given in Equation (12) at time t requires the values of $c^f_{q_r,\pi}$ and Z_π / Z_{q_r} for all $0 \le r \le t$ and integration. We use weighted AIS particles up to step i to estimate the integrand in $\beta^\alpha(t_i)$. We set $t_i = i\delta$ for small $\delta > 0$ and approximate the integral with Riemann sum

$$\int_0^{i\delta} (c^f_{q_r,\pi} Z^\alpha_{q_r} / \sqrt{B} Z^\alpha_\pi) dr \approx \sum_{k=0}^{i} \delta c^f_{q_{k\delta},\pi} Z^\alpha_{q_{k\delta}} / \sqrt{B} Z^\alpha_\pi,$$

where $q_{k\delta} = q^{\alpha\text{-pow}}_{1-\beta^\alpha(k\delta)}$ is the normalized density of power mean path in Equation (11). We set $B = 1$ for simplicity. We can rewrite the above equation incrementally, noting that

$$\begin{aligned}\beta^\alpha((i+1)\delta) &= \beta^\alpha(i\delta) e^{-\int_{i\delta}^{(i+1)\delta} (c^f_{q_r,\pi} Z^\alpha_{q_r} / Z^\alpha_\pi) dr} \\ &\approx \beta^\alpha(i\delta) \exp(-\delta c^f_{q_{i\delta},\pi} Z^\alpha_{q_{i\delta}} / Z^\alpha_\pi). \qquad (13)\end{aligned}$$

In the Algorithm 1, To compute $\exp(-\delta c^f_{q_{i\delta},\pi} Z^\alpha_{q_{i\delta}} / Z^\alpha_\pi)$ we use particles $z_{0:i} \sim q_0 \prod_{k=1}^i \overrightarrow{q}_k$ and their weights $w(z_{0:i}) = \prod_{k=1}^i \tilde{q}_{t_k}(z_{k-1}) / \tilde{q}_{t_{k-1}}(z_{k-1})$ given by the AIS algorithm up to iteration i to estimate $Z_{q_{i\delta}} / Z_\pi$ (line 7-9) and the empirical variance under $q_{i\delta}$ (line 10-11). We reduce the variance of the estimated integrand by reusing the same set of particles to perform both estimations for all transitions. Having these estimates, we can approximate $\beta^\alpha((i+1)\delta)$ recursively from approximation of $\beta^\alpha(i\delta)$ in the previous iteration using Equation (13) (line 12).

Using the constant rate schedule estimate, the next annealing density is updated (line 13-14) and particles $(z^j_i)_j$ are transitioned to their new location with a transition kernel which is invariant with respect to $q_{(i+1)\delta}$ (line 16-17). We update the importance weights with standard AIS procedure (line 18) and repeat the process until convergence.

Algorithm 1 CR-AIS tuning for α-divergences

1: **Input:** Target $\tilde{\pi}$, proposal density q_0, α, δ
2: **Output:** Discretization sequence $(\tau_{i\delta})_i$
3: Set $i \leftarrow 0$, $\beta_0 \leftarrow 1$ and $\tau_0 \leftarrow 0$
4: Draw $z^j_0 \sim q_0(z)$ for $j \in [N]$ and concatenate into $\mathbf{z}_i$.
5: Set $\text{logw}^j \leftarrow -\log q_0(z^j_0)$.
6: **while** not converged and $i <$ max_steps **do**
7: Set $\log \hat{Z}_\pi \leftarrow \text{logsumexp}(\text{logw} + \log \tilde{\pi}(\mathbf{z}_i))$.
8: Set $\log \hat{Z}_{q_{i\delta}} \leftarrow \text{logsumexp}(\text{logw} + \log \tilde{q}_{i\delta}(\mathbf{z}_i))$.
9: Set $r_i \leftarrow \exp(\log \hat{Z}_\pi - \log \hat{Z}_{q_{i\delta}})$.
10: Set $u^j_i \leftarrow \frac{\tilde{\pi}(z^j_i)}{r_i \tilde{q}_{i\delta}(z^j_i)}$ for $j \in [N]$ and concat. to $\mathbf{u}_i$.
11: Set $v_i \leftarrow \widehat{\text{Var}}_{q_{i\delta}}[g(\mathbf{u}_i)]$.
12: Set $\beta_{(i+1)\delta} \leftarrow \beta_{i\delta} \exp(-\delta / v_i r^\alpha_i)$.
13: Set $\tau_{(i+1)\delta} = 1 - \beta_{(i+1)\delta}$.
14: Set $\tilde{q}_{(i+1)\delta} = \tilde{q}^{\alpha\text{-pow}}_{\tau_{(i+1)\delta}}$ from Equation (11).
15: Set $i \leftarrow i + 1$.
16: Construct a $\overrightarrow{q}_i(z|z^j_{i-1})$ invariant w.r.t. $q_{i\delta}$.
17: Draw $z^j_i \sim \overrightarrow{q}_i(z|z^j_{i-1})$ for $j \in [N]$.
18: Set $\text{logw}^j \leftarrow \text{logw}^j + \log \frac{\overleftarrow{q}_i(z^j_{i-1}|z^j_i)}{\overrightarrow{q}_i(z^j_i|z^j_{i-1})}$.
19: **end while**
20: Set $\text{logw}^j \leftarrow \text{logw}^j + \log \tilde{\pi}(z^j_i)$.

To ensure stability of the numerical computations we abort the algorithm when the empirical variance becomes lower than a given threshold. This condition indicates that the last annealing density is sufficiently close to the target. However, the empirical variance given by the particles may be much lower than the true value and mislead the algorithm to terminate the annealing process in a handful of steps especially for larger $|\alpha|$. We recommend to mitigate this problem by constraining the maximum step size in the schedule as we did in our experiments with high dimensional targets.

Another consideration is to use disjoint sets of particles for adjusting the schedule and testing. We recommend to perform the final estimation after the tuning phase and fixing the schedule to ensure the independence between samples and the consistency of the importance weighted estimate.

When Equation (8) does not have a closed form solution, we can use CR-AIS with numerical approximation of its solution instead (line 14 of Algorithm 1). New annealing paths may result in more robust estimation in practice as we can optimize function f more effectively. We leave an efficient implementation of this extension to the future work.

5. Experiments

In this section, we run a number of experiments to illustrate the performance of CR-AIS on support coverage and adaptivity with 2d distributions and we asses its efficiency and accuracy with estimation of the log normal-

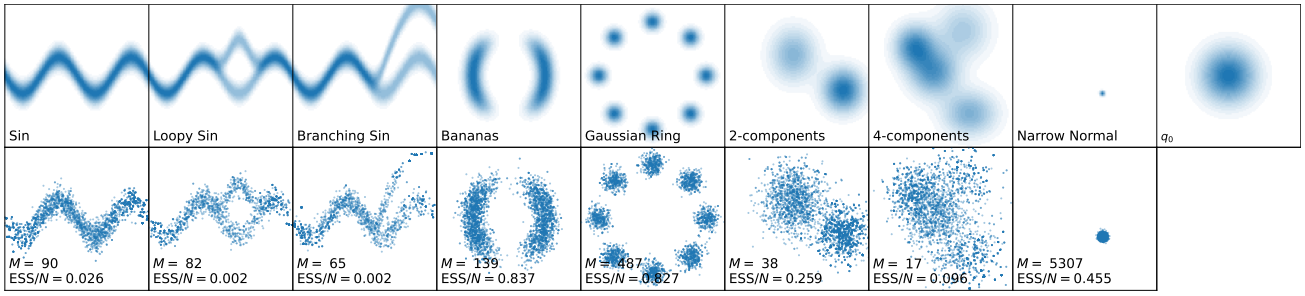

Figure 1. Target distributions and resampled particles with CR-AIS, with geometric mean path and $\delta = 1/32$. CR-AIS adjusts the number of iterations M to the difficulty of the target and covers the support of the targets. The plot in top right corner is the initial distribution used for annealing with the same scale.

ization constant of high dimensional synthetic targets and the posterior of Bayesian models. Code is available at https://github.com/shgoshtasb/cr_ais.

5.1. 2D Distribution Synthesis

Here, we investigate the adaptability of our algorithm to various target distributions. We evaluate CR-AIS on complex 2d distributions which are often used to benchmark sampling (Rezende et al., 2014) and three other distributions. With 2D targets, we can plot the synthesized samples and assess if indeed the particle distribution converges to the targets. We use the inverse KL objective ($\alpha = 0$), initial distribution $\mathcal{N}(0, I)$ and set $\delta = 1/32$ for all the targets as we aim to show the performance differences caused by the difficulty of the sampling task (setup details in Appendix B).

Figure 1 depicts the target distributions next to resampled particles according to the importance weights of the CR-AIS algorithm. The initial distribution is shown in top right corner of Figure 1 with the same scale. Each plot is annotated with the number of AIS iterations, M, and the ESS ratio to number of particles. The algorithm adapts the number of iterations to the difficulty of sampling from the target distribution with the Narrow Gaussian distribution requiring the longest annealing sequence. Additionally, the plots show that particles have reached all the modes of the targets and cover their support. In Appendix D we compare the SMC variant of our algorithm and adaptive SMC on wider multimodal target distributions. Our similar results confirm superiority of the constant rate schedule in mode coverage despite using shorter annealing sequences.

5.2. Adjusting the Schedule to the Annealing Path

In a setup similar to the previous experiment, we investigate how the constant rate schedule adapts across different target distributions and annealing paths. We plot the emerging approximated schedule in CR-AIS for four of the previous targets: the very narrow Gaussian target, Gaussian ring, the dual mode Bananas and the Gaussian mixture distribution with 4 components. We vary α between $\{-0.5, 0, 0.5, 1, 2\}$ to obtain different annealing paths. We compare the schedules to the ones obtained from Adaptive AIS where the schedule is adjusted to decrease CESS at an approximate rate of 0.7. We constrain the maximum step size in Adaptive AIS since large steps cause severe weight degeneracy and premature termination of annealing.

In Figure 2, we show the constant rate schedule $(\tau(i\delta))_i$ emerging from CR-AIS (solid curves) and Adaptive AIS schedules (dashed curves) for different targets and bridging distributions. The constant rate schedule varies considerably between targets and depends explicitly on the similarity of the target distribution and the particle distribution at each time. Consider the second plot from left with $\alpha = 0$. As mentioned before, when majority of particles are in the regions with small π annealing slows down significantly, e.g. in the beginning of annealing for the narrow Gaussian example. It also explains why a much larger number of iterations are required to sample from this target ($M = 5307$) while for the others the number of iterations remain moderate (487, 139 and 17). In contrast, Adaptive AIS has close to linear schedule on the geometric path for 3 of the distributions as CESS depends on weighted averages which may be misleading due to the weight degeneracy problem. For the narrow Gaussian where the target and initial distributions are far apart, Adaptive AIS recovers a schedule similar to CR-AIS.

Across different values of α, the constant rate schedule gradually changes its form depending on the target distribution. For Gaussian mixture, which has overlapping high density regions with the initial distribution, the initial annealing speed reduces monotonically with α. For Gaussian Ring and Bananas, which have modes outside the typical region of q_0, the initial speed grows from $\alpha = 0$ to $\alpha = 0.5$ and decreases for larger α values. Whereas, Adaptive AIS sched-

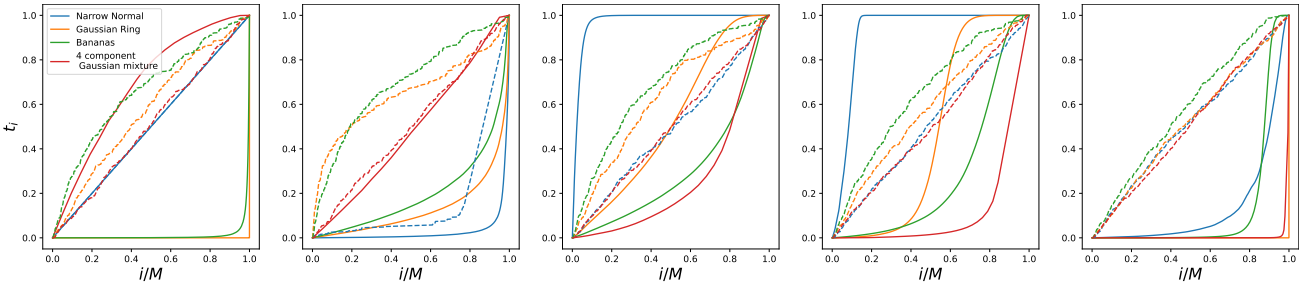

Figure 2. Annealing schedules of CR-AIS (solid curves) and Adaptive AIS (dashed curves) across different annealing paths (α equal to $-0.5, 0, 0.5, 1$ and 2 from left to right) and targets, $\delta = 1/32$. CR-AIS shows higher adaptivity while Adaptive AIS is essentially indifferent to the annealing path.

ule is relatively indifferent to the changes in the annealing path and has to infer the path characteristics through CESS. The high flexibility may lead to stability problems with large variances (e.g. for Gaussian Ring and $\alpha = -0.5$ or $\alpha = 2$) where the schedule grows very slowly and the algorithm reaches the maximum number of iteration, or when the variance is underestimated (e.g. for the Normal distribution with $\alpha = -0.5$) leading to a large step to the target. To avoid these pathologies we recommend to constrain the minimum and maximum of step size as in Adaptive AIS.

5.3. Estimation of $\log Z_\pi$ in High Dimensions

We explore the absolute error of log normalization factor estimation for simple $d = 128$ and 512 dimensional distributions. For the following targets: narrow Gaussian $\mathcal{N}(0, 0.01I)$, a mixture of 8 Gaussian components with variance 1, a standard Laplace distribution and a Student-T distribution with 3 degrees of freedom, we compare CR-AIS with four baselines in Table 2: Adaptive AIS with CESS decrease ratio of 0.6 (Ada. 0.6C), heuristic AIS with linear (Lin.), exponential (Exp.) and sigmoidal (Sigm.) schedules and Monte Carlo Diffusion (MCD) sampler (Doucet et al., 2022a) where the mean and diagonal variance of q_0, the schedule and the transitions are trained for 100 epochs maximizing the evidence lower bound. As CR-AIS and Adaptive AIS generate sequences of varying lengths, for better comparability, we also report a version of the algorithms with interpolated schedules indicated by asterisk (*). See Appendix E for implementation details and further comparison with other adaptive baselines.

The average computation complexity of the sampling algorithms are reported in terms of the number of times $\log \tilde{\pi}$ is evaluated during tuning and testing. This value is proportional to the number of times $\log \tilde{\pi}$ or its gradient are evaluated which are generally the expensive part of the sampling. In Adaptive AIS this corresponds to the number of iterations in the search process of every update to the schedule for every step i during tuning and the final number of discretization steps M for the estimation phase. A parallel measure of complexity counts one schedule update per iteration during tuning in CR-AIS and one schedule update for each pass through the sequence for each annealing step in MCD during training.

MCD has a clear advantage for the Normal target as the parameters of its initial distribution are trained to match the target. However, it's performance drops drastically on the other distributions with insufficient training and it is difficult to justify its expensive training overhead without amortization.

As expected, the performance of the heuristic schedules depend on the target. The exponential schedule is superior for the Normal target in both $d = 128$ and $d = 512$. The linear schedule and the sigmodal schedule are more accurate for the Student-T and the Laplace distributions, respectively, while their ranking changes with d on the Gaussian mixture. Cross validating the heuristic would reduce the efficiency by three folds. On the other hand, at least one of the CR-AIS variations is able to beat the linear, the exponential, and the sigmoidal schedules in 7, 7 and 6 out of the 8 targets with an average overhead of %40 due to tuning with interpolated schedule and %70 without it.

CR-AIS is able to improve over Adaptive AIS while having a higher efficiency. In particular, computation complexity of non-interpolated Adaptive AIS is about $3.5\times$ more than CR-AIS on average, while CR-AIS estimations are more accurate for 5 of the 8 experiments. As the constant rate schedule preserves its form with different δ scales (see e.g. Appendix F) CR-AIS can exploit this property leading to a better performance in comparison to interpolated Adaptive AIS in all 8 of the distributions.

5.4. Bayesian Logistic Regression

In this section, we compare the computation efficiency of CR-AIS to heuristic and Adaptive AIS by evaluating the log marginal likelihood of two Bayesian models. We use two UCI datasets, Pima Indians diabetes dataset ($N = 768$ and $d = 8$) and Sonar dataset ($N = 207$ and $d = 60$) with bi-

Table 2. Absolute $\log Z_\pi$ estimation error for (Top) $d = 128$ and (Bottom) $d = 512$ dimensional distributions with M close to 64 (schedules with * use a shorter sequence for tuning and interpolate the result to $M = 64$). Results are cross validated over different values of α. Smallest error is in bold.

β	NORMAL $\mathcal{N}(0, 0.01I)$		MIXTURE		LAPLACE		STUDENT-T	
	EST. ERR.	COMPUT.	EST. ERR.	COMPUT.	EST. ERR.	COMPUT.	EST. ERR.	COMPUT.
LIN.	991.90 ± 69.87	64.0	230.65 ± 6.16	64.0	0.36 ± 0.22	64.0	1.37 ± 0.19	64.0
SIGM.	907.57 ± 24.04	64.0	270.55 ± 6.42	64.0	0.07 ± 1.14	64.0	1.46 ± 0.14	64.0
EXP.	780.58 ± 40.02	64.0	507.43 ± 15.04	64.0	1.21 ± 0.51	64.0	2.04 ± 0.31	64.0
ADA. 0.6C	853.39 ± 36.87	424.1	268.89 ± 21.24	316.6	0.45 ± 0.75	297.5	1.37 ± 0.24	296.6
ADA. 0.6C*	856.33 ± 29.96	125.8	250.06 ± 6.14	99.6	0.37 ± 0.90	73.4	1.66 ± 0.15	73.8
MCD	**114.18 ± 15.83**	12800.0	680.60 ± 13.68	12800.0	1.01 ± 1.33	12800.0	1.99 ± 1.26	12800.0
CR-AIS (OURS)	788.67 ± 36.25	78.0	308.82 ± 6.81	112.0	0.68 ± 0.31	117.2	**0.69 ± 0.27**	98.4
CR-AIS (OURS)*	807.77 ± 4.36	88.2	**228.54 ± 0.00**	72.0	**0.01 ± 0.00**	72.0	1.53 ± 0.00	73.0
LIN.	5279.55 ± 79.56	64.0	1087.38 ± 22.68	64.0	9.03 ± 0.77	64.0	**8.16 ± 0.45**	64.0
SIGM.	4757.13 ± 59.34	64.0	**1022.60 ± 23.55**	64.0	**6.95 ± 2.48**	64.0	8.71 ± 1.24	64.0
EXP.	4435.05 ± 110.36	64.0	1683.21 ± 13.73	64.0	13.17 ± 0.40	64.0	13.98 ± 1.08	64.0
ADA. 0.6C	5181.95 ± 232.09	503.8	1147.02 ± 142.66	367.2	9.12 ± 3.01	413.8	9.72 ± 2.41	247.7
ADA. 0.6C*	4423.42 ± 204.04	117.4	1272.70 ± 158.61	85.2	8.27 ± 1.33	94.8	9.30 ± 0.56	81.6
MCD	**707.85 ± 110.02**	12800.0	2173.25 ± 43.78	12800.0	16.94 ± 2.51	12800.0	20.95 ± 0.93	12800.0
CR-AIS (OURS)	4413.95 ± 95.75	129.6	1200.12 ± 27.12	140.0	8.75 ± 1.84	110.8	8.40 ± 1.28	98.4
CR-AIS (OURS)*	4546.65 ± 57.80	128.8	1069.32 ± 1.96	75.0	7.93 ± 0.74	78.6	8.98 ± 0.00	73.0

nary labels and a setup similar to (Chopin et al., 2020) with AIS. We consider a Bayesian logistic regression model with normal prior $p(z) = \mathcal{N}(0, 5I)$ and likelihood $p(\mathcal{D}|z) = \prod_n p(y_n|x_n, z)$ for $p(y_n|x_n, z) = \text{Bern}(\sigma(x_n^T z))$. We use CR-AIS to estimate the log marginal likelihood $\log Z = p(\mathcal{D}) = \log \int p(z)p(\mathcal{D}|z)dz$ corresponding to the normalization factor of the posterior distribution of the parameters $\pi(z) = p(z|\mathcal{D}) \propto p(z)p(\mathcal{D}|z)$.

For computation complexity we use a similar measure as described in Section 5.3 and plot the average of estimated $\log Z_\pi$ vs the computation complexity for Pima and Sonar datasets in Figure 3. The estimation of all samplers converges exponentially as the computation budget increases, while CR-AIS has tighter lower bound estimator in comparison to other samplers, especially when computation budget is limited and it roughly requires $\times 4$ fewer $\tilde{\pi}$ evaluations for similar performance as Adaptive AIS.

5.5. Latent Variable Model

We estimate the log marginal likelihood of a Variational AutoEncoder trained on binarized MNIST dataset (Salakhutdinov & Murray, 2008) with a conditional Bernoulli likelihood model. We consider an architecture similar to the one used in (Burda et al., 2016) where the latent variable has $d = 50$ dimensions and both the decoder and the encoder are neural networks with 3 fully-connected layers and we use the same set of hyperparameters as in section 5.3. The estimated lower bound of log marginal likelihood are presented in table 3 for M close to 64 and M close to 512.

With M close to 64, the samplers are far off from the estimation of the variational encoder which is -95.80 nats. The

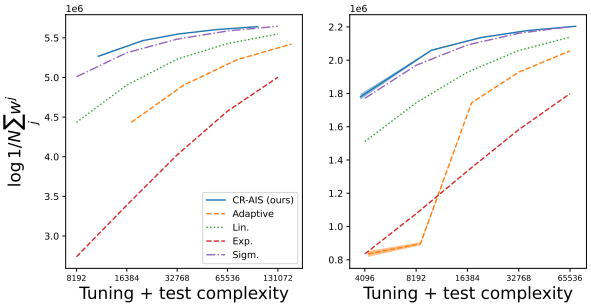

Figure 3. Log marginal likelihood estimates of Bayesian logistic regression model vs computation complexity for CR-AIS, Adaptive AIS with ESS decrease rate of 0.5 on Pima (Left) and Sonar (Right) datasets.

exponential schedule is more accurate than the rest of the samplers. CR-AIS gives a close estimate to it by doubling the computation time and requiring $3\times$ fewer resources in comparison to Adaptive AIS. For the longer sequences with M close to 512, the performance of samplers is harder to distinguish and the gain of tuning diminishes.

6. Related Work and Discussion

Our work is similar to the first order optimization methods used to learn generative probability models. Optimizing functional of probability measures has been studied for decades. Functional gradients are computed in the space of probability distributions endowed with Hilbert structure (Liu & Wang, 2016; Liu, 2017; Dai et al., 2016; Dai, 2018) or Wasserstein structure (Frogner & Poggio, 2020; Lin et al.,

Table 3. Estimated VAE log marginal likelihood with M close to 64 (Top) and M close to 512 (Bottom). Results are cross validated over α. Higher is better.

β	EST. ERR.	M	COMPUT.
LIN.	-141.76 $\pm$ 0.07	64.0	64.0
SIGM.	-130.65 $\pm$ 0.14	64.0	64.0
EXP.	**-121.04** $\pm$ **0.07**	64.0	64.0
ADA. C0.7	-123.04 $\pm$ 3.66	81.6	395.6
CR-AIS (OURS)	-124.63 $\pm$ 0.20	63.4	126.8
LIN.	-106.44 $\pm$ 0.04	512.0	512.0
SIGM.	-104.01 $\pm$ 0.07	512.0	512.0
EXP.	**-102.70** $\pm$ **0.06**	512.0	512.0
ADA. C0.7	-103.26 $\pm$ 0.20	609.7	2560.0
CR-AIS (OURS)	-104.24 $\pm$ 0.11	428.6	857.2

2021). Optimization of deep generative models is generalized as linearization of functional gradients (Chu et al., 2019). However, their application in bridging distributions has not received sufficient attention. These works focus on developing a Wasserstein gradient flow or particle flow to optimally reduce the objective. By contrast our work focuses on optimization of the intermediary unnormalized densities which are required for annealing in particle methods such as Annealed Stein Variational Gradient Descent (D'Angelo & Fortuin, 2021), Parallel Tempering (Earl & Deem, 2005), or Sequential Monte Carlo (Del Moral et al., 2006; Naesseth et al., 2018).

Traditionally, annealing schedule was tuned with ESS or its variants (Jasra et al., 2011; Johansen et al., 2015; Elvira et al., 2018). A variational optimization of the annealing schedule was proposed in (Kiwaki, 2015) with fixed point iteration algorithm to minimize the asymptotic estimation bias/variance. In comparison, CR-AIS uses an analytically interpretable schedule and is able to replace the expensive numerical search or optimization loop with a simple Monte Carlo estimation.

Alternatively, many recent works emerge on combination of AIS and filtering with variational inference (Naesseth et al., 2018; Maddison et al., 2017; Arbel et al., 2021; Thin et al., 2021) or score based generative models (Doucet et al., 2022a;b) to achieve more complex transition kernel and priors and improve posterior approximation. In this line of work, the annealing schedule of a predetermined density path is treated as another set of parameters to optimize using back propagation. This approach results in better amortization for training deep latent variable models and higher log marginal likelihood. End-to-end optimization of the distribution path in AIS has proven to be a more challenging task (Zhang et al., 2021) and is not considered to be effective (Thin et al., 2021; Geffner & Domke, 2021; Goshtasbpour & Perez-Cruz, 2023). Instead, we focus on a greedy optimization approach in terms of the marginal particle distribution divergence with the intended target distribution and we are able to provide a long missing understanding of the reasons underlying the popularity of heuristic annealing paths and demonstrate their limitations due to the greedy nature of the optimization.

7. Conclusion

In this work, we study the connection between the geometric density path in AIS and the functional derivative of inverse KL divergence of marginal particle distribution and the target. We prove that the geometric mean path is the solution to an ODE corresponding to the steepest descent direction of this objective. The analysis can be extended to f-divergences and the ODE has a closed form solution for α-divergences in the form of power mean paths (Brekelmans et al., 2020). We derived constant rate schedule and designed an algorithm that achieves comparable results to the traditional adaptive AIS while avoiding the time consuming search procedure for tuning.

While our theory is motivated by reduction of the immediate bias of the log marginal likelihood estimator, the geometric path is not optimal with respect to the overall sampler bias as it doesn't use information from the possible future steps in the updates. Similarly, power mean path does not translate to an optimal end-to-end statistic of AIS importance weights. However, our optimization method is similar to a continuous time version of Adaptive AIS where instead of searching for the next discretization step size, we optimize the succeeding annealing densities in the function space. As a consequence, we provide a better understanding of the performance of the geometric mean heuristic and demonstrate the underlying reason for its suboptimality. We hope our work helps future research to develop alternative annealing paths with better end-to-end statistics.

Acknowledgements

This work is carried out by Shirin Goshtasbpour while supported by the funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 813999 for this project.

References

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

Brekelmans, R., Masrani, V., Bui, T., Wood, F., Galstyan, A., Steeg, G. V., and Nielsen, F. Annealed Importance Sampling with q-paths. *arXiv preprint arXiv:2012.07823*, 2020.

Burda, Y., Grosse, R. B., and Salakhutdinov, R. Importance weighted autoencoders. In Bengio, Y. and LeCun, Y. (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

Chopin, N., Papaspiliopoulos, O., et al. *An introduction to Sequential Monte Carlo*. Springer, 2020.

Chu, C., Blanchet, J., and Glynn, P. Probability functional descent: A unifying perspective on GANs, variational inference, and reinforcement learning. In *International Conference on Machine Learning*, pp. 1213–1222. PMLR, 2019.

Dai, B. *Learning over functions, distributions and dynamics via stochastic optimization*. PhD thesis, Georgia Institute of Technology, 2018.

Dai, B., He, N., Dai, H., and Song, L. Provable Bayesian inference via particle mirror descent. In *Artificial Intelligence and Statistics*, pp. 985–994. PMLR, 2016.

D'Angelo, F. and Fortuin, V. Annealed Stein variational gradient descent. *arXiv preprint arXiv:2101.09815*, 2021.

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.

Domke, J. and Sheldon, D. R. Importance weighting and variational inference. *Advances in neural information processing systems*, 31, 2018.

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. S., Matthews, A. G. d. G., and Strathmann, H. Annealed importance sampling meets score matching. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022b.

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

Elvira, V., Martino, L., and Robert, C. P. Rethinking the effective sample size. *arXiv preprint arXiv:1809.04129*, 2018.

Frogner, C. and Poggio, T. Approximate inference with wasserstein gradient flows. In *International Conference on Artificial Intelligence and Statistics*, pp. 2581–2590. PMLR, 2020.

Geffner, T. and Domke, J. MCMC variational inference via uncorrected hamiltonian annealing. *Advances in Neural Information Processing Systems*, 34:639–651, 2021.

Gelman, A. and Meng, X.-L. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pp. 163–185, 1998.

Goshtasbpour, S. and Perez-Cruz, F. Optimization of annealed importance sampling hyperparameters. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part V*, pp. 174–190. Springer, 2023.

Grosse, R. B., Maddison, C. J., and Salakhutdinov, R. Annealing between distributions by averaging moments. In *NIPS*, pp. 2769–2777. Citeseer, 2013.

Grosse, R. B., Ghahramani, Z., and Adams, R. P. Sandwiching the marginal likelihood using bidirectional Monte Carlo. *CoRR*, abs/1511.02543, 2015. URL http://arxiv.org/abs/1511.02543.

Grosse, R. B., Ancha, S., and Roy, D. M. Measuring the reliability of MCMC inference with bidirectional Monte Carlo. *Advances in Neural Information Processing Systems*, 29, 2016.

Jasra, A., Stephens, D. A., Doucet, A., and Tsagaris, T. Inference for Lévy-driven stochastic volatility models via adaptive Sequential Monte Carlo. *Scandinavian Journal of Statistics*, 38(1):1–22, 2011.

Johansen, A. M., Aston, J. A., and Zhou, Y. Towards automatic model comparison: An adaptive Sequential Monte Carlo approach. 2015.

Kiwaki, T. Variational optimization of annealing schedules. *arXiv preprint arXiv:1502.05313*, 2015.

Kong, A. A note on importance sampling using standardized weights. *University of Chicago, Dept. of Statistics, Tech. Rep*, 348, 1992.

Lin, A. T., Li, W., Osher, S., and Montúfar, G. Wasserstein proximal of GANs. In *International Conference on Geometric Science of Information*, pp. 524–533. Springer, 2021.

Liu, Q. Stein variational gradient descent as gradient flow. *Advances in neural information processing systems*, 30, 2017.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose Bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

Maddison, C. J., Lawson, J., Tucker, G., Heess, N., Norouzi, M., Mnih, A., Doucet, A., and Teh, Y. Filtering variational objectives. *Advances in Neural Information Processing Systems*, 30, 2017.

Masrani, V., Le, T. A., and Wood, F. The thermodynamic variational objective. *Advances in Neural Information Processing Systems*, 32, 2019.

Masrani, V., Brekelmans, R., Bui, T., Nielsen, F., Galstyan, A., Ver Steeg, G., and Wood, F. q-paths: Generalizing the geometric annealing path using power means. In *Uncertainty in Artificial Intelligence*, pp. 1938–1947. PMLR, 2021.

Naesseth, C., Linderman, S., Ranganath, R., and Blei, D. Variational Sequential Monte Carlo. In *International conference on artificial intelligence and statistics*, pp. 968–977. PMLR, 2018.

Neal, R. M. Sampling from multimodal distributions using tempered transitions. *Statistics and computing*, 6(4):353–366, 1996.

Neal, R. M. Annealed Importance Sampling. *Statistics and computing*, 11(2):125–139, 2001.

Ogata, Y. A Monte Carlo method for high dimensional integration. *Numerische Mathematik*, 55(2):137–157, 1989.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

Salakhutdinov, R. and Murray, I. On the quantitative analysis of deep belief networks. In *Proceedings of the 25th international conference on Machine learning*, pp. 872–879, 2008.

Thin, A., Kotelevskii, N., Doucet, A., Durmus, A., Moulines, E., and Panov, M. Monte Carlo variational auto-encoders. In *International Conference on Machine Learning*, pp. 10247–10257. PMLR, 2021.

Wu, H., Köhler, J., and Noé, F. Stochastic normalizing flows. *arXiv preprint arXiv:2002.06707*, 2020.

Wu, Y., Burda, Y., Salakhutdinov, R., and Grosse, R. On the quantitative analysis of decoder-based generative models. In *International Conference on Learning Representations*, 2017.

Zhang, G., Hsu, K., Li, J., Finn, C., and Grosse, R. B. Differentiable annealed importance sampling and the perils of gradient noise. *Advances in Neural Information Processing Systems*, 34:19398–19410, 2021.

A. Proofs

Here, we provide the proofs of the lemmas and propositions in the paper.

Lemma A.1. *Let $\phi_{t+\epsilon}(z) = \phi_t(z) + \epsilon\eta(z)$ for $\phi_t(z) = \log \tilde{q}_t(z)$ where $\tilde{q}_t(z)$ and $\tilde{\pi}(z)$ are positive unnormalized density functions. Then we have,*

$$\frac{d}{d\epsilon} J_{\mathrm{KL}}[\phi_{t+\epsilon}]\Big|_{\epsilon=0} = \mathrm{Cov}_{q_t}\left[\eta(z), \log \frac{\tilde{q}_t(z)}{\tilde{\pi}(z)}\right], \tag{14}$$

where $\mathrm{Cov}_q[\cdot,\cdot]$ is the covariance under distribution of q and we use the definition of Gâteaux differential for the derivative,

$$\frac{d}{d\epsilon} J_{\mathrm{KL}}[\phi_{t+\epsilon}]\Big|_{\epsilon=0} = \lim_{\epsilon\to 0^+} \frac{J_{\mathrm{KL}}[\phi_t(z) + \epsilon\eta(z)] - J_{\mathrm{KL}}[\phi_t]}{\epsilon}.$$

Proof. We take the derivative as follows,

$$\begin{aligned}
\frac{d}{d\epsilon} J_{\mathrm{KL}}[\phi_{t+\epsilon}]\Big|_{\epsilon=0} =& \frac{\int \tilde{q}_t(z)\eta(z)\log\frac{\tilde{q}_t(z)}{\tilde{\pi}(z)}dz}{\int \tilde{q}_t(z)dz} + \frac{\int \tilde{q}_t(z)\eta(z)dz}{\int \tilde{q}_t(z)dz} \\
&- \frac{\int \tilde{q}_t(z)\log\frac{\tilde{q}_t(z)}{\tilde{\pi}(z)}dz \int \tilde{q}_t(z)\eta(z)dz}{\int \tilde{q}_t(z)dz \int \tilde{q}_t(z)dz} - \frac{\int \tilde{q}_t(z)\eta(z)dz}{\int \tilde{q}_t(z)dz} \\
=& \mathrm{Cov}_{q_t}\left[\eta(z), \log \frac{\tilde{q}_t(z)}{\tilde{\pi}(z)}\right],
\end{aligned}$$

concluding the proof. □

Proposition A.2. *Assume the same conditions as in Lemma 3.1. Additionally, consider the set of smooth perturbation directions with bounded variance*

$$\mathcal{M}_{q_t,\pi} := \{\eta \in \mathcal{C}^1 : \mathrm{Var}_{q_t}[\eta(z)] \le c^{KL}_{q_t,\pi}\},$$

for $B \ge 0$ and $c^{KL}_{q_t,\pi} = B/\mathrm{Var}_{q_t}[\log(\tilde{\pi}(z)/\tilde{q}_t(z))]$. Then the steepest descent direction that minimizes the derivative in Equation (7) in $\mathcal{M}_{q_t,\pi}$ is

$$\eta^*_{q_t,\pi}(z) = \frac{c^{KL}_{q_t,\pi}}{\sqrt{B}} \log \frac{\tilde{\pi}(z)}{\tilde{q}_t(z)} + b, \tag{15}$$

for arbitrary $b \in \mathbb{R}$. A solution to the Ordinary Differential Equation (ODE) $\frac{d}{dt}\phi_t(z) = \eta^*_{q_t,\pi}(z)$ with initial condition $\phi_0(z) = \log \tilde{q}_0(z)$ is the scaled geometric mean path and results in constant rate decrease in the inverse KL divergence in Equation (4).

Proof. It is straight forward to derive the equation for the steepest descent direction using Cauchy-Schwarz inequality. The solution to the ODE is given by

$$\begin{aligned}
\phi_t(z) =& \beta(t)\log\tilde{q}_0(z) + (1-\beta(t))\log\tilde{\pi}(z) \\
&+ \beta(t)\int_0^t \frac{b(r)}{\beta(r)}dr,
\end{aligned} \tag{16}$$

which is equivalent to $\log \tilde{q}^{\mathrm{geom}}_{1-\beta(t)}(z)$ plus a z-independent scale for $t \ge 0$ and for $\beta(t)$ set as

$$\beta^{\mathrm{KL}}(t) := e^{-\int_0^t c^{\mathrm{KL}}_{q_r,\pi} dr/\sqrt{B}}, \tag{17}$$

it leads in constant derivative value form Equation (7). □

Lemma A.3. *Let $\phi_{t+\epsilon(z)} = \phi_t(z) + \epsilon\eta(z)$ for $\phi_t(z) = \log \tilde{q}_t(z)$ where $\tilde{q}_t(z)$ and $\tilde{\pi}(z)$ are positive unnormalized density functions and let $f : \mathbb{R} \to \mathbb{R}$ be convex and differentiable. Then we have,*

$$\frac{d}{d\epsilon} J_f[\phi_{t+\epsilon}]\Big|_{\epsilon=0} = \text{Cov}_{q_t}\left[\eta(z), -g(u_t(z))\right],$$

where $g(u) = u\dot{f}(u) - f(u)$ and $\dot{f}(u) = df(u)/du$. Moreover, consider the set of smooth perturbation directions with bounded variance $\mathcal{M}^f_{q_t,\pi} := \left\{\eta \in \mathcal{C}^1 : \text{Var}_{q_t}[\eta(z)] \le c^f_{q_t,\pi}\right\}$ for $B \ge 0$ and

$$c^f_{q_t,\pi} = B/\text{Var}_{q_t}\left[g(u_t(z))\right].$$

Then the steepest descent direction that minimizes this derivative in $\mathcal{M}^f_{q_t,\pi}$ is

$$\eta^*_{q_t,\pi}(z) = \frac{c^f_{q_t,\pi}}{\sqrt{B}} g(u_t(z)) + b,$$

for arbitrary $b \in \mathbb{R}$.

Proof. With the perturbed negative energy

$$\phi_{t+\epsilon}(z) = \log q_{t+\epsilon}(z) = \phi_t(z) + \epsilon\eta(z)$$

in direction η, $u_t(z)$ is updated to $u_t(z)e^{-\epsilon\eta(z)}\,\mathbb{E}_{q_t}[e^{\epsilon\eta(z)}]$. Consequently,

$$\begin{aligned}
\frac{d}{d\epsilon} J_f[\phi_{t+\epsilon}]\Big|_{\epsilon=0} =& \frac{\int \tilde{q}_t(z)\eta(z) f\left(u_t(z)\right) dz}{\int \tilde{q}_t(z)dz} + \frac{\int \tilde{q}_t(z)\frac{d}{du} f\left(u_t(z)\right) u_t(z)\,\mathbb{E}_{q_t}[\eta(z)]dz}{\int \tilde{q}_t(z)dz} \\
&- \frac{\int \tilde{q}_t(z)\frac{d}{du} f\left(u_t(z)\right) u_t(z)\eta(z)dz}{\int \tilde{q}_t(z)dz} - \frac{\int \tilde{q}_t(z) f\left(u_t(z)\right) dz}{\int \tilde{q}_t(z)dz} \cdot \frac{\int \tilde{q}_t(z)\eta(z)dz}{\int \tilde{q}_t(z)dz} \\
=& \text{Cov}_{q_t}\left[\eta(z), -g(u_t(z))\right].
\end{aligned}$$

The rest of the proof follows due to Cauchy-Schwarz inequality. □

Proposition A.4. *Assume the same conditions as in Lemma 3.3 and $f(u) = (u^\alpha - 1 - \alpha(u-1))/\alpha(\alpha-1)$ for $\alpha \notin \{0,1\}$ or $f(u) = u\log u$ for $\alpha = 1$. Then α-power mean path is the solution to the ODE $\frac{d}{dt}\phi_t(z) = \eta^*_{q_t,\pi}(z)$ with initial condition $\phi_0(z) = \log \tilde{q}_0(z)$ and with a particular schedule it results in constant rate decrease in f-divergence.*

Proof. We prove the proposition for $\alpha \notin \{0,1\}$ in the following which can be easly extended to the case with $\alpha = 1$. Using the derivative of f we have $g(u) = (u^\alpha - 1)/\alpha$, therefore,

$$\eta^*_{q_t,\pi}(z) = \frac{c^f_{q_t,\pi}}{\alpha\sqrt{B}}(u_t^\alpha - 1) + b(t).$$

In the ODE

$$\begin{aligned}
\frac{d}{dt}\log \tilde{q}(z) &= \eta^*_{q_t,\pi}(z) + b(t) \\
&= \frac{c^f_{q_t,\pi}}{\alpha\sqrt{B}}\left(\left(\frac{\tilde{\pi}(z)}{\tilde{q}_t(z)}\frac{Z_{q_t}}{Z_\pi}\right)^\alpha - 1\right) + b(t).
\end{aligned} \tag{18}$$

We use the change of variable $\tilde{s}_t(z) = \tilde{q}_t(z)^\alpha$ and rewrite the ODE as

$$\frac{d}{dt}\tilde{s}_t(z) + \left(\frac{c^f_{q_t,\pi}}{\sqrt{B}} - \alpha b(t)\right)\tilde{s}_t(z) = \frac{c^f_{q_t,\pi} Z^\alpha_{q_t}\tilde{\pi}(z)^\alpha}{\sqrt{B} Z^\alpha_\pi},$$

which has a solution of the form

$$s_t(z) = \beta(t)\left[\tilde{\pi}(z)^\alpha \int_0^t \frac{c^f_{q_r,\pi} Z^\alpha_{q_r}}{\sqrt{B} Z^\alpha_\pi \beta(r)} dr + \tilde{q}_0(z)^\alpha\right]$$

for

$$\beta(t) = e^{\int_0^t (\alpha b(r) - c^f_{q_r,\pi}/\sqrt{B}) dr}.$$

Therefore, without loss of generality, we choose

$$b(t) = -c^f_{q_t,\pi}((Z_{q_t}/Z_\pi)^\alpha - 1)/\alpha\sqrt{B}$$

and get the annealing sequence $\tilde{q}^{\alpha\text{-pow}}_{\tau(t)}(z)$ for the schedule $\tau(t) = 1 - \beta(t)$ and with $\beta(t)$ set to

$$\beta^\alpha(t) := e^{-\int_0^t (c^f_{q_r,\pi} Z^\alpha_{q_r}/\sqrt{B} Z^\alpha_\pi) dr}, \tag{19}$$

The α-divergence decreases with a steady rate along the steepest descent direction. □

B. 2D Distributions Implementation Details

Here we give the implementation details of our experiments on 2d benchmark dataset. We initialize CR-AIS with a standard normal distribution for q_0 which is plotted in the top right corner of Figure 1 for scale and we use the same value of $\delta = 1/32$ for all the targets. Each AIS transition is a single Hamiltonian Monte Carlo (HMC) step with step size 0.5 and $N = 1024$ particles are used to approximate the constant rate schedule. We abort sampling when the empirical variance $\widehat{\text{Var}}_{q_{i\delta}}[g(u)]$ is below 0.001. We use $\alpha = 0$ for the results in Section 5.1.

In Section 5.2, for Adaptive AIS, we tune the schedule on each iteration using binary search with constrained maximum step size set to $1/128$ avoid large steps. This value is chosen to give the search algorithm sufficient flexibility to effect the schedule while larger values would result in big steps and premature annealing.

C. Accuracy of the Approximations in CR-AIS

To assess the accuracy of the approximations used in the CR-AIS algorithm, we depict the Monte Carlo estimation of the objective during a run of the algorithm for the same setup as Section 5.2 in Figure 4. The curves indicate an almost static decrease of the inverse KL divergence and the efficiency of the approximations in CR-AIS with combination of self-normalized normalization factor ratio estimation and Riemann sum.

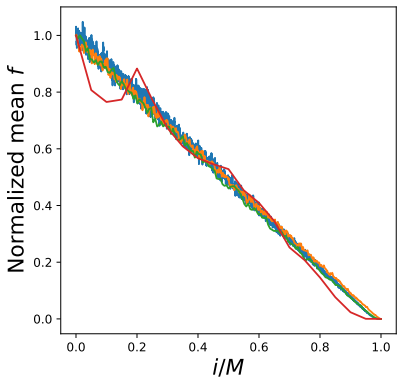

Figure 4. Mean objective vs normalized AIS iterations in four distributions used in Section 5.2, with geometric mean path and $\delta = 1/32$. Mean f is normalized by the initial divergence to fit in the same plot. Best seen in color.

D. Constant Rate Sequential Monte Carlo Experiments

In this section, we adapt Sequential Monte Carlo (SMC) algorithm using the constant rate schedule and compare the performance of sampling to Adaptive SMC. SMC is a modification of AIS with occasional resampling steps to prevent weight degeneracy problem in sequential importance sampling algorithms. In practice, the resampling is performed periodically or adaptively whenever ESS falls below a threshold, the latter providing more sample diversity.

To adapt the schedule in SMC we used weighted estimates of log normalization factor ratio with the SMC importance weights (line 7-9). The rest of the algorithm is similar to Algorithm 1 with the additional branch for resampling after line 18 to duplicate the effective samples. We perform resampling adaptively when ESS is less than 0.9 of the last time resampling was performed.

We construct 4 randomly generated wide Gaussian mixtures uniformly sampling the mean of the components from $[-15, -15]$ and compare the synthesized particles to Adaptive SMC where the schedule is determined with binary search on each iteration to achieve constant rate decrease of 0.9 in ESS. Figure 5 shows the sampled particles from CR-SMC and Adaptive AIS starting from standard normal distribution with similar HMC transitions. Adapative SMC particles tend to group together in a few of the targets' modes, while with CR-SMC the final samples have higher diversity and reach all the target modes in three of the four distributions. This happens despite the fact that CR-SMC uses shorter annealing sequences in all of the distributions.

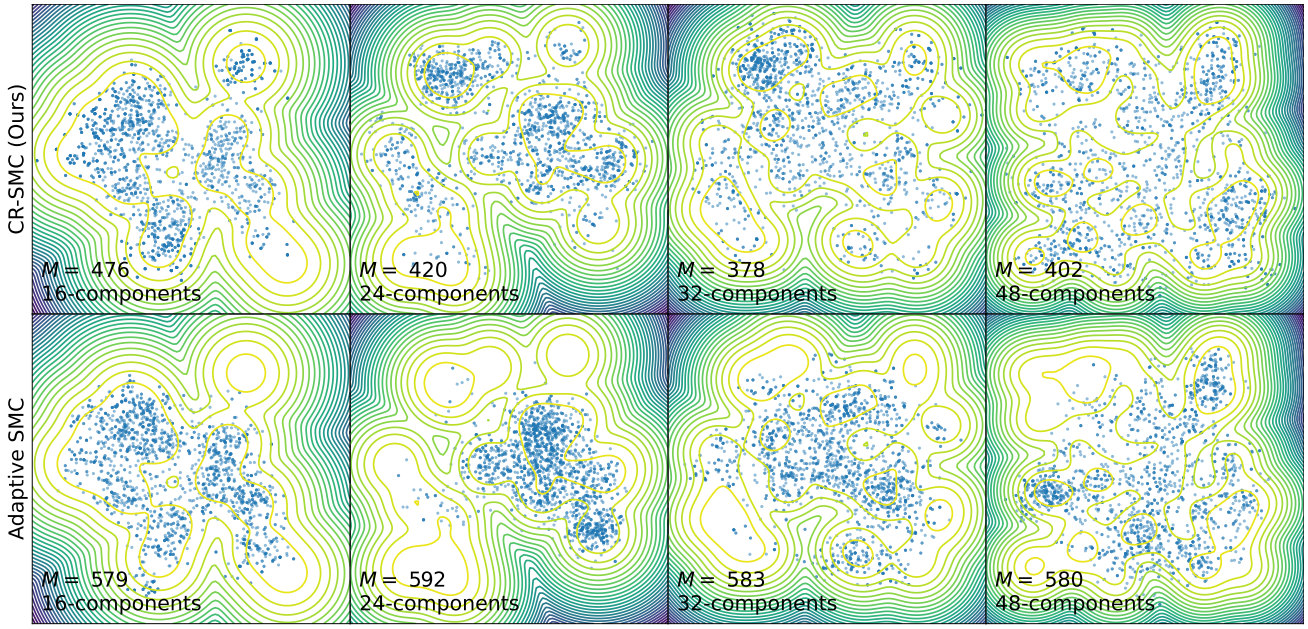

Figure 5. Contour lines of four wide mixture distributions in [-20,20] in each column with samples (blue discs) from CR-SMC (top) and Adaptive SMC (bottom). geometric mean path with $\delta = 1/128$ selected to have roughly the same number of iterations as Adaptive SMC which has worse coverage.

In Table 4 we report the $\log Z_\pi$ estimations with both algorithms on the four targets.

Table 4. Estimated $\log Z$ of four gaussian mixtures with the same order as Figure 5 with 16,24,32 and 48 components using CR-SMC and Adaptive SMC with 0.9 CESS drop resampling and 0.9 ESS drop schedule adaptation. Lower estimation error is in bold.

	TARGET 1	TARGET 2	TARGET 3	TARGET 4
$\log Z$	5.997	6.402	6.690	7.095
ADA 0.9	5.387	**6.285**	5.891	6.808
CR-SMC (OURS)	**5.967**	6.125	**6.634**	**6.974**

E. High dimensional Experiments

We use the same hyperparameters as in Appendix B for the log normalization factor estimation experiments with the following modifications. We set $N = 4096$ and chose α from a grid in $\{-0.5, 0.0, ..., 2.\}$, δ from $\{1/256, ..., 4096\}$ for CR-AIS depending on the target such that sampling budget M will be $M \leq 64$ for interpolation and close to $M = 64$ without interpolation. To generate the results with interpolated schedules, we interpolate the tuned schedules with each δ to match the test-time computation of all samplers to $M = 64$ and sample from the target distribution. We report the $\log 1/N \sum w(z^j_{0:M})$ with smallest estimation error over α and δ. In order to avoid the large steps in Adaptive AIS when ESS is underestimating the variance, we choose a maximum step size in the adaptive AIS from $\{1/512, ..., 1/4\}$ using cross validation in the same manner as CR-AIS. Each algorithm is run with 5 different seeds and the average absolute estimation error is reported.

Below we present more thorough comparison of CR-AIS with adaptive and heuristic AIS benchmarks. The details of the experiment are similar to the setup in Sectionsec:experiment3 with additional variations of Adaptive AIS algorithm. Specifically, in Adaptive AIS, we tune the schedule on each iteration using binary search with constrained maximum step size set to avoid large steps and apply proposed steps which decrease Effective Sample Size (ESS) and Conditional ESS (CESS) (Johansen et al., 2015) constantly with ratio in $\{0.5, 0.6, 0.7, 0.8, 0.9\}$.

Table 5. Absolute $\log Z_\pi$ estimation error for 128-dimensional distributions: Normal $\mathcal{N}(0, 0.01I)$, Gaussian mixtures with 8 components of variance 1, standard Laplace and Student-T distribution with 3 degrees of freedom with M close to 64. Results are cross validated over different values of α. Smallest error is in bold.

β	Normal		Mixture		Laplace		Student-T	
	Est. err.	M	Est. err.	M	Est. err.	M	Est. err.	M
Ada. 0.9c	871.75 ± 25.81	69.8	246.08 ± 33.20	69.2	0.80 ± 0.65	43.2	0.96 ± 0.66	50.0
Ada. 0.6c	853.39 ± 36.87	61.8	268.89 ± 21.24	61.4	0.45 ± 0.75	39.6	1.37 ± 0.24	43.2
Ada. 0.7c	864.52 ± 31.07	61.6	253.38 ± 39.73	70.0	0.12 ± 0.55	69.8	1.13 ± 0.39	43.6
Ada. 0.8c	879.29 ± 22.60	63.2	252.01 ± 33.43	69.0	0.74 ± 0.97	40.8	0.96 ± 1.75	11.4
Ada. 0.9	992.07 ± 33.45	66.4	347.07 ± 29.13	43.0	**0.04 ± 0.77**	44.8	1.00 ± 1.25	42.0
Ada. 0.6	1001.14 ± 72.16	65.6	384.80 ± 24.05	36.8	0.72 ± 0.50	37.2	0.92 ± 0.83	37.2
Ada. 0.7	986.59 ± 53.15	65.8	378.61 ± 23.42	38.0	0.54 ± 0.62	39.4	1.56 ± 0.29	41.2
Ada. 0.8	987.17 ± 29.88	66.8	357.53 ± 22.14	40.6	0.15 ± 0.70	42.0	1.66 ± 0.14	39.8
Lin.	991.90 ± 69.87	64.0	**230.65 ± 6.16**	64.0	0.36 ± 0.22	64.0	1.37 ± 0.19	64.0
Sigm.	907.57 ± 24.04	64.0	270.55 ± 6.42	64.0	0.07 ± 1.14	64.0	1.46 ± 0.14	64.0
Exp.	**780.58 ± 40.02**	64.0	507.43 ± 15.04	64.0	1.21 ± 0.51	64.0	2.04 ± 0.31	64.0
CR-AIS (Ours)	788.67 ± 36.25	39.0	308.82 ± 6.81	56.0	0.68 ± 0.31	58.6	**0.69 ± 0.27**	49.2

Table 6. Absolute $\log Z_\pi$ estimation error for 512-dimensional distributions: Normal $\mathcal{N}(0, 0.01I)$, Gaussian mixtures with 8 components of variance 1, standard Laplace and Student-T distribution with 3 degrees of freedom with M close to 64. Results are cross validated over different values of α. Smallest error is in bold.

β	Normal		Mixture		Laplace		Student-T	
	Est. err.	M	Est. err.	M	Est. err.	M	Est. err.	M
Ada. 0.9c	5163.14 ± 217.21	54.8	1147.02 ± 136.34	65.6	8.30 ± 1.34	67.4	10.39 ± 2.15	40.8
Ada. 0.6c	5181.95 ± 232.09	53.4	1147.02 ± 142.66	66.4	9.12 ± 3.01	66.4	9.72 ± 2.41	37.2
Ada. 0.7c	5297.43 ± 188.64	56.4	1170.21 ± 104.98	66.4	7.01 ± 3.31	64.0	10.66 ± 0.83	39.2
Ada. 0.8c	5195.95 ± 105.26	55.8	1119.56 ± 79.00	69.6	8.27 ± 3.26	64.4	9.92 ± 1.65	38.6
Ada. 0.9	5107.72 ± 142.41	66.0	1662.46 ± 94.87	39.0	9.64 ± 2.18	40.8	9.97 ± 1.01	40.0
Ada. 0.6	5203.06 ± 115.42	65.2	1033.74 ± 42.08	69.2	7.14 ± 2.52	69.8	**7.83 ± 1.26**	69.4
Ada. 0.7	5190.67 ± 112.92	65.2	1095.46 ± 30.55	70.0	10.01 ± 3.15	36.2	10.15 ± 1.36	36.4
Ada. 0.8	5151.53 ± 87.72	65.6	1266.31 ± 67.74	69.8	9.65 ± 1.79	38.6	9.94 ± 1.54	39.4
Lin.	5279.55 ± 79.56	64.0	1087.38 ± 22.68	64.0	9.03 ± 0.77	64.0	8.16 ± 0.45	64.0
Sigm.	4757.13 ± 59.34	64.0	**1022.60 ± 23.55**	64.0	**6.95 ± 2.48**	64.0	8.71 ± 1.24	64.0
Exp.	4435.05 ± 110.36	64.0	1683.21 ± 13.73	64.0	13.17 ± 0.40	64.0	13.98 ± 1.08	64.0
CR-AIS (Ours)	**4413.95 ± 95.75**	64.8	1200.12 ± 27.12	70.0	8.75 ± 1.84	55.4	8.40 ± 1.28	49.2

F. Bayesian Logistic Regression Details

For the Bayesian logistic regression experiment we have the following hyperparameters. We vary δ between $\{64, ..., 2048\}$ for CR-AIS and maximum step size in the range $\{1/65536, ..., 1/2048\}$ for Adaptive AIS with $N = 256$ particles. We chose α from a grid in $[-0.5, 2.0]$ for each value of δ and maximum step size separately using cross validation on the highest estimated log marginal likelihood.

We use standard normal distribution for q_0 and 1-step HMC transitions as before with step size 0.5. Although the transition kernels are considerably simple for a problem of moderate dimensions, it is chosen to compare the adaptability of the algorithms to the posterior distribution. In fact, an annealing sequence with larger M and simple HMC transitions although has the same amount of computations, is more flexible than a shorter annealing sequence with larger number of MCMC steps per transition.

We use Adaptive AIS with ESS decrease rate of 0.5.

The adapted CR-AIS schedule for different values of δ is illustrated in Figure 6(Right) and Figure 7(Right) for Pima and Sonar datasets. As expected, the discretization schedule has a similar pattern for different values of δ. It is possible to exploit this property principally to increase the computation efficiency by interpolating the schedule obtained from tuning with a large δ and elongating the annealing sequence to reach the desired M for the final estimation with negligible impact on the performance.

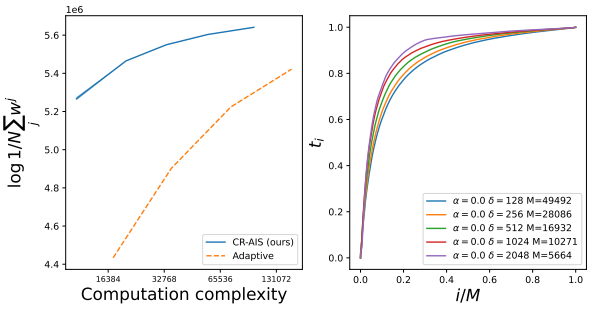

Figure 6. (Left) Log marginal likelihood estimates vs computation complexity for CR-AIS and Adaptive AIS. (Right) CR-AIS discretization schedule for different values of δ on Bayesian logistic regression model of Pima dataset.

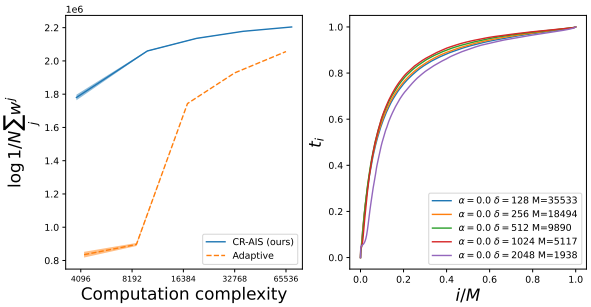

Figure 7. (Left) Log marginal likelihood estimates vs computation complexity for CR-AIS and Adaptive AIS. (Right) CR-AIS discretization schedule for different values of δ on Bayesian logistic regression model of Sonar dataset.