

Non-asymptotic Analysis of Diffusion Annealed Langevin Monte Carlo for Generative Modelling

Paula Cordero-Encinar¹, Ö. Deniz Akyildiz¹, and Andrew B. Duncan^{1, 2}

¹Imperial College London

²The Alan Turing Institute

February 14, 2025

Abstract

We investigate the theoretical properties of general diffusion (interpolation) paths and their Langevin Monte Carlo implementation, referred to as diffusion annealed Langevin Monte Carlo (DALMC), under weak conditions on the data distribution. Specifically, we analyse and provide non-asymptotic error bounds for the annealed Langevin dynamics where the path of distributions is defined as Gaussian convolutions of the data distribution as in diffusion models. We then extend our results to recently proposed heavy-tailed (Student's t) diffusion paths, demonstrating their theoretical properties for heavy-tailed data distributions for the first time. Our analysis provides theoretical guarantees for a class of score-based generative models that interpolate between a simple distribution (Gaussian or Student's t) and the data distribution in *finite time*. This approach offers a broader perspective compared to standard score-based diffusion approaches, which are typically based on a forward Ornstein-Uhlenbeck (OU) noising process.

1 Introduction

Score-based generative models (SGMs) [Ho et al., 2020; Song et al., 2021] have become immensely popular in recent years due to their excellent performance in generating high-quality data. This success has led to widespread adoption across various generative modelling tasks, e.g., image generation [Dhariwal and Nichol, 2021; Rombach et al., 2022; Saharia et al., 2022], audio generation [Ruan et al., 2023], reward maximisation [He et al., 2023; Janner et al., 2022]. Additionally, their remarkable performance has sparked significant interest within the theoretical community to better understand the structure and properties of these models [Benton et al., 2024; Chen et al., 2022, 2023; Lee et al., 2022].

The goal of generative modelling is to learn the underlying probability distribution π_{data} from a given set of samples. Diffusion models, a particular class of SGMs, achieve this by using a *forward process*, typically an OU process, to construct a path of probability distributions from the data distribution towards a simpler one – a Gaussian. The time-reversed process can be characterised [Anderson, 1982] but necessitates the knowledge of the scores of the marginal distributions along this path. These scores are usually intractable - hence they are learnt by noising the data and applying score matching techniques [Hyvärinen and Dayan, 2005; Song et al., 2020; Vincent, 2011]. The learnt scores are then used to sample from the path by discretising the time-reversed diffusion process [Song et al., 2021].

While the forward OU process is mathematically convenient, it does not capture the whole idea of bridging distributions and requires infinite time to interpolate between the data distribution π_{data} and a Gaussian measure. In practice, however, diffusion models consider the evolution of the OU process only up to a finite final time T . Thus, the path does not fully bridge π_{data} and a standard Gaussian. During generation, these models instead evolve samples along a sequence of interpolated distributions

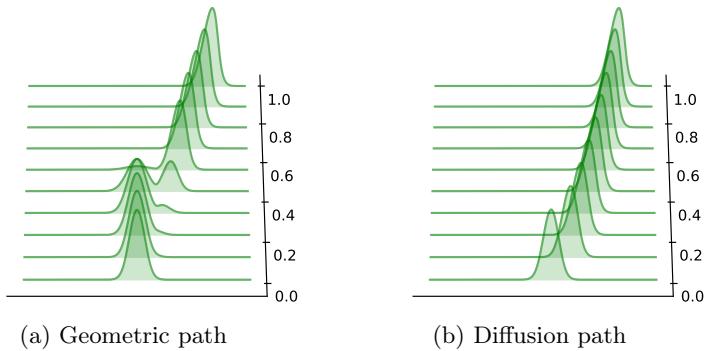


Figure 1: A visual comparison of the geometric path versus the diffusion path for $(\mu_t)_{t \in [0,1]}$. The base distribution is given by $\mu_0 := \mathcal{N}(0, 1)$ and the data distribution, $\mu_1 := \pi_{\text{data}}$, is a mixture of a Gaussian and a smoothed uniform distribution (see Section 3). As observed by Chehab et al. [2025], the geometric path in (a) creates intermediate multimodal distributions which are hard to sample from. In contrast, the diffusion path in (b) stays unimodal throughout, offering more favourable properties.

between the final marginal distribution of the OU process at time T and π_{data} (although, in practice, they are initialised from a Gaussian). Specifically, this interpolation is characterised by defining intermediate random variables $X_t \sim \mu_t$ ¹ [Chehab and Korba, 2024] as

$$X_t = \sqrt{\lambda_t} X + \sqrt{1 - \lambda_t} Z, \quad (1)$$

for $t \in [0, T]$, where $X \sim \pi_{\text{data}}$, $Z \sim \mathcal{N}(0, I)$ is independent of X and a schedule $\lambda_t = \min\{1, e^{-2(T-t)}\}$.

The interpolation perspective of diffusion models has been investigated, see, e.g., Albergo et al. [2023]; Gao et al. [2024]. Notably, the path in Eq. (1) is a special case of the *one-sided stochastic interpolants* [Albergo et al., 2023]. As outlined in these works, the reverse process can be made to exactly interpolate between a base distribution ν and π_{data} in finite time by using an appropriate schedule λ_t and introducing control terms in the corresponding stochastic differential equations (SDEs). Similarly to the score term in diffusion models, these control terms are intractable and need to be learnt.

In this work, we adopt a practical approach to general linear interpolation paths between a simple base distribution ν and π_{data} , that is, $X_t = \sqrt{\lambda_t} X + \sqrt{1 - \lambda_t} Z$, where $X \sim \pi_{\text{data}}$, $Z \sim \nu$ independent of X and $\lambda_t \in [0, 1]$, $\lambda_T = 1$. In particular, we explore the behaviour of Langevin dynamics driven by the gradients of $\log \mu_t$ for $t \in [0, T]$, where μ_t are the intermediate distributions, i.e., $X_t \sim \mu_t$. Our approach is akin to earlier generative modelling methods based on *annealed Langevin dynamics* [Song and Ermon, 2019] which led to the development of diffusion models. However, there has been limited work analysing these methods under minimal assumptions on π_{data} . Block et al. [2022] provide the first theoretical analysis in Wasserstein distance under smoothness and dissipativity of the data distribution. They show that the error depends exponentially on the dimension. In contrast, Lee et al. [2022] provides a non-asymptotic bound in total variation under smoothness conditions and a bounded log-Sobolev constant of the data distribution. Specifically, we make the following contributions.

Contributions

- We provide an analysis of annealed Langevin dynamics methods driven by general linear interpolation paths between ν and π_{data} , which we term diffusion annealed Langevin Monte Carlo (DALMC). In the case where ν is a Gaussian distribution, we derive non-asymptotic convergence bounds in Kullback-Leibler (KL) divergence under different assumptions.

¹In our case, the base (simple) distribution is defined at time 0 as μ_0 , and the data distribution is defined at time T , $\mu_T = \pi_{\text{data}}$. This contrasts with standard diffusion models where the base distribution is defined at time T and the data distribution at time 0.

By assuming that π_{data} has a finite second-order moment M_2 , $\log \pi_{\text{data}}$ has Lipschitz gradients and either π_{data} is strongly convex outside a ball or $\nabla^2 \log \pi_{\text{data}}$ decays to 0 sufficiently fast (as is the case for Student's t -like distributions), we show in Corollary 3.5 that DALMC requires $\mathcal{O}\left(\frac{d(M_2 \vee d)^2 L_{\max}^2}{\varepsilon^6}\right)$ steps to achieve ε^2 -accurate sampling from π_{data} in KL divergence with a sufficiently accurate score estimator. Here, d is the dimension of the data and $L_{\max} := \max_{t \in [0, T]} L_t$ where L_t denotes the Lipschitz constant of $\nabla \log \mu_t$, which we prove to be finite in Lemma 3.2, improving the results of Gao et al. [2024, Proposition 20] under the specified conditions. Furthermore, under slightly less restrictive assumptions involving smoothness of π_{data} with constant L_π , bounded second order moment M_2 and $\mathbb{E}_{\pi_{\text{data}}} \|\nabla \log \pi_{\text{data}}(Y)\|^8 \leq K_\pi^2$, we demonstrate that the data distribution can be approximated to ε^2 -accuracy in KL divergence with $\mathcal{O}\left(\frac{(M_2 \vee d)^2 (d^2 \vee L_\pi^2) d \vee K_\pi) L_\pi}{\varepsilon^6}\right)$ steps. To the best of our knowledge, these are the first results obtained in KL divergence for these Langevin-dynamics driven generative models [Song and Ermon, 2019].

- We then extend this analysis into recent heavy-tailed diffusion models [Pandey et al., 2025] based on Student's t noising distributions, that is, when the base distribution ν is chosen to be a Student's t distribution. In this case, assuming that the data distribution is smooth, has a finite second-order moment and exhibits a tail behaviour similar to that of a multivariate Student's t distribution, we show that DALMC can be used to sample from the data distribution with the same complexity as the Gaussian case. As far as we are aware, this is the first analysis of heavy-tailed diffusion models with explicit complexity estimates.
- We show that, under certain conditions on the covariances, a mixture of Gaussians with different covariances satisfy smoothness conditions and is strongly log-concave outside of a ball, implying a finite log-Sobolev constant. This result is of independent interest, as most analyses of Gaussian mixtures in the literature primarily focus on the equal covariance setting.

The rest of the paper is organised as follows. Section 2 presents our setting and necessary background. Section 3, provides a non-asymptotic analysis of the general diffusion paths with Gaussian base distribution and their implementation via Langevin dynamics. In Section 4, we extend our analysis to heavy-tailed diffusion models. Section 5 discusses related literature, followed by the conclusion.

Notation Let d be the dimension of data. Let A, B be square matrices of the same dimension, we say $A \preccurlyeq B$ if $B - A$ is a positive semidefinite matrix and $\|\cdot\|_F$ denotes the Frobenius norm. For $a, b > 0$, we write $a \lesssim b$ or $a = \mathcal{O}(b)$ to indicate that $a \leq Cb$ for an absolute constant $C \geq 0$, and $a \asymp b$ if $a = \mathcal{O}(b)$ and $b = \mathcal{O}(a)$. For $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ and a probability measure μ on \mathbb{R}^d , we define $\|f\|_{L^2(\mu)} := (\int \|f\|^2 d\mu)^{1/2}$ and $M_2 := \mathbb{E}_{\pi_{\text{data}}} [\|X\|^2]$.

2 Generative Modelling via Diffusion Paths

We present the background and setting for our analysis.

2.1 Diffusion Paths

In practice, implementing the reverse process in diffusion models consists in sampling along a path of probability distributions $(\mu_t)_{t \in [0, T]}$, which starts at a simple distribution μ_0 and ends at an arbitrarily complex data distribution $\mu_T = \pi_{\text{data}}$. In particular, when the forward process is an OU process and evolves the data distribution for time T , the starting distribution of the reversed process takes the form $e^{dT} \pi_{\text{data}}(e^T x) * \mathcal{N}(0, (1 - e^{-2T})I)$ and the interpolated distributions $(\mu_t)_t$ are the marginals of the OU process. Building on this, we can describe a more general version of the probability distribution paths that diffusion models attempt to sample from [Chehab and Korba, 2024], as

$$\mu_t(x) = \frac{\pi_{\text{data}}(x/\sqrt{\lambda_t})}{\lambda_t^{d/2}} * \frac{\nu(x/\sqrt{1-\lambda_t})}{(1-\lambda_t)^{d/2}}, \quad (2)$$

where $*$ denotes the convolution operation, ν describes the base or *noising* distribution, and λ_t is an increasing function called schedule, such that, $\lambda_t \in [0, 1]$ and $\lambda_T = 1$. We refer to the probability path $(\mu_t)_{t \in [0, T]}$ in (2) as the *diffusion path*. In the setting of the OU (i.e. variance preserving) process, λ_t corresponds to $\lambda_t = \min\{1, e^{-2(T-t)}\}$.

The diffusion path with the OU schedule has demonstrated very good performance in the generative modelling literature and has recently started to be explored for sampling [Huang et al., 2024; Richter and Berner, 2024; Vargas et al., 2024]. For instance, Phillips et al. [2024] empirically observed that the diffusion path may have a more favourable geometry for the Langevin sampler than the geometric path, obtained by taking the geometric mean of the base and target distributions, as is typically done in annealing due to the tractability of the score (Figure 1).

While successful, the use of the OU process presents some challenges in practice. As mentioned earlier, the forward OU process cannot reach ν in finite time, meaning that, in theory the reversed path starts from a non-Gaussian distribution μ_0 . However, in practice, the paths are initialised from Gaussians, introducing a bias that is present in error bounds [Benton et al., 2024; Chen et al., 2022, 2023]. In our setting, by selecting an appropriate schedule for the diffusion path (2), which satisfies $\lambda_0 = 0$ and $\lambda_T = 1$, the path of probability distributions $(\mu_t)_{t \in [0, T]}$ can interpolate exactly between $\mu_0 = \nu$ and $\mu_T = \pi_{\text{data}}$ in finite time, unlike the OU process. This formulation is equivalent to that of linear one-sided stochastic interpolants which can also be realised through SDEs [Albergo et al., 2023, Theorem 5.3].

We will next explore an alternative approach for generative modelling with general linear diffusion paths, namely, running annealed Langevin dynamics on paths $(\mu_t)_{t \in [0, T]}$ that are constructed to meet the correct marginals.

2.2 Annealed Langevin Dynamics for Diffusion Paths

For general diffusion paths, the “reverse process” cannot be described by a closed form SDE. While Albergo et al. [2023], estimate the intractable drift term of the SDE using neural networks, their approach can experience numerical instabilities at $t = T$ (see Albergo et al. [2023, Section 6]) due to singularities in the drift term. Therefore, in this work, we focus on *annealed Langevin dynamics* [Song and Ermon, 2019] to explicitly implement a sampler along the diffusion path, avoiding the extra control terms introduced in Albergo et al. [2023]. Note that the score at each time t can be learnt via score matching techniques, as in Song and Ermon [2019].

Our annealed Langevin dynamics consists of running a time-inhomogeneous Langevin SDE, where the drifts are given by the scores of reparametrised probability distributions from the diffusion path $(\hat{\mu}_t = \mu_{\kappa t})_{t \in [0, T/\kappa]}$, for some $0 < \kappa < 1$. That is, we will use the following SDE

$$dX_t = \nabla \log \hat{\mu}_t(X_t) dt + \sqrt{2} dB_t \quad t \in [0, T/\kappa], \quad (3)$$

where $X_0 \sim \mu_0 = \nu$ and $(B_t)_{t \geq 0}$ is a Brownian motion. We refer to (3) as *diffusion annealed Langevin dynamics*. This strategy does provide a viable alternative to implement interpolation paths as the scores can be learnt. In particular, we consider the diffusion annealed Langevin Monte Carlo (DALMC) algorithm given by a simple Euler-Maruyama discretisation of (3) and the use of a score approximation function $s_\theta(x, t)$ [Song and Ermon, 2019]:

$$X_{l+1} = X_l + h_l s_\theta(X_l, t_l) + \sqrt{2h_l} \xi_l, \quad (4)$$

where $h_l > 0$ is the step size, $\xi_k \sim \mathcal{N}(0, I)$, $s_\theta(x, t)$ approximates $\nabla \log \hat{\mu}_t(x)$, $l \in \{1, \dots, M\}$ and $0 = t_0 < \dots < t_M = T/\kappa$ is a discretisation of the interval $[0, T/\kappa]$.

It is important to note that, even if simulated exactly, diffusion annealed Langevin dynamics introduces a bias, as the marginal distributions of the solution of the SDE (3) do not exactly correspond to $(\hat{\mu}_t)_t$, unlike in the stochastic interpolants formulation [Albergo et al., 2023]. One of the contributions of our work will be to quantify this bias non-asymptotically. A key component in determining the effectiveness of the diffusion annealed Langevin dynamics will be the action of the curve of probability measures $\mu = (\mu_t)_{t \in [0, T]}$ interpolating between the base distribution and the data distribution, denoted by $\mathcal{A}(\mu)$. As noted by Guo et al. [2025], the action serves as a measure of the cost of transporting ν to π_{data} along

the given path. Formally, the action of an absolutely continuous curve of probability measures [Lisini, 2007] with finite second-order moment is defined as follows

$$\mathcal{A}(\mu) := \int_0^T \lim_{\delta \rightarrow 0} \frac{W_2(\mu_{t+\delta}, \mu_t)}{|\delta|}.$$

Based on Theorem 1 from Guo et al. [2025], we have that the KL divergence between the path measure of the diffusion annealed Langevin dynamics (3), $\mathbb{P}_{\text{DALD}} = (p_{t,\text{DALD}})_{t \in [0,T/\kappa]}$, and that of a reference SDE such that the marginals at each time have distribution $\hat{\mu}_t$, $\mathbb{P} = (\hat{\mu}_t)_{t \in [0,T/\kappa]}$, can be bounded in terms of the action. In particular, when $p_0 = p_{0,\text{DALD}}$, it follows from Girsanov's theorem that

$$\text{KL}(\mathbb{P} || \mathbb{P}_{\text{DALD}}) \leq \kappa \mathcal{A}(\mu).$$

See Theorem A.3 in Appendix A for the proof and further details. Note that by the data processing inequality, we have that $\text{KL}(\pi_{\text{data}} || p_{T/\kappa, \text{DALD}}) \leq \text{KL}(\mathbb{P} || \mathbb{P}_{\text{DALD}})$, meaning that the KL divergence between the data distribution and the final marginal distribution of the diffusion annealed Langevin dynamics (3) is bounded, provided that the action is finite. In that case by choosing $\kappa = \mathcal{O}(\varepsilon^2 / \mathcal{A}(\mu))$, we ensure that $\text{KL}(\pi_{\text{data}} || p_{T/\kappa, \text{DALD}}) \lesssim \varepsilon^2$.

2.3 Initial Assumptions

In what follows, we will provide an in-depth analysis of the DALMC algorithm when the base distribution ν is Gaussian or multivariate Student's t distribution. The latter relates to recent heavy-tailed diffusion models [Pandey et al., 2025]. Our results in both cases are based on the following assumptions, with additional ones introduced later as necessary.

First, as is typical in the diffusion model literature we require an L^2 accurate score estimator [Chen et al., 2022, 2023].

A 1. *The score approximation function $s_\theta(x, t)$ satisfies*

$$\sum_{l=0}^{M-1} h_l \mathbb{E}_{\hat{\mu}_t} \left[\left\| \nabla \log \hat{\mu}_l(X_{t_l}) - s_\theta(X_{t_l}, t_l) \right\|^2 \right] \leq \varepsilon_{\text{score}}^2.$$

where $0 = t_0 < t_1 < \dots < t_M = T/\kappa$ is a discretisation of the interval $[0, T/\kappa]$.

A 2. *The data distribution π_{data} has a finite second-order moment, that is, $M_2 = \mathbb{E}_{\pi_{\text{data}}}[\|X\|^2] < \infty$.*

3 Gaussian Diffusion Paths

In this section, we focus on analysing algorithms to simulate the diffusion path $(\mu_t)_{t \in [0,T]}$ defined in (2) when the base distribution ν is Gaussian, $\nu \sim \mathcal{N}(m_\nu, \sigma^2 I)$. For simplicity, we will assume that π_{data} has mean 0 and set $m_\nu = 0$. This diffusion path has the remarkable property, illustrated in Figure 1, that when π_{data} has finite log-Sobolev and Poincaré constants, these constants remain uniformly bounded along the entire path, as summarised in the following result.

Proposition 3.1. *If π_{data} has a finite log-Sobolev constant $C_{LSI}(\pi_{\text{data}})$, respectively Poincaré constant $C_{PI}(\pi_{\text{data}})$, the Gaussian diffusion path $(\mu_t)_{t \in [0,T]}$ defined in (2) with base distribution $\nu \sim \mathcal{N}(0, \sigma^2 I)$ satisfies for all $t \in [0, T]$*

$$C_{LSI}(\mu_t) \leq \lambda_t C_{LSI}(\pi_{\text{data}}) + (1 - \lambda_t) C_{LSI}(\nu), \\ C_{PI}(\mu_t) \leq \lambda_t C_{PI}(\pi_{\text{data}}) + (1 - \lambda_t) C_{PI}(\nu),$$

respectively, where $C_{LSI}(\nu) = C_{PI}(\nu) = \sigma^2$.

The proof follows immediately from Chewi [2024, Propositions 2.3.3 and 2.3.7]. This result is highly favourable, as, unlike geometric annealing [Chehab et al., 2025], the log-Sobolev and Poincaré constants remain uniformly bounded along the entire path by the worst constant independently of the distance between π_{data} and ν . We can visually observe this in Figure 1, when the data distribution is given by a mixture of a Gaussian and a smoothed uniform distribution, $\pi_{\text{data}} = (1 - e^{-m^2/4})\mathcal{N}(m, 1) + e^{-m^2/4}u_m$, where u_m is the smoothed uniform distribution on $I_m = [-m, 2m]$ for $m = 10$ [Chehab et al., 2025].

Motivated by this, we analyse the diffusion annealed Langevin dynamics (3) to simulate from (2).

3.1 Analysis of the Gaussian Diffusion Path

We start by analysing the properties of the *Gaussian diffusion path* $(\mu_t)_{t \in [0, T]}$.

Smoothness of $(\mu_t)_t$. We consider the following assumption on the Lipschitz continuity of the scores $\nabla \log \mu_t$.

A 3. *For all $t \in [0, T]$, the scores of the intermediate distributions $\nabla \log \mu_t(x)$ are Lipschitz with finite constant L_t .*

A2 and **A3** are sufficient for one of our non-asymptotic analyses of the Gaussian diffusion path (Theorem 3.4). However, **A3** is generally difficult to verify. Therefore, we introduce two *alternative* assumptions, **A4** and **A5**, which separately ensure that $(\nabla \log \mu_t)_{t \in [0, T]}$ satisfies assumption **A3**. In particular, we show that **A4** is satisfied by a mixture of Gaussians with different covariances, given certain conditions on the covariances. While assumption **A5** is shown to hold for heavy-tailed data distributions.

A 4. *The data distribution π_{data} has density with respect to Lebesgue, which we write $\pi_{\text{data}} \propto e^{-V_\pi}$. The potential V_π has Lipschitz continuous gradients, with Lipschitz constant L_π . In addition, V_π is strongly convex outside of a ball of radius r with convexity parameter $M_\pi > 0$, that is,*

$$\inf_{\|x\| \geq r} \nabla^2 V_\pi \succcurlyeq M_\pi I, \quad \inf_{\|x\| < r} \nabla^2 V_\pi \succcurlyeq -L_\pi I.$$

In Lemma B.1 of the Appendix, we demonstrate that assumption **A4** extends the standard assumption on the data distribution that π_{data} is modelled as a convolution of a compactly supported distribution $\tilde{\pi}$ and a Gaussian, see, e.g., Saremi et al. [2024, Theorem 1] or Grenioux et al. [2024, Assumption 0], under some conditions on the compact support of $\tilde{\pi}$. Additionally, we prove that a mixture of Gaussians with different covariances satisfies assumption **A4** under some mild conditions on the covariances (see Lemma B.4 and Remark B.5 for a further discussion). However, Lemma B.6 shows that, in general, a mixture of Gaussians cannot be expressed as a convolution of a compactly supported measure with a Gaussian. We consider the results regarding the mixture of Gaussians to hold independent significance, as we could not find explicit results in the literature addressing the smoothness properties in this case.

Leveraging the existence of a smooth strongly convex approximation of V_π [Ma et al., 2019] and the Holley-Stroock perturbation lemma [Holley and Stroock, 1987], we show that under **A4**, π_{data} satisfies a log-Sobolev inequality with a finite constant which itself implies a finite Poincaré constant (Lemma B.7) – which is sufficient for Proposition 3.1 to hold. As a consequence, **A4** implies that the data distribution π_{data} has finite second order moment (i.e. **A4** \Rightarrow **A2**).

On the other hand, heavy-tailed data distributions, such as Student's t -like distributions, do not satisfy assumption **A4**, since their potential V_π is not strongly convex outside of a ball. Specifically, the Hessian of the potential tends to zero as $\|x\|$ tends to infinity. We provide the following alternative assumption for heavy-tailed data distributions.

A 5. *The data distribution π_{data} has density with respect to Lebesgue, which we write $\pi \propto e^{-V_\pi}$. The potential V_π has Lipschitz continuous gradients, with Lipschitz constant L_π . In addition, $\nabla^2 V_\pi(x)$ decays to 0 with order $\mathcal{O}(\|x\|^{-2}I)$ as $\|x\|$ tends to ∞ . That is, outside of a ball of radius r we have that*

$$-\frac{I}{\alpha_1 + \alpha_2 \|x\|^2} \preccurlyeq \nabla^2 V_\pi(x) \preccurlyeq \frac{I}{\beta_1 + \beta_2 \|x\|^2} \quad \|x\| > r,$$

where $\alpha_1, \alpha_2, \beta_1, \beta_2 \in \mathbb{R}$.

In Appendix B.2 we show that multivariate Student's t distributions of the form

$$\pi_{\text{data}}(x) \propto \left(1 + \frac{1}{\alpha}(x - \mu)^T \Sigma^{-1}(x - \mu)\right)^{-(\alpha+d)/2}$$

satisfy assumption **A5**.

The following Lemma establishes that assumption **A3** holds when the data distribution satisfies either **A4** or **A5**.

Lemma 3.2. *Under **A4** or **A5**, we have that for all $t \in [0, T]$ the score $\nabla \log \mu_t(x)$ is Lipschitz continuous with constant L_t provided in the proof (i.e. **A4** \Rightarrow **A3** and **A5** \Rightarrow **A3**).*

An important element in the proof, given in Appendix B.3, is the generalisation of the Poincaré inequality for vector-valued functions, which is presented in Lemma B.8. Notably, these bounds improve those in Gao et al. [2024, Proposition 20] under the specified conditions.

It is important to emphasise that a significant number of works in the diffusion models literature, e.g. Chen et al. [2022]; Lee et al. [2022]; Chen et al. [2023], assume that $\nabla \log \mu_t$ is Lipschitz for all t , with the Lipschitz constant bounded over time. In contrast, we have demonstrated that this condition arises naturally under assumptions **A4** or **A5** on the target distribution.

Action of $(\mu_t)_t$. To derive a bound on the action necessary for the convergence analysis, we make the following assumption on the schedule.

A6. *Let $\lambda_t : \mathbb{R}^+ \rightarrow [0, 1]$ be non-decreasing in t and weakly differentiable, such that there exists a constant C_λ satisfying either of the following conditions*

$$\max_{t \in [0, T]} |\partial_t \log \lambda_t| \leq C_\lambda$$

or

$$\max_{t \in [0, T]} \left| \frac{\partial_t \lambda_t}{\sqrt{\lambda_t(1 - \lambda_t)}} \right| \leq C_\lambda.$$

Notably schedules of the form $\lambda_t = 0.5(1 + \cos(\pi(1 - (t/T)^\phi)))$, $0.5(1 + \tanh(\phi(t/T - 0.5)))$ with $\phi \in \mathbb{R}^+$, sigmoid-type schedules, or the schedule corresponding to the OU process $e^{-2(T-t)}$, among others, satisfy the previous assumption. When $\lambda_0 = 0$, the first condition in **A6** requires that the derivative of the schedule at $t = 0$ is close to zero, meaning that the schedule grows very slowly at the beginning. This intuitively captures the importance of the initial stages in the Langevin diffusion generation process. For instance, when the data distribution consists of two distant modes, the diffusion needs to allocate the correct proportion of mass to each mode. During the early stages, as the mass separates towards each mode, employing a slower-increasing schedule can aid in this process. As the mass approaches each mode, the probability of it jumping between modes decreases rapidly, making a slow initial increase essential for effective separation. The second condition ensures that the schedule also becomes flat as it approaches $\lambda_T = 1$. This promotes a more refined and detailed generation process, enabling the model to converge more precisely to the data distribution.

Under assumption **A6** on the schedule, we derive the following bound on the action of $\mu = (\mu_t)_{t \in [0, T]}$.

Lemma 3.3. *If π_{data} and λ_t satisfy assumptions **A2** and **A6**, respectively, the action for the Gaussian diffusion path $\mathcal{A}_\lambda(\mu)$ can be upper bounded by*

$$\mathcal{A}_\lambda(\mu) \lesssim C_\lambda (\mathbb{E}_{\pi_{\text{data}}} [\|X\|^2] + d) \lesssim M_2 \vee d.$$

The proof is given in Appendix B.4. It is worth highlighting that unlike for the geometric path [Guo et al., 2025], for the diffusion path we get an explicit bound on the action under a mild assumption on the schedule. Furthermore, we observe in the proof that selecting the mean and variance of the base distribution ν close to that of the target results in a tighter bound for the action.

3.2 Analysis of the Gaussian DALMC Algorithm

We now analyse the convergence of the DALMC algorithm (4) for a Gaussian base distribution.

Theorem 3.4. *Under **A2**, **A3** and **A6**, the DALMC algorithm (4) initialised at $X_0 \sim \hat{\mu}_0$ and with an approximate score which satisfies **A1**, yields the following bound*

$$\begin{aligned} \text{KL}(\mathbb{P} \parallel \mathbb{Q}_\theta) &\lesssim \left(1 + \frac{L_{\max}^2}{M^2 \kappa^4}\right) \kappa (\mathbb{E}_{\pi_{\text{data}}} [\|X\|^2] + d) \\ &\quad + \frac{d}{M \kappa^2} \left(1 + \frac{L_{\max}}{M \kappa}\right) \int_0^T L_t^2 dt + \varepsilon_{\text{score}}^2, \end{aligned}$$

where $\mathbb{Q}_\theta = (q_{\theta, \lambda_t})_{t \in [0, T]}$ is the path measure of the continuous-time interpolation of (4), \mathbb{P} is that of a reference SDE such that the marginals at each time t have distribution $\hat{\mu}_t$, M denotes the number of steps, $T/\kappa = \sum_{l=1}^M h_l$ and L_t is the Lipschitz constant of $\nabla \log \mu_t$, $L_{\max} = \max_{[0, T]} L_t$.

The proof, included in Appendix B.5, mainly relies on an application of Girsanov's theorem, the bound on the action and the Lipschitzness of $\nabla \log \mu_t$. Additionally, we note in the proof that, smaller step sizes h_l are preferred when the Lipschitz constant L_t is larger to obtain a tighter bound.

This result allows us to provide a bound on the iteration complexity of the DALMC algorithm.

Corollary 3.5. *For $T \geq 1$, $\kappa < 1$ and M , there always exists a sequence of step sizes $h_k = T_k - T_{k-1}$ such that $\sum_{k=1}^M h_k = T/\kappa$. Then, if we take $\kappa = \mathcal{O}\left(\frac{\varepsilon_{\text{score}}^2}{M_2 \vee d}\right)$ and $M = \mathcal{O}\left(\frac{d(M_2 \vee d)^2 L_{\max}^2}{\varepsilon_{\text{score}}^6}\right)$, we have $\text{KL}(\mathbb{P} \parallel \mathbb{Q}_\theta) = \mathcal{O}(\varepsilon_{\text{score}}^2)$. Hence, for any $\varepsilon = \mathcal{O}(\varepsilon_{\text{score}})$, and under assumptions **A2**, **A3** and **A6**, the DALMC algorithm (4) initialised at $X_0 \sim \hat{\mu}_0$ requires at most $\mathcal{O}\left(\frac{d(M_2 \vee d)^2 L_{\max}^2}{\varepsilon^6}\right)$ steps to approximate π_{data} to within ε^2 KL divergence, that is $\text{KL}(\pi_{\text{data}} \parallel q_{\theta, \lambda_T}) \leq \varepsilon^2$, assuming a sufficiently accurate score estimator.*

It is important to note that these bounds are less favourable than those of diffusion models [Benton et al., 2024; Chen et al., 2022, 2023], which explains the success of these models compared to diffusion annealed Langevin-based algorithms. This difference mainly arises because the Langevin SDE implementation (3) introduces an implicit bias, whereas the reverse SDE in diffusion models ensures that the law of the solution of the SDE exactly matches the intermediate marginal distributions. Additionally, the use of an exponential integrator scheme in diffusion models, benefiting from the linear term in the drift of the reverse SDE, contrasts with the Euler-Maruyama discretisation used here, leading to an improvement in the discretisation error.

3.3 Analysis under Relaxed Assumptions

In this section, we introduce a less restrictive assumption for the data distribution that generalises **A4** and **A5**. Under this assumption, we derive an error bound for the DALMC algorithm without relying on the smoothness of $\log \mu_t$ along the diffusion path, in contrast to the proof of Theorem 3.4.

A7. *The data distribution π_{data} has density with respect to Lebesgue, which we write $\pi_{\text{data}} \propto e^{-V_\pi}$, and a finite second order moment. The potential V_π has Lipschitz continuous gradient, with Lipschitz constant L_π , and*

$$\mathbb{E}_{\pi_{\text{data}}} \|\nabla V_\pi(X)\|^8 \leq K_\pi^2.$$

In Appendix B.7, we show that both **A4** and **A5** (with finite second-order moment) imply assumption **A7**. Besides, since under **A7** the data distribution has a finite second-order moment, if the schedule also satisfies **A6**, then the bound on the action established in Lemma 3.3 remains valid. This enables us to obtain the following complexity guarantees for the DALMC algorithm under this new assumption.

Theorem 3.6. Under **A 6** and **A 7**, the DALMC algorithm (4) initialised at $X_0 \sim \hat{\mu}_0$ and with an approximate score which satisfies **A 1** leads to

$$\text{KL}(\mathbb{P} \parallel \mathbb{Q}_\theta) \lesssim \frac{dL_\pi}{M^2\kappa^3} + \frac{(d^2 \vee L_\pi^2 d \vee K_\pi)}{M\kappa^2} \\ + \kappa(\mathbb{E}_{\pi_{\text{data}}}[\|X\|^2] + d) + \varepsilon_{\text{score}}^2,$$

where $\mathbb{Q}_\theta = (q_{\theta, \lambda_t})_{t \in [0, T]}$ is the path measure corresponding to the continuous-time interpolation of algorithm (4), \mathbb{P} is that of a reference SDE such that the marginals at each time t have distribution $\hat{\mu}_t$ and M denotes the number of steps. Therefore, under assumptions **A 6** and **A 7**, the DALMC algorithm (4) initialised at $X_0 \sim \hat{\mu}_0$ with approximate scores requires at most $\mathcal{O}\left(\frac{(M_2 \vee d)^2(d^2 \vee L_\pi^2 d \vee K_\pi)L_\pi}{\varepsilon^6}\right)$ steps to approximate π_{data} to within ε^2 KL divergence, that is $\text{KL}(\pi_{\text{data}} \parallel q_{\theta, \lambda_T}) \leq \varepsilon^2$, assuming a sufficiently accurate score estimator, i.e. $\varepsilon_{\text{score}} = \mathcal{O}(\varepsilon)$. If $M_2 = \mathcal{O}(d)$, $L_\pi = \mathcal{O}(\sqrt{d})$ and $K_\pi = \mathcal{O}(d^2)$, then $M = \mathcal{O}\left(\frac{d^4 L_\pi}{\varepsilon^6}\right)$.

See Appendix B.8 for the proof. Relaxing the assumptions results in a dimensional dependence on the number of steps that is one order worse compared to Corollary 3.5.

4 Heavy-Tailed Diffusion Paths

We now analyse the annealed Langevin diffusion path (3) when the base distribution $\nu \in \mathcal{P}(\mathbb{R}^d)$ is a Student's t -distribution, $\nu \sim t(0, \sigma^2 I, \alpha)$, with tail index $\alpha > 2$

$$\nu(x) \propto \left(1 + \frac{\|x\|^2}{\alpha\sigma^2}\right)^{-(\alpha+d)/2}.$$

It is worth noting that the t -distribution is not a stable distribution, unlike the Gaussian family, meaning that the convolution of two t -distributions is not necessarily a t -distribution. Nadarajah and Dey [2005] provides explicit expressions for the density function of the convolution of one dimensional t -distributions with unit variance, but only when both degrees of freedom are odd. Closed-form expressions cannot be derived in one dimension when either of the two degrees of freedom is even. In the d -dimensional case, only closed forms can be derived when $\alpha + d$ is even.

4.1 Analysis of the Heavy-Tailed Diffusion Path

Smoothness of $(\mu_t)_t$. We require for the discretisation analysis that the intermediate distributions of the heavy-tailed diffusion path satisfy smoothness conditions given in **A3**. We show below that this assumption holds when the data distribution π_{data} satisfies the following conditions.

A 8. The data distribution π_{data} has density with respect to the Lebesgue measure. $\nabla \log \pi_{\text{data}}$ is Lipschitz continuous with constant L_π and $\|\nabla \log \pi_{\text{data}}\|^2 \leq C_\pi$ almost surely.

In particular, Lemma 4.1 below demonstrates that this assumption holds when the data distribution π_{data} can be expressed as the convolution of a compactly supported measure and a t -distribution.

A 9. Let X be a d -dimensional random vector $X \sim \pi_{\text{data}}$, such that $X = U + G$, where $\|U - m_\pi\|^2 \leq dR^2$ holds almost surely, and $G \sim t(0, \tau^2 I, \tilde{\alpha})$ is independent of U .

Lemma C.1 in the Appendix shows that if π_{data} satisfies assumption **A9**, then it has a finite weighted Poincaré constant. This extends the result of Bardet et al. [2018] to convolutions of compactly supported measures with t -distributions. However, unlike the multivariate Gaussian case, our bound on the weighted Poincaré constant is not dimension-free.

The following result shows that **A8** \Rightarrow **A3** and **A9** \Rightarrow **A8**.

Lemma 4.1. Under **A8** and taking the base distribution $\nu \sim t(0, \sigma^2 I, \alpha)$, we have that for all $t \in [0, T]$ $\nabla \log \mu_t(x)$ is Lipschitz (**A3**) with constant L_t provided in the proof. Besides, **A8** holds when π_{data} is a convolution of a compactly supported measure and a multivariate t distribution (**A9**).

The proof is provided in Appendix C.2.

Action of $(\mu_t)_t$. To derive a bound on the action, necessary for the discretisation analysis, we introduce an assumption on the schedule similar to that of **A6**.

A 10. Let $\lambda_t : \mathbb{R}^+ \rightarrow [0, 1]$ be non-decreasing in t and weakly differentiable, such that there exists a constant C_λ satisfying

$$\max_{t \in [0, T]} \left| \frac{\partial_t \lambda_t}{\sqrt{\lambda_t(1 - \lambda_t)}} \right| \leq C_\lambda.$$

Intuitively, schedules with derivatives close to 0 as t approaches 0 and T satisfy the previous assumption. In particular, schedules of the form $\lambda_t = 0.5(1 + \cos(\pi(1 - (t/T)^\phi)))$, $0.5(1 + \tanh(\phi(t - 0.5)))$, where $\phi \in \mathbb{R}^+$, fulfil **A10**. Under the previous assumption, we compute the following bound on the action. The proof is provided in Appendix C.3.

Lemma 4.2. If π_{data} has finite second-order moment (**A2**), $\nu \sim t(0, \sigma^2 I, \alpha)$ with tail index $\alpha > 2$ and λ_t satisfies **A10**, the action for the heavy-tailed diffusion path $\mathcal{A}_\lambda(\mu)$ can be effectively upper bounded as follows

$$\mathcal{A}_\lambda(\mu) \leq \frac{C_\lambda \pi}{8} \left(\mathbb{E}_{\pi_{\text{data}}} [\|X\|^2] + \frac{\sigma^2 d \alpha}{\alpha - 2} \right).$$

4.2 Analysis of the Heavy-Tailed DALMC Algorithm

The following theorem establishes a bound for the discretisation error of the heavy-tailed DALMC algorithm with an approximated score. The proof is given in Appendix C.4.

Theorem 4.3. Assume the data distribution π_{data} satisfies assumption **A2** (finite second-order moment) and **A3** (which holds under **A8**), $\nu \sim t(0, \sigma^2 I, \alpha)$ with $\alpha > 2$ and let the schedule satisfy **A10**, with $\lambda_{\kappa t}/\lambda_{\kappa(t+\delta)} = \mathcal{O}(1 + \delta)$, $\delta \ll 1$. Then, the heavy-tailed DALMC algorithm with an approximated score satisfying **A1** and initialised at $X_0 \sim \hat{\mu}_0$, guarantees that

$$\begin{aligned} \text{KL}(\mathbb{P} \parallel \mathbb{Q}_\theta) &\lesssim \left(1 + \frac{L_{\max}^2}{M^2 \kappa^2} + \frac{1}{M^2 \kappa^4} \right) \kappa (\mathbb{E}_{\pi_{\text{data}}} [\|X\|^2] + d) \\ &\quad + \frac{d}{M \kappa^2} \left(1 + \frac{\alpha}{\alpha - 2} + \frac{L_{\max}}{M \kappa} \right) \int_0^T L_t^2 dt + \varepsilon_{\text{score}}^2, \end{aligned}$$

where $\mathbb{Q}_\theta = (q_{\theta, \lambda_t})_{t \in [0, T]}$ is the path measure of the continuous-time interpolation of (4), \mathbb{P} is that of a reference SDE such that the marginals at each time t have distribution $\hat{\mu}_t$, M denotes the number of steps, $T/\kappa = \sum_{l=1}^M h_l$ and L_t is the Lipschitz constant of $\nabla \log \mu_t$, $L_{\max} = \max_{[0, T]} L_t$. Therefore, under **A2**, **A3** and **A10**, the heavy-tailed DALMC algorithm (4) initialised at $X_0 \sim \hat{\mu}_0$ requires at most $\mathcal{O}\left(\frac{d(M_2 \vee d)^2 L_{\max}^2}{\varepsilon^6}\right)$ steps to approximate π_{data} to within ε^2 KL divergence, that is $\text{KL}(\pi_{\text{data}} \| q_{\theta, \lambda_T}) \leq \varepsilon^2$, assuming a sufficiently accurate score estimator, i.e. $\varepsilon_{\text{score}} = \mathcal{O}(\varepsilon)$.

Note that as the tail index α tends to ∞ , which corresponds to ν approaching a Gaussian distribution, the bound on $\text{KL}(\mathbb{P} \parallel \mathbb{Q}_\theta)$ recovers that of Theorem 3.4. Furthermore, since $\alpha/(\alpha - 2) \leq 3$ for $\alpha > 2$, the iteration complexity of the heavy-tailed DALMC algorithm is identical to that of the Gaussian DALMC algorithm corresponding to the Gaussian diffusion path.

5 Related Work

Score-based generative models. Our approach is similar to earlier generative modelling techniques based on annealed Langevin dynamics [Song and Ermon, 2019], which inspired the advancement of

diffusion models. The existing literature analysing these Langevin Monte Carlo algorithms is limited. Block et al. [2022] derive an error bound in Wasserstein distance that scales exponentially with the data dimension, while Lee et al. [2022] establish a non-asymptotic bound in total variation, which is weaker than our bound in KL, as implied by Pinsker’s inequality.

On the other hand, the convergence of diffusion models [Ho et al., 2020; Song et al., 2021] has been extensively studied. Early results either established non quantitative bounds [Pidstrigach, 2022], relied on restrictive assumptions about the data distribution, such as functional inequalities [Lee et al., 2022], or exhibited exponential dependence on the problem parameters [De Bortoli, 2022]. Recent works have established polynomial convergence bounds under more relaxed assumptions [Benton et al., 2024; Chen et al., 2022, 2023; Li et al., 2024]. In particular, Chen et al. [2022] introduce two bounds on the KL error: a linear bound in the data dimension under smoothness conditions along the entire diffusion path, and a second one which scales quadratically with d , achieved through early stopping and the assumption of a finite second-order moment on π_{data} . In contrast, Benton et al. [2024] provide a bound that is linear in the data dimension, up to logarithmic factors, assuming only that the data distribution has a finite second-order moment. Their proof exploits the specific structure of the OU process to control the error arising from discretising the reverse SDE.

Stochastic interpolants. Stochastic interpolants [Albergo et al., 2023] are generative models that unify flow-based and diffusion-based methods. These models make use of a broad class of continuous-time stochastic processes designed to bridge any two arbitrary probability density functions exactly in finite time, akin to our work. Specifically, the formulation of linear one-sided stochastic interpolants [Albergo et al., 2023; Gao et al., 2024], which interpolate between a Gaussian and the data distribution, is equivalent to the Gaussian diffusion path (2). Unlike our approach, they incorporate intractable control terms into the drift of the SDE to ensure the marginals have the desired distributions. This may result in numerical instabilities caused by singularities in the drift at $t = T$ [Albergo et al., 2023, Section 6]. In contrast, we implement the diffusion path using Langevin dynamics. Furthermore, their theoretical analysis does not include explicit non-asymptotic convergence bounds.

Tempering. Tempering [Geyer, 1992; Marinari and Parisi, 1992; Swendsen and Wang, 1986] is a well-known technique in the sampling literature that involves sampling the system at multiple temperatures: starting with higher temperatures to facilitate transitions between modes, gradually cooling the system to focus on the local structure of the target distribution. The sequence of tempered target distributions is typically defined using the geometric path, as it can be computed in closed form when the target density is known up to a normalising constant. Recently, several works have established theoretical guarantees for the convergence of geometric annealed Langevin Monte Carlo for non-log-concave distributions. In particular, Guo et al. [2025] provides a bound on the KL similar to that of Theorem 3.4. However, they are unable to obtain a closed-form expression for the action of the path. Besides, Chehab et al. [2025] derive upper and lower convergence bounds for the KL of the marginals, based on functional inequalities assumptions. In particular, they demonstrate that in some cases the log-Sobolev constant of the intermediate distributions along the path can deteriorate exponentially compared to those of the base and data distributions, unlike for the diffusion path.

6 Conclusions

In this work we provided a rigorous non-asymptotic analysis of Diffusion Annealed Langevin Monte Carlo (DALMC) for generative modelling, focusing on both Gaussian and heavy-tailed diffusion paths. By examining general diffusion paths that interpolate between complex data distributions and simpler base distributions, we have obtained theoretical insights into the convergence behaviour of DALMC under a range of assumptions. For Gaussian diffusion paths, we derived explicit non-asymptotic path-wise error bounds in KL divergence, improving upon prior results by relaxing smoothness assumptions and addressing the bias introduced through discretisation. Extending the framework to heavy-tailed diffusion paths, such as those based on Student’s t -distributions, we presented the first theoretical guarantees for these models, demonstrating comparable complexity to Gaussian diffusion paths under mild conditions.

Our analysis highlighted how smoothness assumptions, such as Lipschitz continuity of the scores and properties of the data distribution (e.g., bounded second-order moment and convexity or heavy-tailed behaviour), naturally ensure bounded action and efficient convergence. This generalisation broadens the applicability of DALMC beyond the settings considered in prior work. While DALMC introduces some bias compared to reverse SDE implementations, it avoids numerical instabilities and provides a simpler approach, making it a compelling alternative for score-based generative modelling, in certain settings.

Looking ahead, further work could focus on developing more efficient numerical schemes, reducing dimensional dependencies in error bounds, and applying this framework to other generative models.

Acknowledgments

PCE would like to thank Arnaud Guillin, Paul Felix Valsecchi Oliva, Pierre Monmarché and Yanbo Tang for their insightful discussions. PCE is supported by EPSRC through the Modern Statistics and Statistical Machine Learning (StatML) CDT programme, grant no. EP/S023151/1.

References

- Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.
- Luigi Ambrosio and Bernd Kirchheim. Rectifiable sets in metric and banach spaces. *Mathematische Annalen*, 318(3):527–555, 2000.
- Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2008.
- Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.
- D. Bakry and M. Émery. Diffusions hypercontractives. In Jacques Azéma and Marc Yor, editors, *Séminaire de Probabilités XIX 1983/84*, pages 177–206, Berlin, Heidelberg, 1985. Springer Berlin Heidelberg.
- Jean-Baptiste Bardet, Nathaël Gozlan, Florent Malrieu, and Pierre-André Zitt. Functional inequalities for Gaussian convolutions of compactly supported measures: Explicit bounds and dimension dependence. *Bernoulli*, 24(1):333 – 353, 2018.
- Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d -linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*, 2024.
- Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative Modeling with Denoising Auto-Encoders and Langevin Sampling. *arXiv preprint arXiv: 2002.00107*, 2022.
- Patrick Cattiaux, Nathael Gozlan, Arnaud Guillin, and Cyril Roberto. Functional Inequalities for Heavy Tailed Distributions and Application to Isoperimetry. *Electronic Journal of Probability*, 15:346 – 385, 2010a.
- Patrick Cattiaux, Arnaud Guillin, and Li-Ming Wu. A note on Talagrand’s transportation inequality and logarithmic Sobolev inequality. *Probability Theory and Related Fields*, 148(1):285–304, 2010b.
- Omar Chehab and Anna Korba. A practical diffusion path for sampling. In *ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2024.
- Omar Chehab, Anna Korba, Austin Stromme, and Adrien Vacher. Provable Convergence and Limitations of Geometric Tempering for Langevin Dynamics. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, 2022.
- Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023.
- Sinho Chewi. *Log-Concave Sampling. Book draft*. 2024. URL <https://chewisinho.github.io>.
- Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

Jean Alexandre Dieudonné. *Infinitesimal calculus*. Paris : Hermann, 1973.

Yuan Gao, Jian Huang, and Yuling Jiao. Gaussian Interpolation Flows. *Journal of Machine Learning Research*, 25(253):1–52, 2024.

Marta Gentiloni-Silveri and Antonio Ocello. Beyond Log-Concavity and Score Regularity: Improved Convergence Bounds for Score-Based Generative Models in W_2 -distance. *arXiv preprint arXiv:2501.02298*, 2025.

Charles J. Geyer. Practical Markov Chain Monte Carlo. *Statistical Science*, 7(4):473–483, 1992.

Louis Grenioux, Maxence Noble, Marylou Gabrié, and Alain Oliviero Durmus. Stochastic localization via iterative posterior sampling. In *Forty-first International Conference on Machine Learning*, 2024.

Wei Guo, Molei Tao, and Yongxin Chen. Provable Benefit of Annealed Langevin Monte Carlo for Non-log-concave Sampling. In *The Thirteenth International Conference on Learning Representations*, 2025.

Haoran He, Chenjia Bai, Kang Xu, Zhuoran Yang, Weinan Zhang, Dong Wang, Bin Zhao, and Xuelong Li. Diffusion Model is an Effective Planner and Data Synthesizer for Multi-Task Reinforcement Learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Richard Holley and Daniel Stroock. Logarithmic Sobolev inequalities and stochastic Ising models. *Journal of Statistical Physics*, 46(5):1159–1194, 1987.

Xunpeng Huang, Hanze Dong, Yifan Hao, Yian Ma, and Tong Zhang. Reverse Diffusion Monte Carlo. In *The Twelfth International Conference on Learning Representations*, 2024.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with Diffusion for Flexible Behavior Synthesis. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9902–9915. PMLR, 2022.

Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer New York, NY, 1991.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards non-asymptotic convergence for diffusion-based generative models. In *The Twelfth International Conference on Learning Representations*, 2024.

Stefano Lisini. Characterization of absolutely continuous curves in wasserstein spaces. *Calculus of Variations and Partial Differential Equations*, 28(1):85–120, 2007.

Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I. Jordan. Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116(42):20881–20885, 2019.

E. Marinari and G. Parisi. Simulated Tempering: A New Monte Carlo Scheme. *Europhysics Letters*, 19(6):451, 1992.

S. Nadarajah and D.K. Dey. Convolutions of the T distribution. *Computers & Mathematics with Applications*, 49(5):715–721, 2005.

Kushagra Pandey, Jaideep Pathak, Yilun Xu, Stephan Mandt, Michael Pritchard, Arash Vahdat, and Morteza Mardani. Heavy-Tailed Diffusion Models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Angus Phillips, Hai-Dang Dau, Michael John Hutchinson, Valentin De Bortoli, George Deligiannidis, and Arnaud Doucet. Particle Denoising Diffusion Sampler. In *Forty-first International Conference on Machine Learning*, 2024.

Jakiw Pidstrigach. Score-Based Generative Models Detect Manifolds. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Lorenz Richter and Julius Berner. Improved sampling via learned diffusions. In *The Twelfth International Conference on Learning Representations*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. MM-Diffusion: Learning Multi-Modal Diffusion Models for Joint Audio and Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10219–10228, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Saeed Saremi, Ji Won Park, and Francis Bach. Chain of Log-Concave Markov Chains. In *The Twelfth International Conference on Learning Representations*, 2024.

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

Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced Score Matching: A Scalable Approach to Density and Score Estimation. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 574–584. PMLR, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*, 2021.

Robert H. Swendsen and Jian-Sheng Wang. Replica Monte Carlo Simulation of Spin-Glasses. *Phys. Rev. Lett.*, 57:2607–2609, Nov 1986.

Francisco Vargas, Shreyas Padhy, Denis Blessing, and Nikolas Nüsken. Transport meets Variational Inference: Controlled Monte Carlo Diffusions. In *The Twelfth International Conference on Learning Representations*, 2024.

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

A Background

We introduce some concepts from optimal transport and the Girsanov theorem which will be useful for the subsequent analysis.

Optimal transport. Let $v = (v_t : \mathbb{R}^d \rightarrow \mathbb{R}^d)$ be a vector field and $\mu = (\mu_t)_{t \in [a,b]}$ be a curve of probability measures on \mathbb{R}^d with finite second-order moments. μ is generated by the vector field v if the continuity equation

$$\partial_t \mu_t + \nabla \cdot (\mu_t v_t) = 0,$$

holds for all $t \in [a,b]$. The metric derivative of μ at $t \in [a,b]$ is then defined as

$$|\dot{\mu}|_t := \lim_{\delta \rightarrow 0} \frac{W_2(\mu_{t+\delta}, \mu_t)}{|\delta|}.$$

If $|\dot{\mu}|_t$ exists and is finite for all $t \in [a,b]$, we say that μ is an absolutely continuous curve of probability measures. Ambrosio and Kirchheim [2000] establish weak conditions under which a curve of probability measures with finite second-order moments is absolutely continuous.

By Ambrosio et al. [2008, Theorem 8.3.1] we have that among all velocity fields v_t which produce the same flow μ , there is a unique optimal one with smallest $L^p(\mu_t; X)$ -norm. This is summarised in the following lemma.

Lemma A.1 (Lemma 2 from Guo et al. [2025]). *For an absolutely continuous curve of probability measures $\mu = (\mu_t)_{t \in [a,b]}$, any vector field $(v_t)_{t \in [a,b]}$ that generates μ satisfies $|\dot{\mu}|_t \leq \|v_t\|_{L^2(\mu_t)}$ for almost every $t \in [a,b]$. Moreover, there exists a unique vector field v_t^* generating μ such that $|\dot{\mu}|_t = \|v_t^*\|_{L^2(\mu_t)}$ almost everywhere.*

We also introduce the action of the absolutely continuous curve $(\mu_t)_{t \in [a,b]}$ since it will play a key role in our convergence results. In particular, we define the action $\mathcal{A}(\mu)$ as

$$\mathcal{A}(\mu) := \int_a^b |\dot{\mu}|_t^2 dt.$$

Girsanov's theorem. Consider the SDE

$$dX_t = b(X_t, t)dt + \sigma(X_t, t)dB_t,$$

for $t \in [0, T]$, where $(B_t)_{t \in [0,T]}$ is a standard Brownian motion in \mathbb{R}^d . Denote by \mathbb{P}^X the path measure of the solution $X = (X_t)_{t \in [0,T]}$ of the SDE, which characterises the distribution of X over the sample space Ω .

The KL divergence between two path measures can be characterised as a consequence of Girsanov's theorem [Karatzas and Shreve, 1991]. In particular, the following result will be central in our analysis.

Lemma A.2. *Consider the following two SDEs defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$*

$$dX_t = a_t(X)dt + \sqrt{2}dB_t, \quad dY_t = b_t(Y)dt + \sqrt{2}dB_t, \quad t \in [0, T]$$

with the same initial conditions $X_0, Y_0 \sim \mu_0$. Denote by \mathbb{P}^X and \mathbb{P}^Y the path measures of the processes X and Y , respectively. It follows that

$$\text{KL}(\mathbb{P}^X \parallel \mathbb{P}^Y) = \frac{1}{4} \mathbb{E}_{X \sim \mathbb{P}^X} \left[\int_0^T \|a_t(X) - b_t(X)\|^2 dt \right].$$

Preliminary results. Guo et al. [2025, Theorem 1] provide convergence guarantees for the continuous-time geometric annealed Langevin dynamics based on the action of the curve of probability measures given by the geometric mean of the base and target distributions. Their result can be adapted to our setting as follows.

Theorem A.3 (Theorem 1 [Guo et al., 2025]). *Let $\mathbb{P}_{\text{DALD}} = (p_{t,\text{DALD}})_{t \in [0, T/\kappa]}$ be the path measure of the diffusion annealed Langevin dynamics (3), and $\mathbb{P} = (\hat{\mu}_t)_{t \in [0, T/\kappa]}$ that of a reference SDE such that the marginals at each time have distribution $\hat{\mu}_t$. If $p_{0,\text{DALD}} = p_0$, the KL divergence between the path measures is upper bounded by*

$$\text{KL}(\mathbb{P} \parallel \mathbb{P}_{\text{DALD}}) \leq \kappa \mathcal{A}(\mu).$$

Proof. Let \mathbb{P} be the path measure corresponding to the following reference SDE

$$dY_t = (\nabla \log \hat{\mu}_t + v_t)(Y_t)dt + \sqrt{2}dB_t, \quad t \in [0, T/\kappa].$$

The vector field $v = (v_t)_{t \in [0, T/\kappa]}$ is designed such that $Y_t \sim \hat{\mu}_t$ for all $t \in [0, T/\kappa]$. Using the Fokker-Planck equation, we have that

$$\partial_t \hat{\mu}_t = \nabla \cdot (\hat{\mu}_t(\nabla \log \hat{\mu}_t + v_t)) + \Delta \hat{\mu}_t = -\nabla \cdot (\hat{\mu}_t v_t), \quad t \in [0, T/\kappa].$$

This implies that v_t satisfies the continuity equation and hence generates the curve of probability measures $(\hat{\mu}_t)_t$. Leveraging Lemma A.1, we choose v to be the one that minimises the $L^2(\hat{\mu}_t)$ norm, resulting in $\|v_t\|_{L^2(\hat{\mu}_t)} = |\dot{\hat{\mu}}|_t$ being the metric derivative. Using the form of Girsanov's theorem given in Lemma A.2 we have

$$\begin{aligned} \text{KL}(\mathbb{P} \parallel \mathbb{P}_{\text{DALD}}) &= \frac{1}{4} \mathbb{E}_{\mathbb{P}} \left[\int_0^{T/\kappa} \|v_t(X_t)\|^2 dt \right] = \frac{1}{4} \int_0^{T/\kappa} \|v_t(X_t)\|_{L^2(\hat{\mu})}^2 dt = \frac{1}{4} \int_0^{T/\kappa} |\dot{\hat{\mu}}|_t^2 dt \\ &= \frac{\kappa}{4} \int_0^T |\dot{\hat{\mu}}|_t^2 dt = \frac{\kappa \mathcal{A}(\mu)}{4}, \end{aligned}$$

where we have used that $|\dot{\hat{\mu}}|_t = \kappa |\dot{\mu}|_t$ and the change of variable formula. \square

B Proofs of Section 3

B.1 Comments on Assumption A4

A typical assumption in the literature [Grenioux et al., 2024; Saremi et al., 2024] considers the data distribution is given by the convolution of a compactly supported measure and a Gaussian distribution, which can be formalised as follows.

A11. *Let X be a d -dimensional random vector $X \sim \pi_{\text{data}}$, such that $X = U + G$, where $\|U - m_{\pi}\|^2 \leq dR^2$ holds almost surely and $G \sim \mathcal{N}(0, \tau^2 I)$ is independent of U .*

We demonstrate that assumption A11 implies that the potential V_{π} has Lipschitz continuous gradients, satisfies the dissipativity condition and, also ensures that assumption A3 is satisfied. Furthermore, under additional assumptions on the compactly supported measure in A11, we show that A11 entails A4.

Lemma B.1. *Let $\pi_{\text{data}} \propto e^{-V_{\pi}}$, assumption A11 implies that V_{π} has Lipschitz continuous gradients and satisfies the dissipativity inequality*

$$\langle \nabla V_{\pi}(x), x \rangle \geq a_{\pi} \|x\|^2 - b_{\pi},$$

with constants $a_{\pi}, b_{\pi} > 0$. Furthermore, π_{data} has a finite log-Sobolev constant.

Proof. First we show that if $\pi_{\text{data}} \propto e^{-V_\pi}$ satisfies **A11** then V_π is gradient Lipschitz. Let $X \sim \pi_{\text{data}}$, $U \sim \tilde{\pi}$ with compact support and $G \sim \gamma = \mathcal{N}(0, \tau^2 I)$ independent of U . By assumption **A11** we have $X = U + G$ or equivalently $\pi_{\text{data}} = \tilde{\pi} * \gamma$. Using that if g is a differentiable function then $\nabla(f * g) = f * (\nabla g)$, we have that

$$-\nabla \log \pi_{\text{data}}(x) = \frac{1}{\tau^2} (x - \mathbb{E}_{\rho_x}[Y]) \quad (5)$$

$$-\nabla^2 \log \pi_{\text{data}}(x) = \frac{1}{\tau^2} \left(I - \frac{1}{\tau^2} \text{Cov}_{\rho_x}[Y] \right), \quad (6)$$

where $\rho_x(y) \propto \tilde{\pi}(y)\gamma(x-y)$ and $\text{Cov}_{\rho_x}[Y] = \mathbb{E}_{\rho_x}[YY^\top] - \mathbb{E}_{\rho_x}[Y]\mathbb{E}_{\rho_x}[Y]^\top$. Note that $\rho_x(y)$ has bounded support independent of x by **A11**. Therefore, the eigenvalues of $\text{Cov}_{\rho_x}[Y]$ can be upper bounded by a constant independent of x . That is, for any $a \in \mathbb{R}^d$ with $\|a\| = 1$

$$\begin{aligned} -a^\top (\nabla^2 \log \pi_{\text{data}}(x)) a &= \tau^{-2} - \tau^{-4} a^\top \text{Cov}_{\rho_x}[Y] a \geq \tau^{-2} - \tau^{-4} a^\top \mathbb{E}_{\rho_x}[YY^\top] a \\ &\geq \tau^{-2} - \tau^{-4} \mathbb{E}_{\rho_x}[(a^\top Y)^2] \geq \tau^{-2} - \tau^{-4} \mathbb{E}_{\rho_x}[\|Y\|^2] \\ &\geq \tau^{-2} - \tau^{-4} (m_\pi + dR^2), \end{aligned}$$

where we have used Cauchy-Schwarz inequality and **A11**. On the other hand, since the covariance matrix is positive semidefinite, we have

$$-\nabla^2 \log \pi_{\text{data}}(x) \preceq \tau^{-2} I.$$

Therefore, the Hessian $-\nabla^2 \log \pi_{\text{data}}$ satisfies

$$(\tau^{-2} - \tau^{-4} (m_\pi + dR^2)) I \preceq -\nabla^2 \log \pi_{\text{data}}(x) \preceq \tau^{-2} I,$$

proving that $-\nabla \log \pi_{\text{data}}$ is gradient Lipschitz with constant $L_\pi \leq \max\{\tau^{-2}, |\tau^{-2} - \tau^{-4} (m_\pi + dR^2)|\}$.

On the other hand, using the expression for $\nabla V_\pi = -\nabla \log \pi_{\text{data}}$ in (5), we have

$$\begin{aligned} \langle \nabla V_\pi(x), x \rangle &= \frac{1}{\tau^2} \langle x - \mathbb{E}_{\rho_x}[Y], x \rangle = \frac{1}{\tau^2} (\|x\|^2 - \mathbb{E}_{\rho_x}[\langle Y, x \rangle]) \geq \frac{1}{\tau^2} (\|x\|^2 - \|x\| \mathbb{E}_{\rho_x}[\|Y\|]) \\ &\geq \frac{1}{\tau^2} (\|x\|^2 - \|x\| (m_\pi + \sqrt{d}R)) \geq \frac{\|x\|^2}{2\tau} - (m_\pi + \sqrt{d}R)^2, \end{aligned}$$

where we have used that $\rho_x(y) \propto \tilde{\pi}(y)\gamma(x-y)$ has bounded support. This establishes the dissipativity inequality.

Finally, thanks to the dissipativity condition together with Lipschitz gradient, it follows from Cattiaux et al. [2010b] that π_{data} has a finite log-Sobolev constant. \square

We now demonstrate the **A11** implies **A3**.

Lemma B.2. *If π_{data} satisfies assumption **A11**, then the scores $\nabla \log \mu_t$ of the intermediate probability densities of the Gaussian diffusion path are Lipschitz continuous for all t , that is, assumption **A3** is satisfied.*

Proof. Recall that the intermediate random variables of the Gaussian diffusion path, are given by

$$X_t = \sqrt{\lambda_t} X + \sqrt{1 - \lambda_t} \sigma Z$$

where $X \sim \pi_{\text{data}}$ and $Z \sim \mathcal{N}(0, I)$ independent of X . Using assumption **A11**, it follows that

$$X_t \stackrel{d}{=} \sqrt{\lambda_t} U + \sqrt{(1 - \lambda_t)\sigma^2 + \lambda_t \tau^2} Z,$$

where $U \sim \tilde{\pi}$ is compactly supported and $Z \sim \mathcal{N}(0, I)$ independent of $\tilde{\pi}$. By applying the result from the previous lemma (Lemma B.1), we conclude that $\nabla \log \mu_t$, where $\mu_t \sim X_t$, is Lipschitz continuous with constant

$$L_t \leq \max\{\tau_t^{-1}, |\tau_t^{-1} - \tau_t^{-2} (m_\pi + dR^2)|\},$$

where $\tau_t^2 = (1 - \lambda_t)\sigma^2 + \lambda_t \tau^2$. This completes the proof. \square

We note that in general, **A11** does not imply strong convexity outside of a ball. For example, consider the following example in \mathbb{R}^2 . Let $\pi_{\text{data}} = \tilde{\pi} * \gamma$ and $\tilde{\pi} = \frac{1}{2}(\delta_{y_1} + \delta_{y_2})$, where $y_i = (0, (-1)^i R)$. Consider a point in \mathbb{R}^2 of the form $(x, 0)$. We have that the conditional measure $\rho_{(x,0)}$, where $\rho_x(y) \propto \tilde{\pi}(y)\gamma(x - y)$, satisfies

$$\rho_{(x,0)}(y) = \tilde{\pi}(y).$$

Therefore, the covariance term in the expression of the Hessian in (6) is given by

$$\text{Cov}_{\rho_{(x,0)}}(Y) = R^2 \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Substituting this into the expression of the Hessian, it follows that

$$-\nabla^2 \log \pi_{\text{data}}((x, 0)) = \frac{I}{\tau^2} - \frac{R^2}{\tau^4} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Taking R to be greater than τ , we have that $-\nabla^2 \log \pi_{\text{data}}$ evaluated at points on the axis $x = 0$ is not positive definite, meaning that π_{data} cannot be strongly convex outside of any ball. Under the additional assumption that the support of the compactly supported measure $\tilde{\pi}$ is convex and dense in its ambient space, **A11** implies **A4**. This result is formalised in the following lemma.

Lemma B.3. *Let $\pi_{\text{data}} = \tilde{\pi} * \gamma \in \mathcal{P}(\mathbb{R}^d)$ satisfy assumption **A11**. If the support of $\tilde{\pi}$ is convex and dense in \mathbb{R}^d , then π_{data} satisfies assumption **A4**.*

Proof. By Lemma B.1, we have that ∇V_π is Lipschitz continuous. Thus, it remains to show that V_π is strongly convex outside of a ball.

Recall that $\tilde{\pi}$ is supported on a compact set S , that is, $\tilde{\pi}(y) = 0$ for $y \notin S$ and by assumption S is also convex. For $x \in \mathbb{R}^d$, define the function

$$d(x) = \min_{y \in S} \|x - y\|,$$

which is well defined by compactness. Let $y^*(x) \in S$ be the unique (by convexity of S) point where the minimum distance is achieved, i.e., $y^*(x)$ is the projector of x onto S . Then, for every $y \in S$ it holds that

$$\|x - y\| \geq d(x),$$

with equality if and only if $y = y^*(x)$. Consider the convolution kernel $G \sim \gamma(x - y)$ defined as

$$\gamma(x - y) = \frac{1}{(2\pi\tau^2)^{d/2}} e^{-\frac{\|x-y\|^2}{2\tau^2}}.$$

Note that the value at $y = y^*(x)$ is given by

$$\gamma(x - y^*(x)) = \frac{1}{(2\pi\tau^2)^{d/2}} e^{-\frac{d(x)^2}{2\tau^2}}.$$

Besides, for any $y \in S$, we have that

$$d(x) \leq \|x - y\| \leq d(x) + \|y^*(x) - y\|.$$

Because S is a compact set, the term $\|y^*(x) - y\|$ for $y \in S$ is bounded independently of x , therefore we can write

$$\|x - y\| = d(x) + \delta_x(y),$$

for some $\delta_x(y) \geq 0$, with $\delta_x(y)$ if and only if $y = y^*(x)$ and $\delta_x(y) \leq \|y^*(x) - y\|$, which implies that $\delta(y)$ remains uniformly bounded for all $x \in \mathbb{R}^d$, in particular as $\|x\| \rightarrow \infty$. Using this, the convolution kernel can be written as

$$\gamma(x - y) = \frac{1}{(2\pi\tau^2)^{d/2}} e^{-\frac{(d(x) + \delta_x(y))^2}{2\tau^2}}.$$

Thus, we obtain the ratio

$$\frac{\gamma(x-y)}{\gamma(x-y^*(x))} = e^{-\frac{2d(x)\delta_x(y)+\delta_x(y)^2}{2\tau^2}}.$$

Since $\delta_x(y)$ is uniformly bounded for all x , we observe that for $y \neq y^*(x)$, as $\|x\| \rightarrow \infty$ the leading order of the exponent is $d(x)\delta_x(y)$, where $\delta_x(y) > 0$ and $d(x)$ grows with order $\|x\|$. Meaning that the ratio becomes arbitrarily small as $\|x\| \rightarrow \infty$ when $y \neq y^*(x)$. That is, the contribution from $y \neq y^*(x)$ becomes exponentially negligible compared to the contribution from $y^*(x)$.

Given now a bounded test function f , we have

$$\frac{\int f(y)\gamma(x-y)\tilde{\pi}(y)dy}{\int \gamma(x-y)\tilde{\pi}(y)dy} = \frac{\int f(y)e^{-\frac{d(x)\delta_x(y)+\delta_x(y)^2}{2\tau^2}}\gamma(x-y^*(x))\tilde{\pi}(y)dy}{\int e^{-\frac{d(x)\delta_x(y)+\delta_x(y)^2}{2\tau^2}}\gamma(x-y^*(x))\tilde{\pi}(y)dy}.$$

By the dominated convergence theorem, the contribution in both integrals for $y \neq y^*(x)$ vanishes as $\|x\| \rightarrow \infty$. Therefore, we have that for any fixed $\varepsilon > 0$ and any small radius $\delta > 0$, there exists r such that for all $\|x\| > r$, the conditional measure $\rho_x(y) \propto \tilde{\pi}(y)\gamma(x-y)$ satisfies

$$\rho_x(S \setminus B(y^*(x), \delta)) < \varepsilon,$$

where $B(y^*(x), \delta)$ denotes the ball of radius δ centred at $y^*(x)$. Intuitively, this means that for sufficiently large $\|x\|$, almost all the mass of ρ_x is concentrated within an arbitrarily small ball around $y^*(x)$. It is important to note that, due to the assumption that S is dense in \mathbb{R}^d , for any $\delta > 0$, there always exist a point $z \in B(y^*(x), \delta)$ such that $\tilde{\pi}(z) > 0$.

Consequently, the mean $\mu_x = \int_S y\rho_x(dy)$ must be very close to $y^*(x)$, and for any point y in the high-probability region, we have $\|y - \mu(x)\| \leq 2\delta$ (with the worst-case scenario occurring when $\mu(x)$ lies on the edge of $B(y^*(x), \delta)$). This implies that for sufficiently large $\|x\|$, the spread of ρ_x becomes arbitrarily small. In particular, the covariance matrix satisfies

$$\|\text{Cov}_{\rho_x}(Y)\| \leq (2\gamma)^2.$$

Taking the limit as $\|x\| \rightarrow \infty$, we have that $\delta \rightarrow 0$, thus, we obtain that

$$\lim_{\|x\| \rightarrow \infty} \text{Cov}_{\rho_x}(Y) = 0.$$

The final step is to note that

$$\nabla^2 V_\pi(x) = \frac{1}{\tau^2} I - \frac{1}{\tau^4} \text{Cov}_{\rho_x}(Y).$$

□

We now show that a mixture of Gaussians with different covariances satisfies assumption **A4** under mild conditions, but it does not generally satisfy assumption **A11**.

Lemma B.4. *Let $\pi = \sum_{i=1}^M w_i p_i$ be a mixture of Gaussians in \mathbb{R}^d where w_i and p_i denote the weight and the probability density function, respectively, of the i -th component of the mixture which has mean μ_i and covariance Σ_i . If for any pair $\{i, j\}$ with $\Sigma_i \neq \Sigma_j$ there exists a unit vector u such that*

$$u^\top (\Sigma_i^{-1} - \Sigma_j^{-1}) u = 0,$$

but one of the following conditions hold

- (i) u is an eigenvector of $(\Sigma_i^{-1} - \Sigma_j^{-1})$ with eigenvalue 0.
- (ii) $u^\top (\Sigma_i^{-1} \mu_i - \Sigma_j^{-1} \mu_j) \neq 0$.
- (iii) There exists $k \in \{1, \dots, M\}$ such that $u^\top (\Sigma_i^{-1} - \Sigma_k^{-1}) u > 0$ or $u^\top (\Sigma_j^{-1} - \Sigma_k^{-1}) u > 0$.

Then $\nabla \log \pi$ is Lipschitz. Moreover, if also for any pair $\{i, j\}$ with $\Sigma_i \mu_i \neq \Sigma_j \mu_j$ there exists a unit vector u such that $u^\top (\Sigma_i^{-1} - \Sigma_j^{-1}) u = 0$ but either condition (ii) or (iii) hold, then π satisfies assumption **A4**.

Before presenting the proof, we observe that in one dimension, a mixture of Gaussians with different variances always satisfies the condition stated in the previous lemma, and thus assumption **A4** holds.

Proof. We first establish that $\nabla \log \pi$ is Lipschitz continuous by bounding the spectral norm of the Hessian $\nabla^2 \log \pi$. We have the following expressions for $\nabla \log \pi$ and $\nabla^2 \log \pi$

$$\begin{aligned}\nabla \log \pi &= \frac{\sum_i w_i p_i \nabla \log p_i}{\sum_i w_i p_i} \\ \nabla^2 \log \pi &= \frac{\sum_i w_i \nabla^2 p_i}{\sum_i w_i p_i} - \frac{(\sum_i w_i p_i \nabla \log p_i)(\sum_i w_i p_i \nabla \log p_i^\top)}{(\sum_i w_i p_i)^2}.\end{aligned}\quad (7)$$

Observe that

$$\nabla^2 p_i = p_i \nabla^2 \log p_i + p_i \nabla \log p_i (\nabla \log p_i)^\top.$$

Substituting this we have

$$\begin{aligned}-\nabla^2 \log \pi &= \frac{\sum_i w_i p_i \Sigma_i^{-1}}{\sum_i w_i p_i} - \frac{\sum_{i,j} w_i w_j p_i p_j [\nabla \log p_i (\nabla \log p_i)^\top - \nabla \log p_i (\nabla \log p_j)^\top]}{(\sum_i w_i p_i)^2} \\ &= \frac{\sum_i w_i p_i \Sigma_i^{-1}}{\sum_i w_i p_i} - \frac{\sum_{i,j} w_i w_j p_i p_j [\Sigma_i^{-1} (x - \mu_i)(x - \mu_i)^\top \Sigma_i^{-1} - \Sigma_i^{-1} (x - \mu_i)(x - \mu_j)^\top \Sigma_j^{-1}]}{(\sum_i w_i p_i)^2} \\ &= \frac{\sum_i w_i p_i \Sigma_i^{-1}}{\sum_i w_i p_i} - \frac{\sum_{i,j} w_i w_j p_i p_j [\Sigma_i^{-1} x x^\top (\Sigma_i^{-1} - \Sigma_j^{-1}) + \frac{1}{2} (\Sigma_i^{-1} \mu_i - \Sigma_j^{-1} \mu_j)(\Sigma_i^{-1} \mu_i - \Sigma_j^{-1} \mu_j)^\top]}{(\sum_i w_i p_i)^2} \\ &\quad + \frac{1}{2} \frac{\sum_{i,j} w_i w_j p_i p_j [(\Sigma_i^{-1} \mu_i - \Sigma_j^{-1} \mu_j)x^\top (\Sigma_i^{-1} - \Sigma_j^{-1}) + (\Sigma_i^{-1} - \Sigma_j^{-1})x(\Sigma_i^{-1} \mu_i - \Sigma_j^{-1} \mu_j)^\top]}{(\sum_i w_i p_i)^2} \\ &= \frac{\sum_i w_i p_i \Sigma_i^{-1}}{\sum_i w_i p_i} - \frac{1}{2} \frac{\sum_{i,j} w_i w_j p_i p_j [(\Sigma_i^{-1} - \Sigma_j^{-1})x x^\top (\Sigma_i^{-1} - \Sigma_j^{-1}) + (\Sigma_i^{-1} \mu_i - \Sigma_j^{-1} \mu_j)(\Sigma_i^{-1} \mu_i - \Sigma_j^{-1} \mu_j)^\top]}{(\sum_i w_i p_i)^2} \\ &\quad + \frac{\sum_{i,j} w_i w_j p_i p_j [\Sigma_i^{-1} \mu_i x^\top (\Sigma_i^{-1} - \Sigma_j^{-1}) + (\Sigma_i^{-1} - \Sigma_j^{-1})x \mu_i^\top \Sigma_i^{-1}]}{(\sum_i w_i p_i)^2}.\end{aligned}$$

Note that in the case of equal covariances ($\Sigma_i = \Sigma_j$) the terms involving x cancel out. Since the covariance matrices satisfy $\sigma_{i,\min} I \preceq \Sigma_i \preceq \sigma_{i,\max} I$ and the norm of the means $\|\mu_i\|$ is finite for all i , the following terms of the previous expression

$$A_x + B_x = \frac{\sum_i w_i p_i \Sigma_i^{-1}}{\sum_i w_i p_i} - \frac{1}{2} \frac{\sum_{i,j} w_i w_j p_i p_j (\Sigma_i^{-1} \mu_i - \Sigma_j^{-1} \mu_j)(\Sigma_i^{-1} \mu_i - \Sigma_j^{-1} \mu_j)^\top}{(\sum_i w_i p_i)^2}$$

are uniformly bounded above and below for all x . We now focus on the remaining terms which can be rewritten as

$$C_x = -\frac{1}{4} \frac{\sum_{i,j} w_i w_j p_i p_j [M_x + M_x^\top]}{(\sum_i w_i p_i)^2}, \quad M_x = (\Sigma_i^{-1} - \Sigma_j^{-1})x [x^\top (\Sigma_i^{-1} - \Sigma_j^{-1}) - 4\mu_i^\top \Sigma_i^{-1}]$$

aiming to establish an upper bound for the spectral norm of C_x when $\|x\|$ tends to ∞ . Hence, from this point onwards, we consider x such that $\|x\| > \max_i \|\mu_i\|$. Using the triangle inequality and the submultiplicativity property of the spectral norm we obtain

$$\begin{aligned}\|C_x\|_2 &\leq \frac{1}{4} \sum_{i,j} \frac{w_i w_j p_i p_j}{(\sum_i w_i p_i)^2} \|M_x + M_x^\top\|_2 \\ \|M_x + M_x^\top\|_2 &\leq 2 \|\Sigma_i^{-1} - \Sigma_j^{-1}\|_2^2 \|x x^\top\|_2 + 4 \|\Sigma_i^{-1} - \Sigma_j^{-1}\|_2 \|\Sigma_i^{-1}\|_2 \|x \mu_i^\top + \mu_i x^\top\|_2 \\ &\leq 2\|x\|^2 \left(\|\Sigma_i^{-1} - \Sigma_j^{-1}\|_2^2 + 4 \|\Sigma_i^{-1} - \Sigma_j^{-1}\|_2 \|\Sigma_i^{-1}\|_2 \right) \leq 10\|x\|^2 \left(\|\Sigma_i^{-1}\|_2 + \|\Sigma_j^{-1}\|_2 \right)^2.\end{aligned}\quad (8)$$

Let us define the following sets

$$\begin{aligned} D_1 &= \{\{i, j\} | 1 \leq i, j \leq M, \Sigma_i \neq \Sigma_j \text{ and } \nexists u \mid \|u\| = 1 \text{ and } u^\top (\Sigma_i^{-1} - \Sigma_j^{-1}) u = 0\}, \\ D_2 &= \{\{i, j\} | 1 \leq i, j \leq M, \Sigma_i \neq \Sigma_j \text{ and } \exists u \mid \|u\| = 1 \text{ and } u^\top (\Sigma_i^{-1} - \Sigma_j^{-1}) u = 0 \text{ and (i) holds}\}, \\ D_3 &= \{\{i, j\} | 1 \leq i, j \leq M, \Sigma_i \neq \Sigma_j \text{ and } \exists u \mid \|u\| = 1 \text{ and } u^\top (\Sigma_i^{-1} - \Sigma_j^{-1}) u = 0 \text{ and (ii) holds}\}, \\ D_4 &= \{\{i, j\} | 1 \leq i, j \leq M, \Sigma_i \neq \Sigma_j \text{ and } \exists u \mid \|u\| = 1 \text{ and } u^\top (\Sigma_i^{-1} - \Sigma_j^{-1}) u = 0 \text{ and (iii) holds}\}, \end{aligned}$$

We also consider the partition of the unit sphere $S^{d-1} \subset \mathbb{R}^d$ into disjoint subsets

$$S^{d-1} = P_+^{(j,i)} \cup P_-^{(j,i)} \cup P_0^{(j,i)},$$

where $P_+^{(j,i)} = \{u \in S^{d-1} | u^\top (\Sigma_j^{-1} - \Sigma_i^{-1}) u > 0\}$, with $P_-^{(j,i)}$ and $P_0^{(j,i)}$ defined analogously.

We analyse the terms in the sum (8) separately, depending on the set D_k to which the pair $\{i, j\}$ belongs.

(1) Consider the pairs $\{i, j\} \in D_1$, it follows that

$$\sum_{\{i,j\} \in D_1} \frac{w_i w_j p_i p_j}{(\sum_i w_i p_i)^2} \|M_x + M_x^\top\|_2 \leq 10 \max_{\{i,j\} \in D_1} \left\{ \left(\|\Sigma_i^{-1}\|_2 + \|\Sigma_j^{-1}\|_2 \right)^2 \right\} \sum_{\{i,j\} \in D_1} \frac{w_i w_j p_i p_j}{(\sum_i w_i p_i)^2} \|x\|^2.$$

To analyse each term in the previous sum, we first consider the case where one covariance matrix majorises the other. Specifically, without loss of generality, we assume that for the pair $\{i, j\} \in D_1$ we have $\Sigma_i \succ \Sigma_j$, which implies $\Sigma_j^{-1} - \Sigma_i^{-1} \succ \alpha I$ for some $\alpha > 0$. We observe that

$$\begin{aligned} -\frac{1}{2} x^\top (\Sigma_j^{-1} - \Sigma_i^{-1}) x + x^\top (\Sigma_j^{-1} \mu_j - \Sigma_i^{-1} \mu_i) &\leq -\frac{1}{2} \alpha \|x\|^2 + 2\|x\| \left(\|\Sigma_j^{-1/2}\|_2^2 \|\mu_j\| + \|\Sigma_i^{-1/2}\|_2^2 \|\mu_i\| \right) \\ &\leq -\frac{1}{2} \alpha \|x\|^2 + 2\|x\| \left(\sigma_{j,\min}^{-1} \|\mu_j\| + \sigma_{i,\min}^{-1} \|\mu_i\| \right), \end{aligned} \quad (9)$$

which gives

$$\begin{aligned} \frac{w_i w_j p_i p_j \|x\|^2}{(\sum_i w_i p_i)^2} &\leq \frac{w_i w_j p_i p_j \|x\|^2}{(w_i p_i)^2} = \frac{w_j p_j \|x\|^2}{w_i p_i} \\ &= \frac{w_j \det(\Sigma_i)^{1/2}}{w_i \det(\Sigma_j)^{1/2}} e^{-\frac{1}{2} (\mu_j^\top \Sigma_j^{-1} \mu_j - \mu_i^\top \Sigma_i^{-1} \mu_i)} \|x\|^2 e^{-\frac{1}{2} x^\top (\Sigma_j^{-1} - \Sigma_i^{-1}) x + x^\top (\Sigma_j^{-1} \mu_j - \Sigma_i^{-1} \mu_i)} \xrightarrow{\|x\| \rightarrow \infty} 0. \end{aligned} \quad (10)$$

On the other hand, when $\Sigma_j^{-1} - \Sigma_i^{-1}$ is neither positive-definite nor negative-definite, for every $x \in \mathbb{R}^d$ we can write $x = \|x\|u$, where u is a unit vector satisfying $u \in P_+^{(j,i)}$ or $u \in P_-^{(j,i)}$ because $P_0^{(j,i)}$ is empty by definition for $\{i, j\} \in D_1$. These two cases can be treated simultaneously since the indices i, j are interchangeable. Without loss of generality, we assume that $u \in P_+^{(j,i)}$. Following a similar approach to equations (9) and (10), we have

$$\frac{w_i w_j p_i p_j \|x\|^2}{(\sum_i w_i p_i)^2} \leq \frac{w_i w_j p_i p_j \|x\|^2}{(w_i p_i)^2} = \frac{w_j p_j \|x\|^2}{w_i p_i} \xrightarrow{\|x\| \rightarrow \infty} 0.$$

Therefore, for every unit vector $u \in S^{d-1}$, the limit of each term in the sum over $\{i, j\} \in D_1$ along the line $\|x\|u$ is zero as $\|x\|$ tends to ∞ . Since the sum contains finitely many terms, this implies

$$\lim_{\|x\| \rightarrow \infty} f_{D_1}(\|x\|u) = \lim_{\|x\| \rightarrow \infty} \sum_{\{i,j\} \in D_1} \frac{w_i w_j p_i (\|x\|u) p_j (\|x\|u)}{(\sum_i w_i p_i (\|x\|u))^2} \|M_{\|x\|u} + M_{\|x\|u}^\top\|_2 = 0.$$

Since f_{D_1} is a continuous function and S^{d-1} is compact, the behaviour of f_{D_1} can be controlled uniformly across all directions. That is, for every $\varepsilon > 0$ there exists $R > 0$ such that $\|x\| > R$ implies

$$\left| \sum_{\{i,j\} \in D_1} \frac{w_i w_j p_i p_j}{(\sum_i w_i p_i)^2} \|M_x + M_x^\top\|_2 \right| < \varepsilon.$$

(2) For $\{i, j\} \in D_2 \cup D_3 \cup D_4$, following the same reasoning as above, the limit of the spectral norm of each term when $\|x\|$ tends to ∞ along the directions $u \in P_+^{(j,i)}$ or $u \in P_-^{(j,i)}$ is 0. To analyse the limit along the directions $u \in P_0^{(j,i)}$, we consider the following cases.

- $\{i, j\} \in D_2$. For every x such that $x = \|x\|u$ with $u \in P_0^{(j,i)}$, we have that $(\Sigma_i^{-1} - \Sigma_j^{-1})u = 0_d$. Consequently, $M_x = 0_{d \times d}$, which demonstrates that the limit along these directions is 0.
- $\{i, j\} \in D_3$. Take $u \in P_0^{(j,i)}$, by condition (ii) we have that either $u^\top (\Sigma_j^{-1}\mu_j - \Sigma_i^{-1}\mu_i) < 0$ or $u^\top (\Sigma_j^{-1}\mu_j - \Sigma_i^{-1}\mu_i) > 0$. Without loss of generality, assume that $u^\top (\Sigma_j^{-1}\mu_j - \Sigma_i^{-1}\mu_i) < 0$. Then for every $x = \|x\|u$ we have

$$\begin{aligned} \frac{w_i w_j p_i p_j}{(\sum_i w_i p_i)^2} \|M_x + M_x^\top\|_2 &\leq 10 (\|\Sigma_i^{-1}\|_2 + \|\Sigma_j^{-1}\|_2)^2 \frac{w_j p_j}{w_i p_i} \|x\|^2 \\ &\leq 10 (\|\Sigma_i^{-1}\|_2 + \|\Sigma_j^{-1}\|_2)^2 \frac{w_j \det(\Sigma_i)^{1/2}}{w_i \det(\Sigma_j)^{1/2}} e^{-\frac{1}{2}(\mu_j^\top \Sigma_j^{-1} \mu_j - \mu_i^\top \Sigma_i^{-1} \mu_i)} \|x\|^2 e^{\|x\| u^\top (\Sigma_j^{-1}\mu_j - \Sigma_i^{-1}\mu_i)} \xrightarrow{\|x\| \rightarrow \infty} 0. \end{aligned}$$

- $\{i, j\} \in D_4$. Take $u \in P_0^{(j,i)}$, by condition (iii), there exists k such that $u \in P_+^{(j,k)}$ or $u \in P_+^{(i,k)}$. These two cases are symmetric and can be treated together. Without loss of generality, assume $u \in P_+^{(j,k)}$. In this case, following a similar argument to those in equations (9) and (10), we have that for every $x = \|x\|u$

$$\frac{w_i w_j p_i p_j}{(\sum_i w_i p_i)^2} \|M_x + M_x^\top\|_2 \leq 10 (\|\Sigma_i^{-1}\|_2 + \|\Sigma_j^{-1}\|_2)^2 \frac{w_j p_j}{w_k p_k} \|x\|^2 \xrightarrow{\|x\| \rightarrow \infty} 0.$$

Since the sum in equation (8) contains a finite number of terms, we have that for every $u \in S^{d-1}$

$$\lim_{\|x\| \rightarrow \infty} \|C_{\|x\|u}\|_2 = 0.$$

Furthermore, because the function $\|C_x\|_2$ is continuous and S^{d-1} is compact, the limit $\lim_{\|x\| \rightarrow \infty} \|C_x\|_2$ exists and is equal to zero. Consequently, $\|C_x\|_2$ is bounded for all $x \in \mathbb{R}^d$, which concludes that $\nabla \log \pi$ is Lipschitz.

To complete the proof, we need show that $-\nabla \log \pi$ is strongly convex outside of a ball of radius r . Using the same technique as above, we analyse the spectral norm of B_x . Let us define the following sets

$$\begin{aligned} D_5 &= \{\{i, j\} | 1 \leq i, j \leq M, \Sigma_i \mu_i \neq \Sigma_j \mu_j \text{ and } \nexists u \mid \|u\| = 1 \text{ and } u^\top (\Sigma_i^{-1} - \Sigma_j^{-1})u = 0\}, \\ D_6 &= \{\{i, j\} | 1 \leq i, j \leq M, \Sigma_i \mu_i \neq \Sigma_j \mu_j \text{ and } \exists u \mid \|u\| = 1 \text{ and } u^\top (\Sigma_i^{-1} - \Sigma_j^{-1})u = 0 \text{ and (ii) holds}\}, \\ D_7 &= \{\{i, j\} | 1 \leq i, j \leq M, \Sigma_i \mu_i \neq \Sigma_j \mu_j \text{ and } \exists u \mid \|u\| = 1 \text{ and } u^\top (\Sigma_i^{-1} - \Sigma_j^{-1})u = 0 \text{ and (iii) holds}\}, \end{aligned}$$

it follows that

$$\|B_x\|_2 \leq \max_{\{i, j\} \in D_5 \cup D_6 \cup D_7} \|\Sigma_i^{-1}\mu_i - \Sigma_j^{-1}\mu_j\|^2 \sum_{\{i, j\} \in D_5 \cup D_6 \cup D_7} \frac{w_i w_j p_i p_j}{(\sum_i w_i p_i)^2}.$$

By applying the same reasoning as above, we find that for each pair in the sum

$$\lim_{\|x\| \rightarrow \infty} \frac{w_i w_j p_i p_j}{(\sum_i w_i p_i)^2} = 0.$$

Since there is a finite number of pairs $\{i, j\}$ in $D_5 \cup D_6 \cup D_7$, we can conclude that

$$\lim_{\|x\| \rightarrow \infty} \|B_x\|_2 = 0.$$

Thus, as $\|x\|$ tends to ∞ , the only term whose spectral norm does not vanish is

$$A_x = \frac{\sum_i w_i p_i \Sigma_i^{-1}}{\sum_i w_i p_i} \succcurlyeq \frac{\sum_i w_i p_i \sigma_{i,\max}^{-1}}{\sum_i w_i p_i} I \succcurlyeq \min_i \{\sigma_{i,\max}^{-1}\} I.$$

Therefore, we can conclude that π is strongly log-concave outside of a ball, and hence satisfies assumption **A4**. \square

Remark B.5. A concurrent work [Gentiloni-Siliveri and Ocello, 2025] examines the smoothness of a mixture of Gaussians with covariances of the form $\Sigma_i = \sigma_i^2 I$, a specific case that satisfies the assumption of Lemma B.4. However, their result does not extend to the case of general covariance matrices. In particular, the following example, which falls outside the assumptions of Lemma B.4, serves as a counterexample. Let $\pi = \frac{1}{2}(p_1 + p_2)$, where $p_i = \mathcal{N}(\mu, \Sigma_i)$ with $\mu = (1, 0)$ and

$$\Sigma_1^{-1} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad \Sigma_2^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}.$$

We can show that $\nabla^2 \log \pi((x, 0))$ is unbounded when $\|x\| \rightarrow \infty$. To establish this, first note that

$$p_1((x, 0)) = (2\pi\sqrt{3})^{-1}e^{-(x-1)^2}, \quad p_2((x, 0)) = (2\pi\sqrt{6})^{-1}e^{-(x-1)^2}.$$

Substituting this into the expression of the Hessian given above (7) we have

$$-\nabla^2 \log \pi((x, 0)) = \frac{1}{1+\sqrt{2}} \begin{pmatrix} 2\sqrt{2}+1 & \sqrt{2} \\ \sqrt{2} & 2\sqrt{2}+3 \end{pmatrix} - (x-1)^2 \frac{\sqrt{2}}{3+\sqrt{2}} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$

which is clearly unbounded as $\|x\| \rightarrow \infty$, meaning that $\nabla \log \pi$ is not Lipschitz continuous.

Besides, in general a mixture of Gaussians with different covariances does not satisfy assumption **A11**.

Lemma B.6. Let $\pi = \sum_{i=1}^M w_i \mathcal{N}(\mu_i, \Sigma_i)$ be a mixture of Gaussians in \mathbb{R}^d . If either one of the two following assumptions holds

- (i) There exists at least one covariance matrix Σ_i that cannot be expressed as $\Sigma_i = \sigma_i I$.
- (ii) There exists at least one pair $\{i, j\}$ such that $\Sigma_i \neq \Sigma_j$.

Then, π does not satisfy assumption **A11**.

Proof. We want to determine if $\pi(x)$ can be written as

$$\pi(x) = (h * \gamma)(x),$$

where h is a compactly supported measure and γ is a Gaussian distribution $\gamma = \mathcal{N}(0, \tau^2 I)$ for some τ^2 . Assume that $\pi(x) = (h * \gamma)(x)$, we will show that h cannot be compactly supported for any τ^2 .

Since the convolution in real space corresponds to multiplication in Fourier space, we have

$$\hat{\pi}(k) = \hat{h}(k)\hat{\gamma}(k)$$

where $\hat{\pi}(k), \hat{h}(k), \hat{\gamma}(k)$ denote the respective Fourier transforms, which have the following expressions

$$\hat{\pi}(k) = \sum_{i=1}^M w_i e^{-\frac{1}{2}k^\top \Sigma_i k - i\mu_i^\top k}, \quad \hat{\gamma}(k) = e^{-\frac{1}{2}\tau^2 k^\top I k}.$$

Then, the function $\hat{h}(k)$ has to satisfy

$$\hat{h}(k) = \frac{\hat{\pi}(k)}{\hat{\gamma}(k)} = \sum_{i=1}^M w_i e^{-\frac{1}{2}k^\top (\Sigma_i - \tau^2 I) k - i\mu_i^\top k}.$$

Note that τ^2 needs to satisfy $\tau^2 I \preceq \Sigma_i$ for $i = 1, \dots, M$ as otherwise the inverse Fourier transform of \hat{h} would not yield a real-valued function. Under this condition, we have that

$$h(x) = \sum_{i=1}^M w_i \mathcal{N}(\mu_i, \Sigma_i - \tau^2 I)$$

and since, by assumption, there exist either an index i such that $\Sigma_i \neq \sigma_i I$, or a pair $\{i, j\}$ such that $\Sigma_i \neq \Sigma_j$, then h cannot be compactly supported for any choice of τ^2 .

Note that when neither condition (i) nor (ii) holds, we can take $h = \sum_{i=1}^M w_i \delta_{\mu_i}$, where δ_{μ_i} denotes a Dirac delta function centred at μ_i . \square

One final implication of assumption **A4** is provided in the following proposition.

Proposition B.7 (Proposition 1 [Ma et al., 2019]). *If π_{data} satisfies assumption **A4**, then it has a finite log-Sobolev constant $C_{\text{LSI}, \pi_{\text{data}}} \leq \frac{2}{M_\pi} e^{16L_\pi r^2}$.*

Proof. Recall that under **A4** we have that $\pi_{\text{data}} \propto e^{-V_\pi}$ satisfies

$$\inf_{\|x\| \geq r} \nabla^2 V_\pi \succcurlyeq M_\pi I, \quad -L_\pi I \preccurlyeq \nabla^2 V_\pi \preccurlyeq L_\pi I.$$

By Ma et al. [2019, Lemma 1], there exists $\tilde{V}_\pi \in C^1(\mathbb{R}^d)$ such that \tilde{V}_π is $M_\pi/2$ strongly convex on \mathbb{R}^d and has a Hessian $\nabla^2 \tilde{V}_\pi$ that exists everywhere on \mathbb{R}^d . Therefore, using the Bakry-Émery criterion [Bakry and Émery, 1985], $\tilde{\pi} \propto e^{-\tilde{V}_\pi}$ satisfies log-Sobolev inequality with constant $C_{\text{LSI}, \tilde{\pi}} \leq 2/M_\pi$. Moreover, Lemma 1 in Ma et al. [2019] also guarantees that

$$\sup \left(\tilde{V}_\pi(x) - V_\pi \right) - \inf \left(\tilde{V}_\pi(x) - V_\pi \right) \leq 16L_\pi r^2.$$

Applying the Holley-Stroock perturbation principle [Holley and Stroock, 1987], it follows that π_{data} has a finite log-Sobolev constant satisfying

$$C_{\text{LSI}, \pi_{\text{data}}} \leq \frac{2}{M_\pi} e^{16L_\pi r^2}.$$

□

B.2 Comments on Assumption A5

We show below that assumption **A5** is satisfied by multivariate Student's t distributions of the form

$$\pi_{\text{data}}(x) = C_\pi \left(1 + \frac{1}{\alpha} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right)^{-(\alpha+d)/2}, \quad x \in \mathbb{R}^d,$$

where the covariance matrix Σ is a positive definite matrix satisfying $\sigma_{\min} I \preccurlyeq \Sigma \preccurlyeq \sigma_{\max} I$ and $\alpha > 0$ denotes the degrees of freedom. The Hessian of the potential has the following expression

$$-\nabla^2 \log \pi_{\text{data}}(x) = \frac{\alpha + d}{\alpha} \frac{\Sigma^{-1}}{1 + \frac{1}{\alpha} (x - \mu)^\top \Sigma^{-1} (x - \mu)} - \frac{2(\alpha + d)}{\alpha^2} \frac{\Sigma^{-1} (x - \mu) (x - \mu)^\top \Sigma^{-1}}{\left(1 + \frac{1}{\alpha} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right)^2}.$$

The matrix $(x - \mu)(x - \mu)^\top$ is positive semidefinite and satisfies

$$0 \preccurlyeq (x - \mu)(x - \mu)^\top \preccurlyeq \|x - \mu\|^2 I.$$

Since the eigenvalues of the product of symmetric positive semidefinite matrices satisfy the following

$$\lambda_{\min}(ABC) \geq \lambda_{\min}(A)\lambda_{\min}(B)\lambda_{\min}(C), \quad \lambda_{\max}(ABC) \leq \lambda_{\max}(A)\lambda_{\max}(B)\lambda_{\max}(C),$$

we have that

$$0 \preccurlyeq \Sigma^{-1} (x - \mu) (x - \mu)^\top \Sigma^{-1} \preccurlyeq \sigma_{\min}^{-2} \|x - \mu\|^2 I.$$

This leads to

$$\begin{aligned} -\nabla^2 \log \pi_{\text{data}}(x) &\preccurlyeq \frac{\alpha + d}{\alpha} \frac{\Sigma^{-1}}{1 + \frac{1}{\alpha} (x - \mu)^\top \Sigma^{-1} (x - \mu)} \preccurlyeq \frac{(\alpha + d)\sigma_{\min}^{-1}}{\alpha} \frac{I}{1 + \frac{1}{\alpha} \|S^{-1/2}U(x - \mu)\|^2} \\ &\preccurlyeq \frac{(\alpha + d)\sigma_{\min}^{-1}}{\alpha + \sigma_{\max}^{-1} \|x - \mu\|^2} I, \end{aligned}$$

where we have used that $\Sigma^{-1} = U^\top S^{-1}U$, U being an orthogonal matrix and S being diagonal. On the other hand,

$$\begin{aligned} -\nabla^2 \log \pi_{\text{data}}(x) &\succcurlyeq \frac{(\alpha + d)\sigma_{\max}^{-1}}{\alpha + \sigma_{\min}^{-1}\|x - \mu\|^2} I - \frac{2(\alpha + d)\sigma_{\min}^{-2}}{\sigma_{\max}^{-1}} \frac{\sigma_{\max}^{-1}\|x - \mu\|^2}{(\alpha + \sigma_{\max}^{-1}\|x - \mu\|^2)^2} I \\ &\succcurlyeq \frac{(\alpha + d)\sigma_{\max}^{-1}}{\alpha + \sigma_{\min}^{-1}\|x - \mu\|^2} I - \frac{2(\alpha + d)\sigma_{\min}^{-2}}{\sigma_{\max}^{-1}} \frac{I}{\alpha + \sigma_{\max}^{-1}\|x - \mu\|^2}. \\ &\succcurlyeq -\frac{2(\alpha + d)\sigma_{\min}^{-2}}{\sigma_{\max}^{-1}} \frac{I}{\alpha + \sigma_{\max}^{-1}\|x - \mu\|^2}, \end{aligned}$$

which concludes that π_{data} satisfies assumption **A5**.

B.3 Proof of Lemma 3.2

Before jumping into the proof of Lemma 3.2, we provide a generalisation of the Poincaré inequality that will be required later.

Lemma B.8. *Let π be a probability distribution in \mathbb{R}^d with finite Poincaré constant $C_{\text{PI},\pi}$. Then, for all functions $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$, $f \in L^2(\pi)$, it holds that*

$$\text{Cov}_{X \sim \pi}[f(X)] \preccurlyeq C_{\text{PI},\pi} \mathbb{E}_\pi[\nabla f(X) \nabla f(X)^\top].$$

Proof. Given the function $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$, $f \in L^2(\pi)$. For every unit vector $u \in \mathbb{R}^d$, consider the function $u^\top f : \mathbb{R}^d \rightarrow \mathbb{R}$. Since π satisfies a Poincaré inequality we have

$$u^\top \text{Cov}_{X \sim \pi}[f(X)] u = \text{Cov}_{X \sim \pi}[u^\top f(X)] \leq C_{\text{PI},\pi} \mathbb{E}_{X \sim \pi} \left[\left\| \sum_i u_i \nabla f_i(X) \right\|^2 \right] = C_{\text{PI},\pi} u^\top \mathbb{E}_\pi[\nabla f(X) \nabla f(X)^\top] u,$$

which concludes the proof. \square

Proof of Lemma 3.2 under assumption A4. To simplify notation in the proof, we denote π_{data} as π . Let $\mu_t \propto e^{-U_t}$, our aim is to show that ∇U_t is Lipschitz, to do so we are going to show that the Hessian $\nabla^2 U_t$ is bounded for all $t \in [0, T]$. Using that if f and g are differentiable functions then $\nabla(f * g) = (\nabla f) * g = f * (\nabla g)$, we have the following expressions for ∇U_t and $\nabla^2 U_t$

$$\nabla U_t(x) = \frac{1}{\sigma^2(1 - \lambda_t)} (x - \mathbb{E}_{\rho_{t,x}(y)}[Y]) = \frac{1}{\sqrt{\lambda_t}} \mathbb{E}_{\rho_{t,x}(y)} \left[\nabla V_\pi \left(\frac{Y}{\sqrt{\lambda_t}} \right) \right] \quad (11)$$

$$\nabla^2 U_t(x) = \frac{1}{\sigma^2(1 - \lambda_t)} \left(I - \frac{1}{\sigma^2(1 - \lambda_t)} \text{Cov}_{\rho_{t,x}}[Y] \right) \quad (12)$$

$$\nabla^2 U_t(x) = \frac{1}{\lambda_t} \left(\mathbb{E}_{\rho_{t,x}} \left[\nabla^2 V_\pi \left(\frac{Y}{\sqrt{\lambda_t}} \right) \right] - \text{Cov}_{\rho_{t,x}} \left[\nabla V_\pi \left(\frac{Y}{\sqrt{\lambda_t}} \right) \right] \right) \quad (13)$$

$$\nabla^2 U_t(x) = \frac{1}{\sigma^2(1 - \lambda_t)\lambda_t} \text{Cov}_{\rho_{t,x}} \left[Y, \nabla V_\pi \left(\frac{Y}{\sqrt{\lambda_t}} \right) \right], \quad (14)$$

where $\rho_{t,x}(y) \propto e^{-V_\pi(y/\sqrt{\lambda_t}) - \|x - y\|^2/(2\sigma^2(1 - \lambda_t))}$. It is worth mentioning that as $t \rightarrow 0$, $\rho_{t,x}$ tends to a Dirac delta centred at 0, and as $t \rightarrow T$, it approaches a Dirac delta centred at x , both of which have zero variance. Note that (12) and (13) admit the following upper bounds

$$\nabla^2 U_t(x) \preccurlyeq \frac{1}{\sigma^2(1 - \lambda_t)} I \quad (15)$$

$$\nabla^2 U_t(x) \preccurlyeq \frac{L_\pi}{\lambda_t} I, \quad (16)$$

where we have used that the covariance matrix is positive semidefinite. To find a lower bound for $\nabla^2 U_t$ we need to upper bound $\text{Cov}_{\rho_{t,x}}[Y]$ and $\text{Cov}_{\rho_{t,x}}[\nabla V_\pi(Y/\sqrt{\lambda_t})]$. Observe that if $\rho_{t,x}$ satisfies the Poincaré

inequality with constant C_{PI, ρ_t} independent of x , then using the generalisation of the Poincaré inequality for vector-valued random variables given in Lemma B.8, we have

$$\begin{aligned}\text{Cov}_{\rho_{t,x}}[Y] &\leq C_{\text{PI}, \rho_t} I \\ \text{Cov}_{\rho_{t,x}}\left[\nabla V_\pi\left(\frac{Y}{\sqrt{\lambda_t}}\right)\right] &\leq \frac{C_{\text{PI}, \rho_t}}{\lambda_t} \mathbb{E}_{\rho_{t,x}}\left[\nabla^2 V_\pi\left(\frac{Y}{\sqrt{\lambda_t}}\right) \nabla^2 V_\pi\left(\frac{Y}{\sqrt{\lambda_t}}\right)^\top\right] \leq \frac{C_{\text{PI}, \rho_t} L_\pi^2}{\lambda_t} I.\end{aligned}$$

This implies that for each t we have that

$$\nabla^2 U_t(x) \leq \min\left\{\frac{1}{\sigma^2(1-\lambda_t)}, \frac{L_\pi}{\lambda_t}\right\} I =: a_t I \quad (17)$$

$$\nabla^2 U_t(x) \geq \max\left\{\frac{1}{\sigma^2(1-\lambda_t)}\left(1 - \frac{C_{\text{PI}, \rho_t}}{\sigma^2(1-\lambda_t)}\right), -\frac{L_\pi}{\lambda_t}\left(1 + \frac{C_{\text{PI}, \rho_t} L_\pi}{\lambda_t}\right)\right\} I =: b_t I. \quad (18)$$

Therefore, the Lipschitz constant L_t satisfies the following

$$L_t \leq \max\{a_t, |b_t|\}. \quad (19)$$

To conclude the proof we need to check that C_{PI, ρ_t} is independent of x , since otherwise the Poincaré constant can get arbitrarily large and we will not have a meaningful bound for the Hessian. Note that if we denote $\rho_{t,x} \propto e^{-V_{\rho_{t,x}}}$ we have that

$$\left(-\frac{L_\pi}{\lambda_t} + \frac{1}{\sigma^2(1-\lambda_t)}\right) I \leq \nabla^2 V_{\rho_{t,x}}(y) = \frac{1}{\lambda_t} \nabla^2 V_\pi\left(\frac{y}{\sqrt{\lambda_t}}\right) + \frac{I}{\sigma^2(1-\lambda_t)} \leq \left(\frac{L_\pi}{\lambda_t} + \frac{1}{\sigma^2(1-\lambda_t)}\right) I. \quad (20)$$

Note that if $\left(-\frac{L_\pi}{\lambda_t} + \frac{1}{\sigma^2(1-\lambda_t)}\right) > 0$, then $V_{\rho_{t,x}}$ is strongly convex and using Bakry-Émery criterion [Bakry and Émery, 1985] we have that $\rho_{t,x}$ satisfies a log-Sobolev inequality which implies a Poincaré inequality with the same constant. Thus, for t such that

$$\tilde{\lambda} = \frac{\sigma^2 L_\pi}{1 + \sigma^2 L_\pi} < \lambda_t \leq 1,$$

$\rho_{t,x}$ satisfies the Poincaré inequality with constant $\left(\frac{1}{\sigma^2(1-\lambda_t)} - \frac{L_\pi}{\lambda_t}\right)^{-1}$ independent of x , which tends to 0 as λ_t tends to 1.

On the other hand, using that the potential V_π is strongly convex outside of a ball of radius r , we have that for $\|y\| > \sqrt{\lambda_t}r$

$$\left(\frac{M_\pi}{\lambda_t} + \frac{1}{\sigma^2(1-\lambda_t)}\right) I \leq \nabla^2 V_{\rho_{t,x}}(y) \leq \left(\frac{L_\pi}{\lambda_t} + \frac{1}{\sigma^2(1-\lambda_t)}\right) I. \quad (21)$$

Equations (20)-(21) imply that for $\lambda_t > 0$, $\nabla \log \rho_{t,x}$ is Lipschitz continuous and $\rho_{t,x}$ is strongly log-concave outside of a ball of radius $r_t = \sqrt{\lambda_t}r$. Similarly to the proof in Lemma B.7, the existence of a smooth strongly convex approximation of $V_{\rho_{t,x}}$ and the Holley-Stroock perturbation lemma [Holley and Stroock, 1987] imply that $\rho_{t,x}$ satisfies a log-Sobolev inequality and hence a Poincaré inequality with constant

$$C_{\text{PI}, \rho_t} \leq 2 \left(\frac{M_\pi}{\lambda_t} + \frac{1}{\sigma^2(1-\lambda_t)}\right)^{-1} e^{16(L_\pi + \frac{\lambda_t}{\sigma^2(1-\lambda_t)})r^2}, \quad (22)$$

independent of x . Observe that when λ_t tends to 0 the upper bound of the Poincaré constant C_{PI, ρ_t} also tends to 0. Therefore, for $\lambda_t \in [0, \tilde{\lambda}]$ C_{PI, ρ_t} is bounded by (22), while for $\lambda_t \in (\tilde{\lambda}, 1]$, C_{PI, ρ_t} is bounded by

$$C_{\text{PI}, \rho_t} \leq \min\left\{\left(\frac{1}{\sigma^2(1-\lambda_t)} - \frac{L_\pi}{\lambda_t}\right)^{-1}, 2 \left(\frac{M_\pi}{\lambda_t} + \frac{1}{\sigma^2(1-\lambda_t)}\right)^{-1} e^{16(L_\pi + \frac{\lambda_t}{\sigma^2(1-\lambda_t)})r^2}\right\}.$$

□

It is important to note that, in the proof of Lemma 3.2 under assumption **A5** below, we only rely on the lower bound of the Hessian from **A5**. If we omit the upper bound on the Hessian, **A5** becomes a generalisation of **A4**. The stronger assumption in **A4** results in tighter bounds for the Lipschitz constants along the diffusion path, thereby improving upon those in Gao et al. [2024], as our bounds are non-vacuous for all $t \in [0, T]$.

*Proof of Lemma 3.2 under assumption **A5**.* Let $\mu_t \propto e^{-U_t}$. Note that the expressions for the Hessian of U_t given in (12)-(14) remain valid in this case. Consequently, the bounds provided in (15)-(16) also hold.

On the other hand, if $\rho_{t,x}(y) \propto e^{-V_\pi(y/\sqrt{\lambda_t}) - \|x-y\|^2/(2\sigma^2(1-\lambda_t))}$ satisfies a Poincaré inequality with constant C_{PI,ρ_t} independent of x , then the bounds (17)-(19) are valid. Therefore, to conclude that $\nabla \log \mu_t$ is Lipschitz we need to show that C_{PI,ρ_t} is independent of x . Note that under assumption **A5**, we have that

$$\left(-\frac{L_\pi}{\lambda_t} + \frac{1}{\sigma^2(1-\lambda_t)}\right) I \preccurlyeq \nabla^2 V_{\rho_{t,x}}(y) \preccurlyeq \left(\frac{L_\pi}{\lambda_t} + \frac{1}{\sigma^2(1-\lambda_t)}\right) I,$$

and for y such that $\|y\| > \sqrt{\lambda_t}r$

$$\left(-\frac{1}{\lambda_t \alpha_1 + \alpha_2 \|y\|^2} + \frac{1}{\sigma^2(1-\lambda_t)}\right) I \preccurlyeq \nabla^2 V_{\rho_{t,x}}(y) \preccurlyeq \left(\frac{1}{\lambda_t \beta_1 + \beta_2 \|y\|^2} + \frac{1}{\sigma^2(1-\lambda_t)}\right) I.$$

Note that if $1 \geq \lambda_t > \frac{\sigma^2 L_\pi}{1+\sigma^2 L_\pi} = \tilde{\lambda}$, then $V_{\rho_{t,x}}$ is strongly convex and $\rho_{t,x}$ satisfies a Poincaré inequality with constant $\left(\frac{1}{\sigma^2(1-\lambda_t)} - \frac{L_\pi}{\lambda_t}\right)^{-1}$, which tends to 0 as λ_t tends to 1. On the other hand, define

$$\tilde{r}_t^2 = \max\{\lambda_t r^2, (2\sigma^2(1-\lambda_t) - \lambda_t \alpha_1) \alpha_2^{-1}\}.$$

We have that $V_{\rho_{t,x}}$ is strongly convex outside of a ball of radius \tilde{r}_t . That is, for $\|y\| > \tilde{r}_t$, it follows that $\nabla^2 V_{\rho_{t,x}} \succcurlyeq I/(2\sigma^2(1-\lambda_t))$. Therefore, as in Lemma B.7, leveraging the existence of a smooth strongly convex approximation of $V_{\rho_{t,x}}$ and Holley-Stroock perturbation lemma [Holley and Stroock, 1987], $\rho_{t,x}$ satisfies a Poincaré inequality with constant

$$C_{\text{PI},\rho_t} \leq 2 \left(\frac{1}{2\sigma^2(1-\lambda_t)}\right)^{-1} e^{16\left(\frac{L_\pi}{\lambda_t} + \frac{1}{\sigma^2(1-\lambda_t)}\right)\tilde{r}_t^2}, \quad (23)$$

independent of x . Therefore, for $\lambda_t \in (0, \tilde{\lambda}]$ C_{PI,ρ_t} is bounded by (23), while for $\lambda_t \in (\tilde{\lambda}, 1]$, C_{PI,ρ_t} is bounded by

$$C_{\text{PI},\rho_t} \leq \min \left\{ \left(\frac{1}{\sigma^2(1-\lambda_t)} - \frac{L_\pi}{\lambda_t}\right)^{-1}, 2 \left(\frac{1}{2\sigma^2(1-\lambda_t)}\right)^{-1} e^{16\left(\frac{L_\pi}{\lambda_t} + \frac{1}{\sigma^2(1-\lambda_t)}\right)\tilde{r}_t^2} \right\}.$$

Finally, observe that if $\lambda_0 = 0$, then $\mu_0 = \nu = \mathcal{N}(0, \sigma^2 I)$, which implies that $\nabla \log \mu_0$ is Lipschitz continuous. This concludes that $\nabla \log \mu_t$ is Lipschitz continuous for all $t \in [0, T]$. \square

B.4 Proof of Lemma 3.3

Proof. When the schedule satisfies $\max_{t \in [0, T]} |\partial_t \log \lambda_t| \leq C_\lambda$, we consider the reparametrised version of μ_t in terms of the schedule λ_t , denoted as $\tilde{\mu}_\lambda$ and let $X_\lambda \sim \tilde{\mu}_\lambda$ and $X_{\lambda+\delta} \sim \tilde{\mu}_{\lambda+\delta}$. Recall that

$$X_\lambda = \sqrt{\lambda} X + \sqrt{1-\lambda} \sigma Z \quad (24)$$

where $X \sim \pi_{\text{data}}$ and $Z \sim \mathcal{N}(0, I)$ are independent from each other. We introduce a new random variable \tilde{X} , independent from Z , that follows a Gaussian distribution, $\tilde{X} \sim N(0, \sigma_\pi^2 I)$, satisfying

$$\sigma_\pi = \arg \min_{\hat{\sigma} \geq \sigma} W_2(\mathcal{L}(X), \mathcal{L}(Y_{\hat{\sigma}})), \quad \text{where } Y_{\hat{\sigma}} \sim \mathcal{N}(0, \hat{\sigma}^2 I), \quad Z \perp\!\!\!\perp Y_{\hat{\sigma}}.$$

Furthermore, we select \tilde{X} to be the specific random variable that attains the minimal coupling with X , that is, $W_2^2(\mathcal{L}(X), \mathcal{L}(\tilde{X})) = \mathbb{E}[\|X - \tilde{X}\|^2]$. Using the random variable \tilde{X} , we can rewrite (24) as

$$\begin{aligned} X_\lambda &= \sqrt{\lambda}(X - \tilde{X}) + \sqrt{\lambda}\tilde{X} + \sqrt{1-\lambda}\sigma Z \\ &\stackrel{d}{=} \sqrt{\lambda}(X - \tilde{X}) + \sqrt{\lambda\sigma_\pi^2 + (1-\lambda)\sigma^2}Y, \end{aligned}$$

where $Y \sim \mathcal{N}(0, I)$. The Wasserstein-2 distance between $\tilde{\mu}_\lambda$ and $\tilde{\mu}_{\lambda+\delta}$ is given by

$$\begin{aligned} W_2^2(\tilde{\mu}_\lambda, \tilde{\mu}_{\lambda+\delta}) &\leq \mathbb{E}[\|X_\lambda - X_{\lambda+\delta}\|^2] \\ &\leq 2\mathbb{E}\left[\left\|(\sqrt{\lambda} - \sqrt{\lambda+\delta})(X - \tilde{X})\right\|^2\right] + 2\mathbb{E}\left[\left\|\left(\sqrt{\lambda\sigma_\pi^2 + (1-\lambda)\sigma^2} - \sqrt{(\lambda+\delta)\sigma_\pi^2 + (1-\lambda-\delta)\sigma^2}\right)Y\right\|^2\right] \\ &= 2(\sqrt{\lambda} - \sqrt{\lambda+\delta})^2\mathbb{E}[\|X - \tilde{X}\|^2] + 2\left(\sqrt{\lambda\sigma_\pi^2 + (1-\lambda)\sigma^2} - \sqrt{(\lambda+\delta)\sigma_\pi^2 + (1-\lambda-\delta)\sigma^2}\right)^2 d. \end{aligned}$$

Using the definition of the metric derivative we have

$$|\dot{\mu}|_\lambda^2 = \lim_{\delta \rightarrow 0} \frac{W_2^2(\tilde{\mu}_\lambda, \tilde{\mu}_{\lambda+\delta})}{\delta^2} \leq \frac{\mathbb{E}[\|X - \tilde{X}\|^2]}{2\lambda} + \frac{(\sigma_\pi^2 - \sigma^2)^2}{2(\lambda\sigma_\pi^2 + \sigma^2(1-\lambda))}d = \frac{\mathbb{E}[\|X - \tilde{X}\|^2]}{2\lambda} + \frac{(\sigma_\pi^2 - \sigma^2)^2}{2(\sigma^2 + \lambda(\sigma_\pi^2 - \sigma^2))}d.$$

Since $\mu_t = \tilde{\mu}_{\lambda_t}$, we have that $|\dot{\mu}|_t = |\dot{\tilde{\mu}}|_\lambda |\partial_t \lambda_t|$. Using the assumption on the schedule we have the following expression for the action

$$\begin{aligned} \mathcal{A}_\lambda(\mu) &= \int_0^T |\dot{\mu}|_t^2 dt = \int_0^T |\dot{\tilde{\mu}}|_\lambda^2 |\partial_t \lambda_t|^2 dt \\ &\leq \int_0^T \left(\frac{\mathbb{E}[\|X - \tilde{X}\|^2]}{2\lambda_t} + \frac{(\sigma_\pi^2 - \sigma^2)^2}{2(\sigma^2 + \lambda_t(\sigma_\pi^2 - \sigma^2))}d \right) |\partial_t \lambda_t|^2 dt \\ &= \int_0^T \left(\frac{\mathbb{E}[\|X - \tilde{X}\|^2]}{2} + \frac{(\sigma_\pi^2 - \sigma^2)^2}{2(\sigma^2/\lambda_t + \sigma_\pi^2 - \sigma^2)}d \right) |\partial_t \log \lambda_t| |\partial_t \lambda_t| dt \\ &\leq C_\lambda \int_0^T \left(\frac{\mathbb{E}[\|X - \tilde{X}\|^2]}{2} + \frac{(\sigma_\pi^2 - \sigma^2)^2}{2(\sigma^2/\lambda_t + \sigma_\pi^2 - \sigma^2)}d \right) |\partial_t \lambda_t| dt \\ &= C_\lambda \int_{\lambda_0}^1 \left(\frac{\mathbb{E}[\|X - \tilde{X}\|^2]}{2} + \frac{(\sigma_\pi^2 - \sigma^2)^2}{2(\sigma^2/\lambda + (\sigma_\pi^2 - \sigma^2))}d \right) d\lambda \\ &\leq \frac{C_\lambda}{2} \left(\mathbb{E}[\|X - \tilde{X}\|^2] + d \left(\sigma_\pi^2 - \sigma^2 + \sigma^2 \log \frac{\sigma^2}{\sigma_\pi^2} \right) \right) \\ &\leq \frac{C_\lambda}{2} (2\mathbb{E}[\|X\|^2] + d(3\sigma_\pi^2 - \sigma^2)), \end{aligned}$$

where in the last line we have used that $\sigma_\pi \geq \sigma$. Note that by setting $\sigma = \sigma_\pi$, the second term in the penultimate expression cancels out, resulting in

$$\mathcal{A}_\lambda(\mu) \leq \frac{C_\lambda}{2} \mathbb{E}[\|X - \tilde{X}\|^2],$$

where we chose \tilde{X} such that $\mathbb{E}[\|X - \tilde{X}\|^2] = W_2^2(\mathcal{L}(X), \mathcal{L}(\tilde{X}))$ is minimised.

On the other hand, if the schedule satisfies $\max_{t \in [0, T]} \left| \frac{\partial_t \lambda_t}{\sqrt{\lambda_t(1-\lambda_t)}} \right| \leq C_\lambda$, the Wasserstein-2 distance

between $\tilde{\mu}_\lambda$ and $\tilde{\mu}_{\lambda+\delta}$ is given by

$$\begin{aligned} W_2^2(\tilde{\mu}_\lambda, \tilde{\mu}_{\lambda+\delta}) &\leq \mathbb{E} [\|X_\lambda - X_{\lambda+\delta}\|^2] \\ &= \mathbb{E} \left[\left\| (\sqrt{\lambda} - \sqrt{\lambda+\delta}) X \right\|^2 \right] + \mathbb{E} \left[\left\| (\sqrt{1-\lambda} - \sqrt{1-\lambda-\delta}) \sigma Z \right\|^2 \right] \\ &= (\sqrt{\lambda} - \sqrt{\lambda+\delta})^2 \mathbb{E} [\|X\|^2] + (\sqrt{1-\lambda} - \sqrt{1-\lambda-\delta})^2 \sigma^2 d. \end{aligned}$$

Using the definition of the metric derivative we have

$$|\dot{\mu}|_\lambda^2 = \lim_{\delta \rightarrow 0} \frac{W_2^2(\tilde{\mu}_\lambda, \tilde{\mu}_{\lambda+\delta})}{\delta^2} \leq \frac{\mathbb{E} [\|X\|^2]}{4\lambda} + \frac{\sigma^2 d}{4(1-\lambda)}.$$

Therefore, we have the following expression for the action

$$\begin{aligned} \mathcal{A}_\lambda(\mu) &= \int_0^T |\dot{\mu}|_t^2 dt = \int_0^T |\dot{\mu}|_\lambda^2 |\partial_t \lambda_t|^2 dt \\ &\lesssim \int_0^T \left(\frac{\mathbb{E} [\|X\|^2]}{4\lambda_t} + \frac{\sigma^2}{4(1-\lambda_t)} d \right) |\partial_t \lambda_t|^2 dt \\ &= \int_0^T \left(\frac{\mathbb{E} [\|X\|^2] \sqrt{1-\lambda_t}}{4\sqrt{\lambda_t}} + \frac{\sigma^2 \sqrt{\lambda_t}}{4\sqrt{1-\lambda_t}} d \right) \left| \frac{\partial_t \lambda_t}{\sqrt{\lambda_t(1-\lambda_t)}} \right| |\partial_t \lambda_t| dt \\ &\leq C_\lambda \int_0^T \left(\frac{\mathbb{E} [\|X\|^2] \sqrt{1-\lambda_t}}{4\sqrt{\lambda_t}} + \frac{\sigma^2 \sqrt{\lambda_t}}{4\sqrt{1-\lambda_t}} d \right) |\partial_t \lambda_t| dt \\ &\leq C_\lambda \int_0^1 \left(\frac{\mathbb{E} [\|X\|^2] \sqrt{1-\lambda}}{4\sqrt{\lambda}} + \frac{\sigma^2 \sqrt{\lambda}}{4\sqrt{1-\lambda}} d \right) d\lambda \\ &\leq \frac{C_\lambda \pi}{8} (\mathbb{E} [\|X\|^2] + \sigma^2 d). \end{aligned}$$

□

B.5 Proof of Theorem 3.4

Proof. First, consider a modified version of the DALMC algorithm with exact scores, that is,

$$X_{l+1} = X_l + h_l \nabla \log \hat{\mu}_l(X_l) + \sqrt{2h_l} \xi_l, \quad (25)$$

where $h_l > 0$ is the step size, $\xi_k \sim \mathcal{N}(0, I)$, $\hat{\mu}_t = \mu_{\kappa t}$, $l \in \{1, \dots, M\}$ and $0 = t_0 < t_1 < \dots < t_M = T/\kappa$ is a discretisation of the interval $[0, T/\kappa]$. Let \mathbb{Q} be the path measure associated with the continuous-time interpolation of this auxiliary algorithm which corresponds to the SDE

$$dX_t = \nabla \log \hat{\mu}_{t-}(X_{t-}) dt + \sqrt{2} dB_t, \quad t \in [0, T/\kappa],$$

where given a discretisation of the interval $[0, T/\kappa]$, $0 = t_0 < t_1 < \dots < t_M = T/\kappa$, we define $t_- := t_{l-1}$ when $t \in [t_{l-1}, t_l)$ for $l = 1, \dots, M$. On the other hand, let \mathbb{P} be the path measure corresponding to the following reference SDE

$$dY_t = (\nabla \log \hat{\mu}_t + v_t)(Y_t) dt + \sqrt{2} dB_t, \quad t \in [0, T/\kappa].$$

The vector field $v = (v_t)_{t \in [0, T/\kappa]}$ is designed such that $Y_t \sim \hat{\mu}_t$ for all $t \in [0, T/\kappa]$. Using the Fokker-Planck equation, we have that

$$\partial_t \hat{\mu}_t = \nabla \cdot (\hat{\mu}_t (\nabla \log \hat{\mu}_t + v_t)) + \Delta \hat{\mu}_t = -\nabla \cdot (\hat{\mu}_t v_t), \quad t \in [0, T/\kappa].$$

This implies that v_t satisfies the continuity equation and hence generates the curve of probability measures $(\hat{\mu}_t)_t$. Leveraging Lemma A.1, we choose v to be the one that minimises the $L^2(\hat{\mu}_t)$ norm, resulting in

$\|v_t\|_{L^2(\hat{\mu}_t)} = \left| \dot{\hat{\mu}}_t \right|$ being the metric derivative. Using the form of Girsanov's theorem given in Lemma A.2 we have

$$\begin{aligned} \text{KL}(\mathbb{P} \parallel \mathbb{Q}) &= \frac{1}{4} \int_0^{T/\kappa} \mathbb{E}_{\mathbb{P}} \left[\|\nabla \log \hat{\mu}_t(X_t) - \nabla \log \hat{\mu}_{t-}(X_{t-}) + v_t(X_t)\|^2 \right] dt \\ &\lesssim \int_0^{T/\kappa} \mathbb{E}_{\mathbb{P}} \left[\|\nabla \log \hat{\mu}_t(X_t) - \nabla \log \hat{\mu}_t(X_{t-})\|^2 \right] dt + \int_0^{T/\kappa} \|v_t\|_{L^2(\hat{\mu}_t)}^2 dt \\ &\quad + \int_0^{T/\kappa} \mathbb{E}_{\mathbb{P}} \left[\|\nabla \log \hat{\mu}_t(X_{t-}) - \nabla \log \hat{\mu}_{t-}(X_{t-})\|^2 \right] dt \\ &\leq \sum_{l=1}^M \int_{t_{l-1}}^{t_l} L_{\kappa t}^2 \mathbb{E}_{\mathbb{P}} \left[\|X_t - X_{t-}\|^2 \right] dt + \int_0^{T/\kappa} \|v_t\|_{L^2(\hat{\mu}_t)}^2 dt + \int_0^{T/\kappa} \mathbb{E}_{\mathbb{P}} \left[\left\| \nabla \log \frac{\hat{\mu}_t(X_{t-})}{\hat{\mu}_{t-}(X_{t-})} \right\|^2 \right] dt, \quad (26) \end{aligned}$$

where we have used that $\nabla \log \hat{\mu}_t$ is Lipschitz with constant $L_{\kappa t}$. First, we bound the change in the score function $\mathbb{E}_{\mathbb{P}} \left[\left\| \nabla \log \frac{\hat{\mu}_t(X_{t-})}{\hat{\mu}_{t-}(X_{t-})} \right\|^2 \right]$. Let $t \geq t_-$, we can write

$$\hat{\mu}_{t-} = T \sqrt{\frac{\lambda_{\kappa t}}{\lambda_{\kappa t-}}} \# \hat{\mu}_t * \mathcal{N} \left(0, \left(\sqrt{1 - \lambda_{\kappa t-}} - \sqrt{\frac{(1 - \lambda_{\kappa t})\lambda_{\kappa t-}}{\lambda_{\kappa t}}} \right)^2 \sigma^2 I \right),$$

where the pushforward $T_\alpha \#$ is defined as $T_\alpha \# \mu(x) = \alpha^d \mu(\alpha x)$. Using Lee et al. [2022, Lemma C.12] we have

$$\left\| \nabla \log \frac{\hat{\mu}_t(X_{t-})}{\hat{\mu}_{t-}(X_{t-})} \right\|^2 \lesssim L_{\kappa t}^2 d \gamma_t + L_{\kappa t}^2 \left(\sqrt{\frac{\lambda_{\kappa t}}{\lambda_{\kappa t-}}} - 1 \right)^2 \|X_{t-}\|^2 + \left(\sqrt{\frac{\lambda_{\kappa t}}{\lambda_{\kappa t-}}} - 1 + L_{\kappa t} \gamma_t \right)^2 \|\nabla \log \hat{\mu}_t(X_{t-})\|^2,$$

where

$$\gamma_t = \left(\sqrt{1 - \lambda_{\kappa t-}} - \sqrt{\frac{(1 - \lambda_{\kappa t})\lambda_{\kappa t-}}{\lambda_{\kappa t}}} \right)^2 \lesssim 1 - \frac{\lambda_{\kappa t-}}{\lambda_{\kappa t}}.$$

Let C_λ introduced in A6, we have

$$\gamma_t \leq 1 - \frac{\lambda_{\kappa t-}}{\lambda_{\kappa t}} \lesssim C_\lambda h_l, \quad \left(\sqrt{\frac{\lambda_{\kappa t}}{\lambda_{\kappa t-}}} - 1 \right)^2 \lesssim C_\lambda^2 h_l^2, \quad \left(\sqrt{\frac{\lambda_{\kappa t}}{\lambda_{\kappa t-}}} - 1 + L_{\kappa t} \gamma_t \right)^2 \lesssim h_l^2 L_{\kappa t}^2.$$

In addition, by choosing an appropriate step size, as will be shown in Corollary 3.5, we can bound $h_l^2 L_{\kappa t}^2 \lesssim 1$.

Given that $X_t = \sqrt{\lambda_t} X + \sqrt{1 - \lambda_t} \sigma Z$ for $X_t \sim \hat{\mu}_t$, we derive the following moment bound

$$\mathbb{E}_{\mathbb{P}} \left[\|X_{t-}\|^2 \right] = \mathbb{E}_{\mathbb{P}} \left[\left\| \sqrt{\lambda_{\kappa t-}} X + \sqrt{1 - \lambda_{\kappa t-}} Z \right\|^2 \right] = \lambda_{\kappa t-} \mathbb{E}_{\pi_{\text{data}}} \left[\|X\|^2 \right] + (1 - \lambda_{\kappa t-}) \sigma^2 d \lesssim \mathbb{E}_{\pi_{\text{data}}} \left[\|X\|^2 \right] + d.$$

To bound $\mathbb{E}_{\mathbb{P}} \left[\|\nabla \log \hat{\mu}_t(X_{t-})\|^2 \right]$, recall from (17) that $\mu_t \propto e^{-U_t}$ satisfies $\nabla^2 U_t \preceq L_t I$. Therefore, using Chewi [2024, Lemma 4.0.1] it holds that

$$\begin{aligned} \mathbb{E}_{\mathbb{P}} \left[\|\nabla \log \hat{\mu}_t(X_{t-})\|^2 \right] &\leq \mathbb{E}_{\mathbb{P}} \left[\|\nabla \log \hat{\mu}_t(X_t)\|^2 \right] + \mathbb{E}_{\mathbb{P}} \left[\|\nabla \log \hat{\mu}_t(X_t) - \nabla \log \hat{\mu}_t(X_{t-})\|^2 \right] \\ &\leq L_{\kappa t} d + L_{\kappa t}^2 \mathbb{E}_{\mathbb{P}} \left[\|X_t - X_{t-}\|^2 \right]. \end{aligned}$$

This implies that

$$\mathbb{E}_{\mathbb{P}} \left[\left\| \nabla \log \frac{\hat{\mu}_t(X_{t-})}{\hat{\mu}_{t-}(X_{t-})} \right\|^2 \right] \lesssim d h_l L_{\kappa t}^2 + h_l^2 L_{\kappa t}^2 \left(\mathbb{E}_{\pi_{\text{data}}} \left[\|X\|^2 \right] + d \right) + h_l^2 L_{\kappa t}^2 \left(L_{\kappa t} d + L_{\kappa t}^2 \mathbb{E}_{\mathbb{P}} \left[\|X_t - X_{t-}\|^2 \right] \right).$$

Substituting this expression into (26), we have

$$\begin{aligned} \text{KL}(\mathbb{P} \parallel \mathbb{Q}) &\lesssim \sum_{l=1}^M \int_{t_{l-1}}^{t_l} L_{\kappa t}^2 \mathbb{E}_{\mathbb{P}} [\|X_t - X_{t-}\|^2] dt + \int_0^{T/\kappa} \|v_t\|_{L^2(\hat{\mu}_t)}^2 dt \\ &+ \sum_{l=1}^M \int_{t_{l-1}}^{t_l} (dh_l L_{\kappa t}^2 + h_l^2 L_{\kappa t}^2 (\mathbb{E}_{\pi_{\text{data}}} [\|X\|^2] + d) + dh_l^2 L_{\kappa t}^3) dt. \end{aligned}$$

To bound $\mathbb{E}_{\mathbb{P}} [\|X_t - X_{t-}\|^2]$ we note that under \mathbb{P} , for $t \in [t_{l-1}, t_l)$, we have

$$X_t - X_{t-} = \int_{t_{l-1}}^t (\nabla \log \hat{\mu}_\tau + v_\tau)(X_\tau) d\tau + \sqrt{2}(B_t - B_{t_{l-1}}).$$

Therefore,

$$\begin{aligned} \mathbb{E}_{\mathbb{P}} [\|X_t - X_{t-}\|^2] &\lesssim \mathbb{E}_{\mathbb{P}} \left[\left\| \int_{t_{l-1}}^t (\nabla \log \hat{\mu}_\tau + v_\tau)(X_\tau) d\tau \right\|^2 \right] + \mathbb{E}_{\mathbb{P}} [\|\sqrt{2}(B_t - B_{t_{l-1}})\|^2] \\ &\lesssim (t - t_{l-1}) \int_{t_{l-1}}^t (\|\nabla \log \hat{\mu}_\tau\|_{L^2(\hat{\mu}_\tau)}^2 + \|v_\tau\|_{L^2(\hat{\mu}_\tau)}^2) d\tau + d(t - t_{l-1}) \\ &\lesssim h_l \int_{t_{l-1}}^{t_l} (\|\nabla \log \hat{\mu}_\tau\|_{L^2(\hat{\mu}_\tau)}^2 + \|v_\tau\|_{L^2(\hat{\mu}_\tau)}^2) d\tau + dh_l, \end{aligned}$$

where the second inequality arises from the application of the Cauchy-Schwarz inequality, and the last inequality is due to the definition $h_l = t_l - t_{l-1}$. Taking the integral over $t \in [t_{l-1}, t_l]$, it follows

$$\int_{t_{l-1}}^{t_l} L_{\kappa t}^2 \mathbb{E}_{\mathbb{P}} [\|X_t - X_{t-}\|^2] dt \lesssim \left(\int_{t_{l-1}}^{t_l} L_{\kappa t}^2 dt \right) \left(h_l \int_{t_{l-1}}^{t_l} (\|\nabla \log \hat{\mu}_t\|_{L^2(\hat{\mu}_t)}^2 + \|v_t\|_{L^2(\hat{\mu}_t)}^2) dt + dh_l \right).$$

Putting this together we have

$$\begin{aligned} &\sum_{l=1}^M \int_{t_{l-1}}^{t_l} L_{\kappa t}^2 \mathbb{E}_{\mathbb{P}} [\|X_t - X_{t-}\|^2] dt + \int_0^{T/\kappa} \|v_t\|_{L^2(\hat{\mu}_t)}^2 dt \\ &\lesssim \sum_{l=1}^M \left(\int_{t_{l-1}}^{t_l} L_{\kappa t}^2 dt \right) \left(h_l \int_{t_{l-1}}^{t_l} (\|\nabla \log \hat{\mu}_t\|_{L^2(\hat{\mu}_t)}^2 + \|v_t\|_{L^2(\hat{\mu}_t)}^2) dt + dh_l \right) + \int_0^{T/\kappa} \|v_t\|_{L^2(\hat{\mu}_t)}^2 dt \\ &\lesssim \sum_{l=1}^M \left(\int_{t_{l-1}}^{t_l} L_{\kappa t}^2 dt \right) \left(h_l \int_{t_{l-1}}^{t_l} (dL_{\kappa t} + \|v_t\|_{L^2(\hat{\mu}_t)}^2) dt + dh_l \right) + \int_0^{T/\kappa} \|v_t\|_{L^2(\hat{\mu}_t)}^2 dt \\ &\lesssim \sum_{l=1}^M \left(1 + h_l^2 \max_{[t_{l-1}, t_l]} L_t^2 \right) \int_{t_{l-1}}^{t_l} |\dot{\hat{\mu}}|_t^2 dt + \left(dh_l \int_{t_{l-1}}^{t_l} L_{\kappa t}^2 dt \right) \left(1 + h_l \max_{[t_{l-1}, t_l]} L_t \right). \end{aligned} \tag{27}$$

This results in the following bound for the KL divergence between \mathbb{P} and \mathbb{Q} :

$$\begin{aligned} \text{KL}(\mathbb{P} \parallel \mathbb{Q}) &\lesssim \sum_{l=1}^M \left(1 + h_l^2 \max_{[t_{l-1}, t_l]} L_t^2 \right) \int_{t_{l-1}}^{t_l} |\dot{\hat{\mu}}|_t^2 dt + \left(dh_l \int_{t_{l-1}}^{t_l} L_{\kappa t}^2 dt \right) \left(1 + h_l \max_{[t_{l-1}, t_l]} L_t \right) \\ &+ \sum_{l=1}^M \int_{t_{l-1}}^{t_l} (dh_l L_{\kappa t}^2 + h_l^2 L_{\kappa t}^2 (\mathbb{E}_{\pi_{\text{data}}} [\|X\|^2] + d) + dh_l^2 L_{\kappa t}^3) dt \end{aligned}$$

Note that intuitively we want to take smaller steps $h_l = t_l - t_{l-1}$ when L_t is larger. Define $h =$

$\max_{l \in \{1, \dots, M\}} h_l$, we can further simplify the previous expression to obtain

$$\begin{aligned} \text{KL}(\mathbb{P} \parallel \mathbb{Q}) &\lesssim \sum_{l=1}^M (1 + h^2 L_{\max}^2) \int_{t_{l-1}}^{t_l} \left| \dot{\hat{\mu}}_t \right|^2 dt + dh_l (1 + h L_{\max}) \int_{t_{l-1}}^{t_l} L_{\kappa t}^2 dt + \frac{T h^2 L_{\max}^2}{\kappa} \left(\mathbb{E}_{\pi_{\text{data}}} [\|X\|^2] + d \right) \\ &= (1 + h^2 L_{\max}^2) \int_0^{T/\kappa} \left| \dot{\hat{\mu}}_t \right|^2 dt + dh (1 + h L_{\max}) \int_0^{T/\kappa} L_{\kappa t}^2 dt + \frac{T h^2 L_{\max}^2}{\kappa} \left(\mathbb{E}_{\pi_{\text{data}}} [\|X\|^2] + d \right) \\ &= (1 + h^2 L_{\max}^2) \kappa \mathcal{A}_\lambda(\mu) + \frac{dh}{\kappa} (1 + h L_{\max}) \int_0^T L_t^2 dt + \frac{T h^2 L_{\max}^2}{\kappa} \left(\mathbb{E}_{\pi_{\text{data}}} [\|X\|^2] + d \right). \end{aligned}$$

The step size h can be expressed in terms of the number of steps M and κ as $h \asymp \frac{1}{M\kappa}$. Therefore, we have

$$\begin{aligned} \text{KL}(\mathbb{P} \parallel \mathbb{Q}) &\lesssim \left(1 + \frac{L_{\max}^2}{M^2 \kappa^2} \right) \kappa \mathcal{A}_\lambda(\mu) + \frac{d}{M \kappa^2} \left(1 + \frac{L_{\max}}{M \kappa} \right) \int_0^T L_t^2 dt + \frac{L_{\max}^2}{M^2 \kappa^3} \left(\mathbb{E}_{\pi_{\text{data}}} [\|X\|^2] + d \right) \\ &\lesssim \left(1 + \frac{L_{\max}^2}{M^2 \kappa^4} \right) \kappa C_\lambda (\mathbb{E}_{\pi_{\text{data}}} [\|X\|^2] + d) + \frac{d}{M \kappa^2} \left(1 + \frac{L_{\max}}{M \kappa} \right) \int_0^T L_t^2 dt, \end{aligned}$$

where we have used the bound on the action derived in Lemma 3.3 and $T = \mathcal{O}(1)$.

To derive the previous bound, we have assumed that the score of the intermediate distributions $\nabla \log \hat{\mu}_t$, can be computed exactly. In practice, however, we use an approximation, introducing an additional error term into the analysis. Let $s_\theta(\cdot, t)$ denote our estimator for $\nabla \log \hat{\mu}_t$ and let \mathbb{Q}_θ be the path measure of the continuous-time interpolation of the DALMC algorithm (4). We conclude that

$$\begin{aligned} \text{KL}(\mathbb{P} \parallel \mathbb{Q}_\theta) &= \frac{1}{4} \int_0^{T/\kappa} \mathbb{E}_{\mathbb{P}} \left[\left\| \nabla \log \hat{\mu}_t(X_t) - s_\theta(X_{t-}, t-) + v_t(X_t) \right\|^2 \right] dt \\ &\lesssim \int_0^{T/\kappa} \mathbb{E}_{\mathbb{P}} \left[\left\| \nabla \log \hat{\mu}_t(X_t) - \nabla \log \hat{\mu}_{t-}(X_{t-}) + v_t(X_t) \right\|^2 \right] dt \\ &\quad + \int_0^{T/\kappa} \mathbb{E}_{\mathbb{P}} \left[\left\| \nabla \log \hat{\mu}_{t-}(X_{t-}) - s_\theta(X_{t-}, t-) \right\|^2 \right] dt \\ &\lesssim \left(1 + \frac{L_{\max}^2}{M^2 \kappa^4} \right) \kappa C_\lambda (\mathbb{E}_{\pi_{\text{data}}} [\|X\|^2] + d) + \frac{d}{M \kappa^2} \left(1 + \frac{L_{\max}}{M \kappa} \right) \int_0^T L_t^2 dt \\ &\quad + \sum_{l=0}^{M-1} h_l \mathbb{E}_{\hat{\mu}_t} \left[\left\| \nabla \log \hat{\mu}_l(X_{t_l}) - s_\theta(X_{t_l}, t_l) \right\|^2 \right] \\ &\lesssim \left(1 + \frac{L_{\max}^2}{M^2 \kappa^4} \right) \kappa C_\lambda (\mathbb{E}_{\pi_{\text{data}}} [\|X\|^2] + d) + \frac{d}{M \kappa^2} \left(1 + \frac{L_{\max}}{M \kappa} \right) \int_0^T L_t^2 dt + \varepsilon_{\text{score}}^2 \\ &\lesssim \left(1 + \frac{L_{\max}^2}{M^2 \kappa^4} \right) \kappa (M_2 \vee d) + \frac{d}{M \kappa^2} \left(1 + \frac{L_{\max}}{M \kappa} \right) L_{\max}^2 + \varepsilon_{\text{score}}^2, \end{aligned}$$

where we have used the control of the score approximation given in assumption A1. \square

B.6 Proof of Corollary 3.5

Proof. Based on the KL bound established in Theorem 3.4, we can obtain the iteration complexity of the DALMC algorithm (4). Observe that by selecting

$$\kappa = \mathcal{O} \left(\frac{\varepsilon_{\text{score}}^2}{M_2 \vee d} \right), \quad M = \mathcal{O} \left(\frac{d(M_2 \vee d)^2 L_{\max}^2}{\varepsilon_{\text{score}}^6} \right),$$

it follows that $\text{KL}(\mathbb{P} \parallel \mathbb{Q}_\theta) \leq \varepsilon_{\text{score}}^2$. Therefore, for any $\varepsilon = \mathcal{O}(\varepsilon_{\text{score}})$, the DALMC algorithm requires at most

$$M = \mathcal{O} \left(\frac{d(M_2 \vee d)^2 L_{\max}^2}{\varepsilon^6} \right)$$

steps to approximate π_{data} to within ε^2 in KL divergence. \square

B.7 Comment on Assumption A7

We show below that both assumptions **A4** and **A5** independently imply **A7**.

A4 \Rightarrow A7: First, recall that we proved in Lemma B.7 that if π_{data} satisfies **A4**, then it has finite log-Sobolev and Poincaré constants, thereby ensuring $\mathbb{E}_{\pi_{\text{data}}} \|X\|^2$ is finite. Besides, observe that

$$\mathbb{E}_{\pi_{\text{data}}} \|\nabla V_\pi(X)\|^8 \lesssim \|\nabla V_\pi(0)\|^8 + L_\pi^8 \mathbb{E}_{\pi_{\text{data}}} \|X\|^8.$$

Using Poincaré inequality, it follows

$$\begin{aligned} \mathbb{E}_{\pi_{\text{data}}} \|X\|^4 &= \text{Var}_{\pi_{\text{data}}} (\|X\|^2) + (\mathbb{E}_{\pi_{\text{data}}} \|X\|^2)^2 \lesssim C_{\text{PI}} \mathbb{E}_{\pi_{\text{data}}} \|X\|^2 + (\mathbb{E}_{\pi_{\text{data}}} \|X\|^2)^2, \\ \mathbb{E}_{\pi_{\text{data}}} \|X\|^6 &= \text{Var}_{\pi_{\text{data}}} (\|X\|^3) + (\mathbb{E}_{\pi_{\text{data}}} \|X\|^3)^2 \lesssim C_{\text{PI}} \mathbb{E}_{\pi_{\text{data}}} \|X\|^4 + (\mathbb{E}_{\pi_{\text{data}}} \|X\|^4)^{3/2}, \\ \mathbb{E}_{\pi_{\text{data}}} \|X\|^8 &= \text{Var}_{\pi_{\text{data}}} (\|X\|^4) + (\mathbb{E}_{\pi_{\text{data}}} \|X\|^4)^2 \lesssim C_{\text{PI}} \mathbb{E}_{\pi_{\text{data}}} \|X\|^6 + (\mathbb{E}_{\pi_{\text{data}}} \|X\|^4)^2, \end{aligned}$$

which demonstrates that **A4** implies **A5**.

A5 \Rightarrow A7: If π_{data} satisfies **A5** we have that $\nabla^2 V_{\pi_{\text{data}}}$ has asymptotic order $\mathcal{O}(\|x\|^{-2}I)$, hence by Dieudonné [1973] $\|\nabla V_{\pi_{\text{data}}}\|$ has asymptotic order $\mathcal{O}(\|x\|^{-1})$. Therefore, $\mathbb{E}_{\pi_{\text{data}}} \|\nabla V_\pi(X)\|^8$ is guaranteed to be bounded, which concludes that **A5** implies **A7**.

B.8 Proof of Theorem 3.6

Proof. Let \mathbb{Q} denote the path measure associated with the continuous-time interpolation of the modified DALMC algorithm with exact scores (25), which corresponds to the SDE

$$dX_t = \nabla \log \hat{\mu}_{t-}(X_{t-}) dt + \sqrt{2} dB_t, \quad t \in [0, T/\kappa],$$

where given a discretisation of the interval $[0, T/\kappa]$, $0 = t_0 < t_1 < \dots < t_M = T/\kappa$, we define $t_- := t_{l-1}$ when $t \in [t_{l-1}, t_l)$ for $l = 1, \dots, M$. On the other hand, let \mathbb{P} be the path measure corresponding to the following reference SDE

$$dX_t = (\nabla \log \hat{\mu}_t + v_t)(X_t) dt + \sqrt{2} dB_t, \quad t \in [0, T/\kappa].$$

As in the proof of Theorem 3.4, the vector field $v = (v_t)_{t \in [0, T/\kappa]}$ is designed such that $X_t \sim \hat{\mu}_t$ for all $t \in [0, T/\kappa]$ and $\|v_t\|_{L^2(\hat{\mu}_t)} = |\dot{\hat{\mu}}_t|$. Using Girsanov's theorem we have

$$\begin{aligned} \text{KL}(\mathbb{P} \parallel \mathbb{Q}) &= \frac{1}{4} \int_0^{T/\kappa} \mathbb{E}_{\mathbb{P}} \left[\|\nabla \log \hat{\mu}_t(X_t) - \nabla \log \hat{\mu}_{t-}(X_{t-}) + v_t(X_t)\|^2 \right] dt \\ &\lesssim \int_0^{T/\kappa} \mathbb{E}_{\mathbb{P}} \left[\|\nabla \log \hat{\mu}_t(X_t) - \nabla \log \hat{\mu}_{t-}(X_{t-})\|^2 \right] dt + \int_0^{T/\kappa} \|v_t\|_{L^2(\hat{\mu}_t)}^2 dt. \end{aligned}$$

Note that the first term involves both a time and space discretisation error. Inspired by Chen et al. [2022], we start analysing

$$E_{t,s} = \mathbb{E}_{\mathbb{P}} \left[\|\nabla \log \hat{\mu}_t(X_t) - \nabla \log \hat{\mu}_s(X_s)\|^2 \right],$$

with $0 \leq s < t \leq 1$. Recall that we can write

$$X_s = \sqrt{\frac{\lambda_{\kappa s}}{\lambda_{\kappa t}}} X_t + \sigma \left(\sqrt{1 - \frac{\lambda_{\kappa s}}{\lambda_{\kappa t}}} - \sqrt{\frac{(1 - \lambda_{\kappa t})\lambda_{\kappa s}}{\lambda_{\kappa t}}} \right) Z = \alpha_{t,s} X_t + \tilde{\sigma}_{t,s} Z,$$

where $Z \sim \mathcal{N}(0, I)$ independent of X_t . So, we have

$$\hat{\mu}_s(x) \propto \int \alpha_{t,s}^{-d} \hat{\mu}_t \left(\frac{y}{\alpha_{t,s}} \right) e^{-\frac{\|x-y\|^2}{2\tilde{\sigma}_{t,s}^2}} dy = \int \hat{\mu}_t(y) e^{-\frac{\|x-\alpha_{t,s}y\|^2}{2\tilde{\sigma}_{t,s}^2}} dy.$$

Therefore, similarly to the proof of Lemma 3.2, we can express the score of $\hat{\mu}_s$ in terms of that of $\hat{\mu}_t$,

$$\nabla \log \hat{\mu}_s(x) = \alpha_{t,s}^{-1} \mathbb{E}_{Y \sim \varphi_{t|x,s}} [\nabla \log \hat{\mu}_t(Y)|x,s],$$

where $\varphi_{t|x,s}(y) \propto \hat{\mu}_t(y) e^{-\frac{\|x - \alpha_{t,s}y\|^2}{2\hat{\sigma}_{t,s}^2}}$. Substituting this we have

$$\begin{aligned} E_{t,s} &\leq 2\mathbb{E}_{\mathbb{P}} \left[\|\nabla \log \hat{\mu}_s(X_s) - \nabla \log \hat{\mu}_t(\alpha_{t,s}^{-1}X_s)\|^2 \right] + 2\mathbb{E}_{\mathbb{P}} \left[\|\nabla \log \hat{\mu}_t(X_t) - \nabla \log \hat{\mu}_t(\alpha_{t,s}^{-1}X_s)\|^2 \right] \\ &= 2\mathbb{E}_{\mathbb{P}} \left[\left\| \mathbb{E}_{\varphi_{t|x,s}} [(\alpha_{t,s}^{-1} - 1)\nabla \log \hat{\mu}_t(Y) + \nabla \log \hat{\mu}_t(Y) - \nabla \log \hat{\mu}_t(\alpha_{t,s}^{-1}X_s)] \right\|^2 \right] \\ &\quad + 2\mathbb{E}_{\mathbb{P}} \left[\|\nabla \log \hat{\mu}_t(X_t) - \nabla \log \hat{\mu}_t(\alpha_{t,s}^{-1}X_s)\|^2 \right] \\ &\leq 4(\alpha_{t,s}^{-1} - 1)^2 \mathbb{E} [\|\nabla \log \hat{\mu}_t(X_t)\|^2] + 6 \mathbb{E}_{\mathbb{P}} \left[\|\nabla \log \hat{\mu}_t(X_t) - \nabla \log \hat{\mu}_t(\alpha_{t,s}^{-1}X_s)\|^2 \right]. \end{aligned} \quad (28)$$

Using the score expressions provided in Lemma 3.2 Eq. (11), we can bound $\mathbb{E} [\|\nabla \log \hat{\mu}_t(X_t)\|^2]$ as follows

$$\begin{aligned} \mathbb{E}_{\hat{\mu}_t} [\|\nabla \log \hat{\mu}_t(X_t)\|^2] &\leq \frac{1}{\sigma^2(1 - \lambda_{kt})} \mathbb{E}_{\hat{\mu}_t} \mathbb{E}_{\rho_{t,x_t}} \left[\left\| \frac{X_t - Y}{\sigma\sqrt{1 - \lambda_{kt}}} \right\|^2 \right] = \frac{d}{\sigma^2(1 - \lambda_{kt})}, \\ \mathbb{E}_{\hat{\mu}_t} [\|\nabla \log \hat{\mu}_t(X_t)\|^2] &\leq \frac{1}{\lambda_{kt}} \mathbb{E}_{\hat{\mu}_t} \mathbb{E}_{\rho_{t,x_t}} \left[\left\| \nabla V_{\pi} \left(\frac{Y}{\sqrt{\lambda_{kt}}} \right) \right\|^2 \right] = \frac{1}{\lambda_{kt}} \mathbb{E}_{\pi_{\text{data}}} [\|\nabla V_{\pi}(Y)\|^2] \leq \frac{L_{\pi}}{\lambda_{kt}} d, \end{aligned}$$

where we have used the Lipschitzness of π_{data} . This provides, $\mathbb{E} [\|\nabla \log \hat{\mu}_t(X_t)\|^2] \leq \min \left\{ \frac{d}{\sigma^2(1 - \lambda_{kt})}, \frac{L_{\pi}d}{\lambda_{kt}} \right\}$. Following a similar argument to Chen et al. [2022, Lemma 13], we study the second term in (28),

$$\begin{aligned} \mathbb{E} \left[\|\nabla \log \hat{\mu}_t(X_t) - \nabla \log \hat{\mu}_t(\alpha_{t,s}^{-1}X_s)\|^2 \right] &= \mathbb{E} \left[\left\| \int_0^1 \nabla^2 \log \hat{\mu}_t(X_t + aZ_{t,s}) Z_{t,s} da \right\|^2 \right] \\ &\leq \int_0^1 \mathbb{E} [\|\nabla^2 \log \hat{\mu}_t(X_t + aZ_{t,s}) Z_{t,s}\|^2] da, \end{aligned}$$

where $Z_{t,s} = \alpha_{t,s}^{-1}X_s - X_t \sim \mathcal{N} \left(0, \left(\frac{\hat{\sigma}_{t,s}}{\alpha_{t,s}} \right)^2 I \right)$ is independent of X_t . For simplicity, we denote

$$\hat{\sigma}_{t,s} = \frac{\tilde{\sigma}_{t,s}}{\alpha_{t,s}} = \sigma \left(\sqrt{\frac{(1 - \lambda_{ks})\lambda_{kt}}{\lambda_{ks}}} - \sqrt{1 - \lambda_{kt}} \right).$$

To bound $\mathbb{E} [\|\nabla^2 \log \hat{\mu}_t(X_t + aZ_{t,s}) Z_{t,s}\|^2]$ we consider the following change of variable

$$\begin{aligned} \mathbb{E} \left[\|\nabla^2 \log \hat{\mu}_t(X_t + aZ_{t,s}) Z_{t,s}\|^2 \right] &= \mathbb{E} \left[\left\| \nabla^2 \log \hat{\mu}_t(X_t) Z_{t,s} \right\|^2 \frac{dP_{X_t+aZ_{t,s},Z_{t,s}}(X_t, Z_{t,s})}{dP_{X_t,Z_{t,s}}(X_t, Z_{t,s})} \right] \\ &\lesssim \left(\mathbb{E} \left[\left\| \nabla^2 \log \hat{\mu}_t(X_t) Z_{t,s} \right\|^4 \right] \mathbb{E} \left(\frac{dP_{X_t+aZ_{t,s},Z_{t,s}}(X_t, Z_{t,s})}{dP_{X_t,Z_{t,s}}(X_t, Z_{t,s})} \right)^2 \right)^{1/2}. \end{aligned}$$

Let $M_t = \nabla^2 \log \hat{\mu}_t(X_t)$ ($\nabla^2 \log \hat{\mu}_t(X_t)$) $^{\top}$ and $N_{t,s} = Z_{t,s} Z_{t,s}^{\top}$, which by definition they are independent. We now need to bound the two previous factors. By the properties of the tensor product, we have

$$\mathbb{E} \left[\left\| \nabla^2 \log \hat{\mu}_t(X_t) Z_{t,s} \right\|^4 \right] = \mathbb{E} \left[\text{Tr} (M_t^{\top} Z_{t,s})^2 \right] = \langle \mathbb{E} M_t \otimes M_t, \mathbb{E} Z_{t,s} \otimes Z_{t,s} \rangle.$$

Using the properties of the χ^2 distribution, we obtain

$$\mathbb{E} (Z_{t,s} \otimes Z_{t,s})_{i_1, i_2, i_3, i_4} = \begin{cases} 3\hat{\sigma}_{t,s}^4, & i_1 = i_2 = i_3 = i_4, \\ \hat{\sigma}_{t,s}^4, & i_1 \neq i_2, (i_1, i_2) = (i_3, i_4) \text{ or } (i_1, i_2) = (i_4, i_3), \\ 0, & \text{otherwise.} \end{cases}$$

Substituting this we have

$$\begin{aligned} \mathbb{E} \left\| \nabla^2 \log \hat{\mu}_t(X_t) Z_{t,s} \right\|^4 &\lesssim \hat{\sigma}_{t,s}^4 \left(\sum_{(i_1, i_2) = (i_3, i_4)} + \sum_{(i_1, i_2) = (i_4, i_3)} \right) \mathbb{E} (M_t \otimes M_t)_{i_1, i_2, i_3, i_4} \\ &\lesssim \hat{\sigma}_{t,s}^4 \sum_{(i_1, i_2) = (i_3, i_4)} \mathbb{E} (M_t \otimes M_t)_{i_1, i_2, i_3, i_4} \lesssim \hat{\sigma}_{t,s}^4 \mathbb{E} \|M_t\|_F^2 \lesssim \hat{\sigma}_{t,s}^4 \mathbb{E} \|\nabla^2 \log \hat{\mu}_t(X_t)\|_F^4. \end{aligned}$$

Applying Lemma 12 from Chen et al. [2022], it follows that

$$\mathbb{E} \|\nabla^2 \log \hat{\mu}_t(X_t)\|_F^4 \lesssim \left(\frac{d}{\sigma^2(1 - \lambda_{\kappa t})} \right)^4.$$

This bound becomes arbitrarily large as $\lambda_{\kappa t}$ tends to 1, however, using the alternative expression for the Hessian provided in (13), we can write

$$\nabla^2 \log \hat{\mu}_t(X_t) = -\frac{1}{\lambda_{\kappa t}} \left(\mathbb{E}_{\rho_{t,X_t}} \left[\nabla^2 V_\pi \left(\frac{Y}{\sqrt{\lambda_{\kappa t}}} \right) \right] - \text{Cov}_{\rho_{t,X_t}} \left[\nabla V_\pi \left(\frac{Y}{\sqrt{\lambda_{\kappa t}}} \right) \right] \right),$$

where $\rho_{t,X_t}(y) \propto e^{-V_\pi(y/\sqrt{\lambda_{\kappa t}}) - \|x-y\|^2/(2\sigma^2(1-\lambda_{\kappa t}))}$. Due to the Lipschitzness of π_{data} , we have

$$-L_\pi I \preccurlyeq \mathbb{E}_{\rho_{t,X_t}} \left[\nabla^2 V_\pi \left(\frac{Y}{\sqrt{\lambda_{\kappa t}}} \right) \right] \preccurlyeq L_\pi I,$$

where the Frobenius norm of the identity matrix is $\|I\|_F = \sqrt{d}$. For the covariance term we proceed as follows

$$\begin{aligned} \frac{1}{\lambda_{\kappa t}^4} \mathbb{E}_{\hat{\mu}_t} \left\| \text{Cov}_{\rho_{t,X_t}} \left[\nabla V_\pi \left(\frac{Y}{\sqrt{\lambda_{\kappa t}}} \right) \right] \right\|_F^4 &\leq \frac{1}{\lambda_{\kappa t}^4} \mathbb{E}_{\hat{\mu}_t} \mathbb{E}_{\rho_{t,X_t}} \left\| \nabla V_\pi \left(\frac{Y}{\sqrt{\lambda_{\kappa t}}} \right) \nabla V_\pi \left(\frac{Y}{\sqrt{\lambda_{\kappa t}}} \right)^\top \right\|_F^4 \\ &= \frac{1}{\lambda_{\kappa t}^4} \mathbb{E}_{T_{\lambda_{\kappa t}^{-1/2} \# \pi_{\text{data}}}} \left\| \nabla V_\pi \left(\frac{Y}{\sqrt{\lambda_{\kappa t}}} \right) \nabla V_\pi \left(\frac{Y}{\sqrt{\lambda_{\kappa t}}} \right)^\top \right\|_F^4 = \frac{1}{\lambda_{\kappa t}^4} \mathbb{E}_{T_{\lambda_{\kappa t}^{-1/2} \# \pi_{\text{data}}}} \left\| \nabla V_\pi \left(\frac{Y}{\sqrt{\lambda_{\kappa t}}} \right) \right\|^8 \\ &= \frac{1}{\lambda_{\kappa t}^4} \mathbb{E}_{\pi_{\text{data}}} \|\nabla V_\pi(Y)\|^8 \leq \frac{K_\pi^2}{\lambda_{\kappa t}^4}. \end{aligned}$$

Therefore, we have

$$\mathbb{E} \|\nabla^2 \log \hat{\mu}_t(X_t)\|_F^4 \lesssim \min \left\{ \frac{d^4}{\sigma^8(1 - \lambda_{\kappa t})^4}, \frac{L_\pi^4 d^2 + K_\pi^2}{\lambda_{\kappa t}^4} \right\}.$$

Next, we bound the term concerning the change of variable

$$\begin{aligned} \mathbb{E} \left(\frac{dP_{X_t+aZ_{t,s},Z_{t,s}}(X_t, Z_{t,s})}{dP_{X_t,Z_{t,s}}(X_t, Z_{t,s})} \right)^2 &= \mathbb{E} \left(\frac{dP_{X_t+aZ_{t,s}|Z_{t,s}}(X_t|Z_{t,s})}{dP_{X_t|Z_{t,s}}(X_t|Z_{t,s})} \right)^2 \leq \mathbb{E} \left(\frac{dP_{X_t+aZ_{t,s}|Z_{t,s},X_{T_\kappa}}(X_t|Z_{t,s}, X_{T_\kappa})}{dP_{X_t|Z_{t,s},X_{T_\kappa}}(X_t|Z_{t,s}, X_{T_\kappa})} \right)^2 \\ &= \mathbb{E} \left(\frac{dP_{X_t+aZ_{t,s}|Z_{t,s},X_{T_\kappa}}(X_t|Z_{t,s}, X_{T_\kappa})}{dP_{X_t|X_{T_\kappa}}(X_t|X_{T_\kappa})} \right)^2, \end{aligned}$$

where we have used the data processing inequality and $X_{T_\kappa} \sim \pi_{\text{data}}$. Since $X_t + aZ_{t,s}|(Z_{t,s}, X_{T_\kappa}) \sim \mathcal{N}(\sqrt{\lambda_{\kappa t}}X_{T_\kappa} + aZ_{t,s}, \sigma^2(1 - \lambda_{\kappa t}))$ and $X_t|X_{T_\kappa} \sim \mathcal{N}(\sqrt{\lambda_{\kappa t}}X_{T_\kappa}, \sigma^2(1 - \lambda_{\kappa t}))$, we can explicitly compute the previous expression, as it corresponds to the χ^2 divergence between two Gaussians, that is,

$$\mathbb{E} \left(\frac{dP_{X_t+aZ_{t,s},Z_{t,s}}(X_t, Z_{t,s})}{dP_{X_t,Z_{t,s}}(X_t, Z_{t,s})} \right)^2 \leq \mathbb{E} \exp \left(\frac{a^2 \|Z_{t,s}\|^2}{\sigma^2(1 - \lambda_{\kappa t})} \right) \leq \left(1 - 2a^2 \left(1 - \sqrt{\frac{\lambda_{\kappa t}(1 - \lambda_{\kappa s})}{\lambda_{\kappa s}(1 - \lambda_{\kappa t})}} \right)^2 \right)^{-d/2},$$

where we have used the expression of the moment generating function of a χ^2 distribution. Under the assumption on the schedule and for $t - s \ll 1$, it follows that

$$\mathbb{E} \left(\frac{dP_{X_s+aZ_{t,s},Z_{t,s}}(X_s, Z_{t,s})}{dP_{X_s,Z_{t,s}}(X_s, Z_{t,s})} \right)^2 \lesssim 1.$$

Putting all this together, we have

$$\begin{aligned}\mathbb{E} \left[\|\nabla \log \hat{\mu}_t(X_t) - \nabla \log \hat{\mu}_t(\alpha_{t,s}^{-1} X_s)\|^2 \right] &\lesssim \hat{\sigma}_{t,s}^2 \min \left\{ \frac{d^2}{\sigma^4(1-\lambda_{\kappa t})^2}, \frac{L_\pi^2 d + K_\pi}{\lambda_{\kappa t}^2} \right\} \\ &\lesssim \sigma^2(t-s) \min \left\{ \frac{d^2}{\sigma^4(1-\lambda_{\kappa t})^2}, \frac{L_\pi^2 d \vee K_\pi}{\lambda_{\kappa t}^2} \right\}\end{aligned}$$

for $t-s \ll 1$. Substituting this into (28), it follows

$$E_{t,s} \lesssim (t-s)^2 d \min \left\{ \frac{1}{\sigma^2(1-\lambda_{\kappa t})}, \frac{L_\pi}{\lambda_{\kappa t}} \right\} + \sigma^2(t-s) \min \left\{ \frac{d^2}{\sigma^4(1-\lambda_{\kappa t})^2}, \frac{L_\pi^2 d \vee K_\pi}{\lambda_{\kappa t}^2} \right\}.$$

Therefore, this results into the following bound for the KL

$$\begin{aligned}\text{KL}(\mathbb{P} \parallel \mathbb{Q}) &\lesssim \sum_{l=1}^M \int_{t_{l-1}}^{t_l} \mathbb{E}_{\mathbb{P}} \left[\|\nabla \log \hat{\mu}_t(X_t) - \nabla \log \hat{\mu}_{t-}(X_{t-})\|^2 \right] dt + \int_0^{T/\kappa} \|v_t\|_{L^2(\hat{\mu}_t)}^2 dt \\ &\lesssim \sum_{l=1}^M h_l^3 d \min \left(\frac{1}{\sigma^2(1-\lambda_{\kappa t_l})}, \frac{L_\pi}{\lambda_{\kappa t_{l-1}}} \right) + \sigma^2 h_l^2 \min \left(\frac{d^2}{\sigma^4(1-\lambda_{\kappa t_l})^2}, \frac{L_\pi^2 d \vee K_\pi}{\lambda_{\kappa t_{l-1}}^2} \right) + \kappa \mathcal{A}_\lambda(\mu).\end{aligned}$$

Let \mathbb{Q}_θ be the path measure of the continuous-time interpolation of the DALMC algorithm (4). When implementing the algorithm with step sizes $h_l \asymp 1/(M\kappa)$, we have

$$\begin{aligned}\text{KL}(\mathbb{P} \parallel \mathbb{Q}_\theta) &\lesssim \text{KL}(\mathbb{P} \parallel \mathbb{Q}) + \sum_{l=0}^{M-1} h_l \mathbb{E}_{\hat{\mu}_t} \left[\|\nabla \log \hat{\mu}_l(X_{t_l}) - s_\theta(X_{t_l}, t_l)\|^2 \right] \\ &\lesssim \frac{d L_\pi}{M^2 \kappa^3} + \frac{(d^2 \vee L_\pi^2 d \vee K_\pi)}{M \kappa^2} + \kappa (\mathbb{E}_{\pi_{\text{data}}} [\|X\|^2] + d) + \varepsilon_{\text{score}}^2,\end{aligned}$$

where we have used the bound on the action given in Lemma 4.2. We can conclude that by taking

$$\kappa = \mathcal{O} \left(\frac{\varepsilon_{\text{score}}^2}{M_2 \vee d} \right), \quad M = \mathcal{O} \left(\frac{(M_2 \vee d)^2 (d^2 \vee L_\pi^2 d \vee K_\pi) L_\pi}{\varepsilon_{\text{score}}^6} \right),$$

we guarantee that $\text{KL}(\mathbb{P} \parallel \mathbb{Q}_\theta) \leq \varepsilon_{\text{score}}^2$. Therefore, for any $\varepsilon = \mathcal{O}(\varepsilon_{\text{score}})$, the DALMC algorithm under relaxed assumptions requires at most

$$M = \mathcal{O} \left(\frac{(M_2 \vee d)^2 (d^2 \vee L_\pi^2 d \vee K_\pi) L_\pi}{\varepsilon^6} \right)$$

steps to approximate π_{data} to within ε^2 in KL divergence. Note that if $M_2 = \mathcal{O}(d)$, $L_\pi = \mathcal{O}(\sqrt{d})$ and $K_\pi = \mathcal{O}(d^2)$, then the number of steps is of order

$$M = \mathcal{O} \left(\frac{T^2 d^4 L_\pi}{\varepsilon^6} \right).$$

□

C Proofs of Section 4

C.1 Comments on Assumption A9

Lemma C.1. *If $\tilde{\pi}$ is supported in a closed Euclidean ball $B_d(0, R)$ and $\gamma \sim t(0, \tau^2 I, \alpha)$, then $\pi = \tilde{\pi} * \gamma$ satisfies a weighted Poincaré inequality.*

Proof. Recall that

$$\pi = \tilde{\pi} * \gamma = \int_{\mathbb{R}^d} \gamma_x d\tilde{\pi}(x),$$

where $\gamma_x \sim t(x, \sigma^2 I, \alpha)$. Following Bardet et al. [2018], the variance of a function $f \in L^2(\pi)$ can be decomposed as

$$\text{Var}_{\tilde{\pi} * \gamma}(f) = \int_{\mathbb{R}^d} \text{Var}_{\gamma_x}(f) d\tilde{\pi}(x) + \text{Var}_{\tilde{\pi}} \left(x \mapsto \int f d\gamma_x \right) := A + B.$$

Since γ_x satisfies a weighted Poincaré inequality with constant $C_{\text{PI}, \gamma}$ and weight function $\omega = 1 + \|x\|^2$ [Cattiaux et al., 2010a, Proposition 2.17], the first term A is bounded by

$$A \leq C_{\text{PI}, \gamma} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |\nabla f|^2 \omega d\gamma_x d\tilde{\pi}(x) = C_{\text{PI}, \gamma} \int_{\mathbb{R}^d} |\nabla f|^2 \omega d(\tilde{\pi} * \gamma).$$

For the second term B , consider $g : x \mapsto \int_{\mathbb{R}^d} f d\gamma_x$. Using this, B can be rewritten as

$$B = \frac{1}{2} \int \int_{\mathbb{R}^d \times \mathbb{R}^d} (g(x) - g(y))^2 d\tilde{\pi}(x) d\tilde{\pi}(y),$$

where using Cauchy-Schwartz inequality

$$(g(x) - g(y))^2 \leq \text{Var}_{\gamma_x}(f) \text{Var}_{\gamma_x} \left(1 - \frac{d\gamma_y}{d\gamma_x} \right).$$

For the first factor, we reapply the weighted Poincaré inequality for the t distribution γ_x . The second factor is the χ^2 divergence between the t distributions γ_x and γ_y .

$$\begin{aligned} \chi^2(\gamma_x, \gamma_y) &= \int_{\mathbb{R}^d} \left(\frac{\gamma_x(z)}{\gamma_y(z)} - 1 \right)^2 \gamma_y(z) dz = \int_{\mathbb{R}^d} \frac{\gamma_x(z)^2}{\gamma_y(z)} dz - 1 = \int_{\mathbb{R}^d} \left(\frac{\alpha + (z-x)^\top \Sigma^{-1}(z-x)}{\alpha + (z-y)^\top \Sigma^{-1}(z-y)} \right)^{-(\alpha+d)/2} \gamma_x(z) dz - 1 \\ &= \int_{\mathbb{R}^d} \left(1 + \frac{(z-y)^\top \Sigma^{-1}(z-y) - (z-x)^\top \Sigma^{-1}(z-x)}{\alpha + (z-x)^\top \Sigma^{-1}(z-x)} \right)^{(\alpha+d)/2} \gamma_x(z) dz - 1 \\ &= \int_{\mathbb{R}^d} \left(1 + \frac{y^\top \Sigma^{-1}y - x^\top \Sigma^{-1}x - 2z^\top \Sigma^{-1}(y-x)}{\alpha + (z-x)^\top \Sigma^{-1}(z-x)} \right)^{(\alpha+d)/2} \gamma_x(z) dz - 1 \\ &= \int_{A=\{\|z-x\|^2 > 1\}} \left(1 + \frac{y^\top \Sigma^{-1}y - x^\top \Sigma^{-1}x - 2z^\top \Sigma^{-1}(y-x)}{\alpha + (z-x)^\top \Sigma^{-1}(z-x)} \right)^{(\alpha+d)/2} \gamma_x(z) dz \\ &\quad + \int_{\mathbb{R}^d \setminus A} \left(1 + \frac{y^\top \Sigma^{-1}y - x^\top \Sigma^{-1}x - 2z^\top \Sigma^{-1}(y-x)}{\alpha + (z-x)^\top \Sigma^{-1}(z-x)} \right)^{(\alpha+d)/2} \gamma_x(z) dz - 1. \end{aligned}$$

For $z \in \mathbb{R}^d$, the following holds

$$(z-x)^\top \Sigma^{-1}(z-x) = (z-x)^\top U^\top S^{-1/2} S^{-1/2} U(z-x) = \|S^{-1/2} U(z-x)\|^2 \geq 0.$$

The second integral can be upper bounded as follows

$$\begin{aligned} &\int_{\mathbb{R}^d \setminus A} \left(1 + \frac{y^\top \Sigma^{-1}y - x^\top \Sigma^{-1}x - 2z^\top \Sigma^{-1}(y-x)}{\alpha + (z-x)^\top \Sigma^{-1}(z-x)} \right)^{(\alpha+d)/2} \gamma_x(z) dz \\ &= \int_{\mathbb{R}^d \setminus A} \left(\frac{\alpha + (z-y)^\top \Sigma^{-1}(z-y)}{\alpha + (z-x)^\top \Sigma^{-1}(z-x)} \right)^{(\alpha+d)/2} \gamma_x(z) dz \leq \int_{\mathbb{R}^d \setminus A} \left(\frac{\alpha + \sigma_{\min}^{-1} \|z-y\|^2}{\alpha} \right)^{(\alpha+d)/2} \gamma_x(z) dz \\ &\leq \int_{\mathbb{R}^d \setminus A} \left(\frac{\alpha + \sigma_{\min}^{-1} 2(\|x-y\|^2 + \|z-x\|^2)}{\alpha} \right)^{(\alpha+d)/2} \gamma_x(z) dz \leq \left(\frac{\alpha + \sigma_{\min}^{-1} 2(4R^2 + 1)}{\alpha} \right)^{(\alpha+d)/2} = \kappa_R, \end{aligned}$$

where we have used that $\|x - y\|$ is bounded by $2R$. On the other hand, for $z \in A$ we have

$$\begin{aligned} 1 + \frac{y^\top \Sigma^{-1} y - x^\top \Sigma^{-1} x - 2z^\top \Sigma^{-1}(y-x)}{\alpha + (z-x)^\top \Sigma^{-1}(z-x)} &= 1 + \frac{(y-x)^\top \Sigma^{-1}(y-x) - 2(z-x)^\top \Sigma^{-1}(y-x)}{\alpha + (z-x)^\top \Sigma^{-1}(z-x)} \\ &\leq 1 + \frac{(y-x)^\top \Sigma^{-1}(y-x)}{\alpha + (z-x)^\top \Sigma^{-1}(z-x)} + \frac{2\|(z-x)^\top \Sigma^{-1}(y-x)\|}{(z-x)^\top \Sigma^{-1}(z-x)} \\ &\leq 1 + \frac{(y-x)^\top \Sigma^{-1}(y-x)}{\alpha + (z-x)^\top \Sigma^{-1}(z-x)} + \frac{2\|S^{-1/2}U(z-x)\|\|S^{-1/2}U(y-x)\|}{\|S^{-1/2}U(z-x)\|^2} \\ &= 1 + \frac{(y-x)^\top \Sigma^{-1}(y-x)}{\alpha + (z-x)^\top \Sigma^{-1}(z-x)} + \frac{2\|S^{-1/2}U(y-x)\|}{\|S^{-1/2}U(z-x)\|} \leq 1 + \frac{4R^2\|S^{-1/2}\|^2}{\alpha} + \frac{4R\|S^{-1/2}\|}{\sigma_{\max}^{-1/2}} \leq 1 + \frac{4R^2}{\alpha\sigma_{\min}} + \frac{4R\sigma_{\max}^{1/2}}{\sigma_{\min}^{1/2}}. \end{aligned}$$

Therefore, we obtain

$$\int_{A=\{\|z-x\|^2>1\}} \left(1 + \frac{y^\top \Sigma^{-1} y - x^\top \Sigma^{-1} x - 2z^\top \Sigma^{-1}(y-x)}{\alpha + (z-x)^\top \Sigma^{-1}(z-x)}\right)^{(\alpha+d)/2} \gamma_x(z) dz \leq \left(1 + \frac{4R^2}{\alpha\sigma_{\min}} + \frac{4R\sigma_{\max}^{1/2}}{\sigma_{\min}^{1/2}}\right)^{(\alpha+d)/2} = \beta_R$$

This shows that $\chi^2(\gamma_x, \gamma_y)$ is upper bounded with dependence on R of the form $R^{(\alpha+d)}$. In particular, B satisfies the following bound

$$B \leq C_{\text{PI},\gamma}(\kappa_R + \beta_R - 1) \int_{\mathbb{R}^d} |\nabla f|^2 \omega d(\tilde{\pi} * \gamma).$$

Therefore, the measure $\tilde{\pi} * \gamma$ satisfies a weighted Poincaré inequality with constant

$$C_{\text{PI}} \leq C_{\text{PI},\gamma}(\kappa_R + \beta_R).$$

□

C.2 Proof of Lemma 4.1

Proof. First note that a d -dimensional Student's t -distribution $\varphi_{\mu,\sigma^2} \sim t(\mu, \sigma^2 I, \alpha)$ satisfies

$$\nabla^2 \log \varphi_{\mu,\sigma^2}(x) = (\alpha + d) \left(\frac{I}{\alpha\sigma^2 + \|x - \mu\|^2} - \frac{(x - \mu)(x - \mu)^\top}{(\alpha\sigma^2 + \|x - \mu\|^2)^2} \right).$$

Hence, we have

$$\begin{aligned} \nabla^2 \log \varphi_{\mu,\sigma^2}(x) &\preccurlyeq \frac{\alpha + d}{\alpha\sigma^2} I, \\ \nabla^2 \log \varphi_{\mu,\sigma^2}(x) &\succcurlyeq -\frac{(\alpha + d)\|x - \mu\|^2}{(\alpha\sigma^2 + \|x - \mu\|^2)^2} I \succcurlyeq -\frac{\alpha + d}{2\alpha\sigma^2} I, \end{aligned}$$

which shows that $\nabla \log \varphi_{\mu,\sigma^2}$ is Lipschitz. Let $X_t \sim \mu_t$, $\varphi_{\sigma^2} \sim t(0, \sigma^2 I, \alpha)$ and for simplicity denote $\pi = \pi_{\text{data}}$, we have that

$$\mu_t(x) = \frac{1}{\lambda_t^{d/2}} \pi \left(\frac{x}{\sqrt{\lambda_t}} \right) * \frac{1}{(1 - \lambda_t)^{d/2}} \varphi_{\sigma^2} \left(\frac{x}{\sqrt{1 - \lambda_t}} \right) = \int \frac{1}{\lambda_t^{d/2}} \pi \left(\frac{y}{\sqrt{\lambda_t}} \right) \varphi_{\sigma^2(1 - \lambda_t)}(x - y) dy,$$

where $\varphi_{\sigma^2(1 - \lambda_t)} \sim t(0, \sigma^2(1 - \lambda_t) I, \alpha)$. The Hessian $\nabla^2 \log \mu_t(x)$ is then given by

$$\nabla^2 \log \mu_t(x) = \mathbb{E}_{Y \sim \rho_{t,x}} [\nabla^2 \log \varphi_{\sigma^2(1 - \lambda_t)}(x - Y)] + \text{Cov}_{Y \sim \rho_{t,x}} [\nabla \log \varphi_{\sigma^2(1 - \lambda_t)}(x - Y)],$$

where $\rho_{t,x}(y) \propto \frac{1}{\lambda_t^{d/2}} \pi \left(\frac{y}{\sqrt{\lambda_t}} \right) \varphi_{\sigma^2(1 - \lambda_t)}(x - y)$. Using that $\nabla \log \varphi_{\sigma^2}$ is Lipschitz with constant L_σ , the first term can be as follows

$$-\frac{L_\sigma}{1 - \lambda_t} I \preccurlyeq \mathbb{E}_{Y \sim \rho_{t,x}} [\nabla^2 \log \varphi_{\sigma^2(1 - \lambda_t)}(x - Y)] \preccurlyeq \frac{L_\sigma}{1 - \lambda_t} I.$$

Focusing on the covariance term we have that

$$0 \preccurlyeq \text{Cov}_{Y \sim \rho_{t,x}} [\nabla \log \varphi_{\sigma^2(1-\lambda_t)}(x - Y)] \preccurlyeq \mathbb{E}_{Y \sim \rho_{t,x}} [\|\nabla \log \varphi_{\sigma^2(1-\lambda_t)}(x - Y)\|^2] I.$$

The expectation can be bounded independently of x as follows

$$\begin{aligned} \mathbb{E}_{Y \sim \rho_{t,x}} [\|\nabla \log \varphi_{\sigma^2(1-\lambda_t)}(x - Y)\|^2] &= (\alpha + d)^2 \mathbb{E}_{Y \sim \rho_{t,x}} \left[\frac{\frac{\|x - Y\|^2}{(\alpha \sigma^2(1-\lambda_t))^2}}{\left(1 + \frac{\|x - Y\|^2}{\alpha \sigma^2(1-\lambda_t)}\right)^2} \right] \\ &\leq \frac{(\alpha + d)^2}{2} \mathbb{E}_{Y \sim \rho_{t,x}} \left[\frac{1}{\alpha \sigma^2(1 - \lambda_t) + \|x - Y\|^2} \right] \leq \frac{(\alpha + d)^2}{2\alpha \sigma^2(1 - \lambda_t)}. \end{aligned} \quad (29)$$

Therefore, we have that

$$-\frac{L_\sigma}{1 - \lambda_t} I \preccurlyeq \nabla^2 \log \mu_t(x) \preccurlyeq \left(\frac{L_\sigma}{1 - \lambda_t} + \frac{(\alpha + d)^2}{2\alpha \sigma^2(1 - \lambda_t)} \right) I. \quad (30)$$

On the other hand, under assumption **A8** we have that $\nabla \log \pi$ is L_π -Lipschitz and $\|\nabla \log \pi\|^2 \leq C_\pi$, then

$$\begin{aligned} -\frac{L_\pi}{\lambda_t} I &\preccurlyeq \nabla^2 \log \mu_t(x) = \frac{1}{\lambda_t} \mathbb{E}_{Y \sim \rho_{t,x}} \left[\nabla^2 \log \pi \left(\frac{Y}{\sqrt{\lambda_t}} \right) \right] + \frac{1}{\lambda_t} \text{Cov}_{Y \sim \rho_{t,x}} \left[\nabla \log \pi \left(\frac{Y}{\lambda_t} \right) \right] \\ &\preccurlyeq \frac{1}{\lambda_t} \left(L_\pi + \mathbb{E}_{\rho_{t,x}} \left[\left\| \nabla \log \pi \left(\frac{Y}{\lambda_t} \right) \right\|^2 \right] \right) I \preccurlyeq \frac{L_\pi + C_\pi}{\lambda_t} I, \end{aligned}$$

where $\rho_{t,x}(y) \propto \frac{1}{\lambda_t^{d/2}} \pi \left(\frac{y}{\sqrt{\lambda_t}} \right) \varphi_{\sigma^2(1-\lambda_t)}(x - y)$. Putting this together with (30), we obtain

$$\max \left\{ -\frac{L_\sigma}{1 - \lambda_t}, -\frac{L_\pi}{\lambda_t} \right\} I \preccurlyeq \nabla^2 \log \mu_t(x) \preccurlyeq \min \left\{ \left(\frac{L_\sigma}{1 - \lambda_t} + \frac{(\alpha + d)^2}{2\alpha \sigma^2(1 - \lambda_t)} \right), \left(\frac{L_\pi + C_\pi}{\lambda_t} \right) \right\} I, \quad (31)$$

which concludes that $\nabla \log \mu_t$ is Lipschitz for all $t \in [0, T]$ with constant

$$L_t \leq \min \left\{ \left(\frac{L_\sigma}{1 - \lambda_t} + \frac{(\alpha + d)^2}{2\alpha \sigma^2(1 - \lambda_t)} \right), \left(\frac{L_\pi + C_\pi}{\lambda_t} \right) \right\}.$$

We now prove the second part of Lemma 4.1, that is, Assumption **A8** is satisfied when $\pi(x) = \tilde{\pi} * \varphi_{\tau^2}(x)$, where $\tilde{\pi}$ is compactly supported and $\varphi_{\tau^2} \sim t(0, \tau^2 I, \tilde{\alpha})$ (Assumption **A9**). In this case, we can write

$$\|\nabla \log \pi(x)\|^2 = \|\mathbb{E}_{Y \sim \hat{\rho}_{t,x}} [\nabla \log \varphi_{\tau^2}(x - Y)]\|^2 \leq \mathbb{E}_{Y \sim \hat{\rho}_{t,x}} [\|\nabla \log \varphi_{\tau^2}(x - Y)\|^2] \leq \frac{(\alpha + d)^2}{2\tilde{\alpha}\tau^2} = C_\pi,$$

where $\hat{\rho}_{t,x} \propto \hat{\pi}(y) \varphi_{\tau^2}(x - y)$ and we have used the same trick as in (29). Denote by L_τ the Lipschitz constant of φ_{τ^2} . The Hessian can be upper and lower bounded as follows

$$\begin{aligned} -L_\tau I &\preccurlyeq \nabla^2 \log \pi(x) = \mathbb{E}_{Y \sim \hat{\rho}_{t,x}} [\nabla^2 \log \varphi_{\tau^2}(x - Y)] + \text{Cov}_{Y \sim \hat{\rho}_{t,x}} [\nabla \log \varphi_{\tau^2}(x - Y)] \\ &\preccurlyeq \left(L_\tau + \mathbb{E}_{Y \sim \hat{\rho}_{t,x}} [\|\nabla \log \varphi_{\tau^2}(x - Y)\|^2] \right) I \leq \left(L_\tau + \frac{(\tilde{\alpha} + d)^2}{2\tilde{\alpha}\tau^2} \right) I, \end{aligned}$$

which shows that $\nabla \log \pi$ is Lipschitz with constant $L_\pi = L_\tau + \frac{(\tilde{\alpha} + d)^2}{2\tilde{\alpha}\tau^2}$. Finally, observe that exploiting Assumption **A9** we can get a more refined Lipschitz constant for $\nabla \log \mu_t$ than that of (31). That is,

$$-\frac{L_\tau}{\lambda_t} \preccurlyeq \nabla^2 \log \mu_t(x) = \mathbb{E}_{Y \sim \tilde{\rho}_{t,x}} [\nabla^2 \log \varphi_{\tau^2 \lambda_t}(x - Y)] + \text{Cov}_{Y \sim \tilde{\rho}_{t,x}} [\nabla \log \varphi_{\tau^2 \lambda_t}(x - Y)] \preccurlyeq \left(\frac{L_\tau}{\lambda_t} + \frac{(\tilde{\alpha} + d)^2}{2\tilde{\alpha}\tau^2 \lambda_t} \right) I,$$

where $\tilde{\rho}_{t,x}(y) \propto \left(\frac{1}{\lambda_t^{d/2}} \pi \left(\frac{y}{\sqrt{\lambda_t}} \right) * \frac{1}{(1 - \lambda_t)^{d/2}} \varphi_{\sigma^2} \left(\frac{y}{\sqrt{1 - \lambda_t}} \right) \right) \varphi_{\tau^2 \lambda_t}(x - y)$. This combined with (30) leads to

$$\max \left\{ -\frac{L_\sigma}{1 - \lambda_t}, -\frac{L_\tau}{\lambda_t} \right\} I \preccurlyeq \nabla^2 \log \mu_t(x) \preccurlyeq \min \left\{ \left(\frac{L_\sigma}{1 - \lambda_t} + \frac{(\alpha + d)^2}{2\alpha \sigma^2(1 - \lambda_t)} \right), \left(\frac{L_\tau}{\lambda_t} + \frac{(\tilde{\alpha} + d)^2}{2\tilde{\alpha}\tau^2 \lambda_t} \right) \right\} I,$$

which shows that $\nabla \log \mu_t$ is Lipschitz for all $t \in [0, T]$. \square

C.3 Proof of Lemma 4.2

Proof. Consider the reparametrised version of μ_t in terms of the schedule λ_t , denoted as $\tilde{\mu}_\lambda$ and let $X_\lambda \sim \tilde{\mu}_\lambda$ and $X_{\lambda+\delta} \sim \tilde{\mu}_{\lambda+\delta}$. Recall that

$$X_\lambda = \sqrt{\lambda}X + \sqrt{1-\lambda}\sigma Z$$

where $X \sim \pi_{\text{data}}$ and $Z \sim t(0, I, \alpha)$. The Wasserstein-2 distance between $\tilde{\mu}_\lambda$ and $\tilde{\mu}_{\lambda+\delta}$ is given by

$$\begin{aligned} W_2^2(\tilde{\mu}_\lambda, \tilde{\mu}_{\lambda+\delta}) &\leq \mathbb{E} [\|X_\lambda - X_{\lambda+\delta}\|^2] \\ &= \mathbb{E} \left[\left\| (\sqrt{\lambda} - \sqrt{\lambda+\delta})X \right\|^2 \right] + \mathbb{E} \left[\left\| (\sqrt{1-\lambda} - \sqrt{1-\lambda-\delta})\sigma Z \right\|^2 \right] \\ &= (\sqrt{\lambda} - \sqrt{\lambda+\delta})^2 \mathbb{E} [\|X\|^2] + (\sqrt{1-\lambda} - \sqrt{1-\lambda-\delta})^2 \frac{\sigma^2 d\alpha}{\alpha-2}. \end{aligned}$$

Using the definition of the metric derivative we have

$$|\dot{\tilde{\mu}}|_\lambda^2 = \lim_{\delta \rightarrow 0} \frac{W_2^2(\tilde{\mu}_\lambda, \tilde{\mu}_{\lambda+\delta})}{\delta^2} \leq \frac{\mathbb{E} [\|X\|^2]}{4\lambda} + \frac{1}{4(1-\lambda)} \frac{\sigma^2 d\alpha}{\alpha-2}.$$

Since $\mu_t = \tilde{\mu}_{\lambda_t}$, we have that $|\dot{\mu}|_t = |\dot{\tilde{\mu}}|_\lambda |\partial_t \lambda_t|$. Using assumption **A10** for the schedule, we have the following expression for the action

$$\begin{aligned} \mathcal{A}_\lambda(\mu) &= \int_0^T |\dot{\mu}|_t^2 dt = \int_0^T |\dot{\tilde{\mu}}|_\lambda^2 |\partial_t \lambda_t|^2 dt \\ &\lesssim \int_0^T \left(\frac{\mathbb{E} [\|X\|^2]}{4\lambda_t} + \frac{\sigma^2 \alpha}{4(1-\lambda_t)(\alpha-2)} d \right) |\partial_t \lambda_t|^2 dt \\ &= \int_0^T \left(\frac{\mathbb{E} [\|X\|^2] \sqrt{1-\lambda_t}}{4\sqrt{\lambda_t}} + \frac{\sigma^2 \alpha \sqrt{\lambda_t}}{4\sqrt{1-\lambda_t}(\alpha-2)} d \right) \left| \frac{\partial_t \lambda_t}{\sqrt{\lambda_t(1-\lambda_t)}} \right| |\partial_t \lambda_t| dt \\ &\leq C_\lambda \int_0^T \left(\frac{\mathbb{E} [\|X\|^2] \sqrt{1-\lambda_t}}{4\sqrt{\lambda_t}} + \frac{\sigma^2 \alpha \sqrt{\lambda_t}}{4\sqrt{1-\lambda_t}(\alpha-2)} d \right) |\partial_t \lambda_t| dt \\ &\leq C_\lambda \int_0^1 \left(\frac{\mathbb{E} [\|X\|^2] \sqrt{1-\lambda}}{4\sqrt{\lambda}} + \frac{\sigma^2 \alpha \sqrt{\lambda}}{4\sqrt{1-\lambda}(\alpha-2)} d \right) d\lambda \\ &\leq \frac{C_\lambda \pi}{8} \left(\mathbb{E} [\|X\|^2] + \frac{\sigma^2 d\alpha}{\alpha-2} \right). \end{aligned}$$

□

C.4 Proof of Theorem 4.3

Lemma C.2. Suppose that $p(x) \propto e^{-V(x)}$ is a probability density on \mathbb{R}^d , where $\nabla V(x)$ is Lipschitz continuous with constant L and let $\varphi_{\sigma^2}(x)$ be the density function of a Student's t distribution $t(0, \sigma^2 I, \alpha)$. Then

$$\left\| \nabla \log \frac{p(x)}{p * \varphi_{\sigma^2}(x)} \right\| \leq L \mathbb{E}_{Y|x} [\|Y - x\|],$$

where the distribution of $Y|x$ is of the form $\frac{p(y)\varphi_{\sigma^2}^2(x-y)}{p(x)*\varphi_{\sigma^2}^2(x)}$.

Proof. Observe that

$$\nabla \log p * \varphi_{\sigma^2}(x) = - \frac{\int_{\mathbb{R}^d} \nabla V(y) e^{-V(y)} \left(1 + \frac{1}{\alpha} \frac{\|y-x\|^2}{\sigma^2} \right)^{-\frac{\alpha+d}{2}} dy}{\int_{\mathbb{R}^d} e^{-V(y)} \left(1 + \frac{1}{\alpha} \frac{\|y-x\|^2}{\sigma^2} \right)^{-\frac{\alpha+d}{2}} dy} = - \frac{\int_{\mathbb{R}^d} \nabla V(y) p(y) \varphi_{\sigma^2}(x-y) dy}{\int_{\mathbb{R}^d} p(y) \varphi_{\sigma^2}(x-y) dy} = -\mathbb{E}_{\gamma_{x,\sigma^2}} [\nabla V(Y)],$$

where γ_{x,σ^2} denotes the probability density

$$\gamma_{x,\sigma^2}(y) = \frac{p(y)\varphi_{\sigma^2}(x-y)}{p(x)*\varphi_{\sigma^2}(x)}.$$

Using the Lipschitzness of ∇V , we have

$$\left\| \nabla \log \frac{p(x)}{p * \varphi_{\sigma^2}(x)} \right\| = \left\| \mathbb{E}_{\gamma_{x,\sigma^2}} [\nabla V(Y) - \nabla V(x)] \right\| \leq L \mathbb{E}_{\gamma_{x,\sigma^2}} [\|Y - x\|].$$

□

Lemma C.3. *With the setting in Lemma C.2. Denote $p_\lambda(x) = \lambda^d p(\lambda x)$ for $\lambda \geq 1$. Then*

$$\left\| \nabla \log \frac{p(x)}{p_\lambda * \varphi_{\sigma^2}(x)} \right\| \lesssim \lambda(\lambda-1)\|x\| + (\lambda-1)\|\nabla V(x)\| + \lambda^2 L \mathbb{E}_{Y|x} [\|Y - x\|],$$

where the law of $Y|x$ is given by $\frac{p_\lambda(y)\varphi_{\sigma^2}^2(x-y)}{p_\lambda(x)*\varphi_{\sigma^2}^2(x)}$.

Proof. Using the triangle inequality,

$$\left\| \nabla \log \frac{p(x)}{p_\lambda * \varphi_{\sigma^2}(x)} \right\| \leq \left\| \nabla \log \frac{p(x)}{p_\lambda(x)} \right\| + \left\| \nabla \log \frac{p_\lambda(x)}{p_\lambda * \varphi_{\sigma^2}(x)} \right\|.$$

The first term can be bounded as

$$\begin{aligned} \left\| \nabla \log \frac{p(x)}{p_\lambda(x)} \right\| &= \|\lambda \nabla V(\lambda x) - \nabla V(x)\| \leq \|\lambda \nabla V(\lambda x) - \lambda \nabla V(x)\| + \|\lambda \nabla V(x) - \nabla V(x)\| \\ &\leq \lambda(\lambda-1)\|x\| + (\lambda-1)\|\nabla V(x)\|. \end{aligned}$$

By the result in Lemma C.2, we have the following bound for the second term

$$\left\| \nabla \log \frac{p_\lambda(x)}{p_\lambda * \varphi_{\sigma^2}(x)} \right\| \lesssim \lambda^2 L \mathbb{E}_{Y|x} [\|Y - x\|],$$

where we have used that $\lambda \nabla V(\lambda x)$ is $\lambda^2 L$ -Lipschitz and $Y|x$ has a distribution of the form

$$\frac{p_\lambda(y)\varphi_{\sigma^2}^2(x-y)}{p_\lambda(x)*\varphi_{\sigma^2}^2(x)}.$$

□

Proof of Theorem 4.3. Similarly to the proof of Theorem 3.4, using Girsanov's theorem, we have that the following bound for $\text{KL}(\mathbb{P}||\mathbb{Q})$.

$$\begin{aligned} \text{KL}(\mathbb{P}||\mathbb{Q}) &= \frac{1}{4} \int_0^{T/\kappa} \mathbb{E}_{\mathbb{P}} \left[\left\| \nabla \log \hat{\mu}_t(X_t) - \nabla \log \hat{\mu}_{t-}(X_{t-}) + v_t(X_t) \right\|^2 \right] dt \\ &\leq \sum_{l=1}^M \int_{t_{l-1}}^{t_l} L_{\kappa t}^2 \mathbb{E}_{\mathbb{P}} \left[\|X_t - X_{t-}\|^2 \right] dt + \int_0^{T/\kappa} \|v_t\|_{L^2(\hat{\mu}_t)}^2 dt + \int_0^{T/\kappa} \mathbb{E}_{\mathbb{P}} \left[\left\| \nabla \log \frac{\hat{\mu}_t(X_{t-})}{\hat{\mu}_{t-}(X_{t-})} \right\|^2 \right] dt, \quad (32) \end{aligned}$$

where we have used that $\nabla \log \hat{\mu}_t$ is Lipschitz with constant $L_{\kappa t}$. First, we bound the change in the score function $\mathbb{E}_{\mathbb{P}} \left[\left\| \nabla \log \frac{\hat{\mu}_t(X_{t-})}{\hat{\mu}_{t-}(X_{t-})} \right\|^2 \right]$. Let $t \geq t_-$, we can write

$$\hat{\mu}_{t-} = T \sqrt{\frac{\lambda_{\kappa t}}{\lambda_{\kappa t-}}} \# \hat{\mu}_t * t \left(0, \left(\sqrt{1 - \lambda_{\kappa t-}} - \sqrt{\frac{(1 - \lambda_{\kappa t})\lambda_{\kappa t-}}{\lambda_{\kappa t}}} \right)^2 \sigma^2 I, \alpha \right) = T \sqrt{\frac{\lambda_{\kappa t}}{\lambda_{\kappa t-}}} \# \hat{\mu}_t * t (0, \gamma_t \sigma^2 I, \alpha),$$

where the pushforward $T_\lambda \#$ is defined as $T_\lambda \# \mu(x) = \lambda^d \mu(\lambda x)$ and we have abused notation by identifying $t(0, \gamma_t \sigma^2 I, \alpha)$ with its density function. Using the result in Lemma C.3, we have

$$\begin{aligned} \left\| \nabla \log \frac{\hat{\mu}_t(X_{t_-})}{\hat{\mu}_{t_-}(X_{t_-})} \right\|^2 &\lesssim \frac{\lambda_{\kappa t}}{\lambda_{\kappa t_-}} \left(\sqrt{\frac{\lambda_{\kappa t}}{\lambda_{\kappa t_-}}} - 1 \right)^2 \|X_{t_-}\|^2 + \left(\sqrt{\frac{\lambda_{\kappa t}}{\lambda_{\kappa t_-}}} - 1 \right)^2 \|\nabla \log \hat{\mu}_t(X_{t_-})\|^2 \\ &+ \left(\frac{\lambda_{\kappa t}}{\lambda_{\kappa t_-}} \right)^2 L_{\kappa t}^2 \mathbb{E}_{Y|X_{t_-}} [\|Y - X_{t_-}\|^2], \end{aligned}$$

where the distribution of $Y|X_t$ is given by

$$Y|X_{t_-} \sim \frac{T \sqrt{\frac{\lambda_{\kappa t}}{\lambda_{\kappa t_-}}} \# \hat{\mu}_t(y) \varphi_{\gamma_t \sigma^2}(X_{t_-} - y)}{T \sqrt{\frac{\lambda_{\kappa t}}{\lambda_{\kappa t_-}}} \# \hat{\mu}_t * \varphi_{\gamma_t \sigma^2}(X_{t_-})} = \frac{T \sqrt{\frac{\lambda_{\kappa t}}{\lambda_{\kappa t_-}}} \# \hat{\mu}_t(y) \varphi_{\gamma_t \sigma^2}(X_{t_-} - y)}{\hat{\mu}_{t_-}(X_{t_-})},$$

where $\varphi_{\gamma_t \sigma^2}$ is the density function of a Student's t distribution of the form $t(0, \gamma_t \sigma^2 I, \alpha)$. Therefore, we have that

$$\mathbb{E}_{\mathbb{P}} \left[\mathbb{E}_{Y|X_{t_-}} [\|Y - X_{t_-}\|^2] \right] = \mathbb{E}_{X_{t_-}, Y} \|Y - X_{t_-}\|^2,$$

where the joint distribution of $(X_{t_-}, Y) \sim \rho_{(X_{t_-}, Y)}(x, y)$ is of the form

$$\rho_{(X_{t_-}, Y)}(x, y) \propto T \sqrt{\frac{\lambda_{\kappa t}}{\lambda_{\kappa t_-}}} \# \hat{\mu}_t(y) \varphi_{\gamma_t \sigma^2}(x - y).$$

Using a change of measure, it follows that Y is independent of $X_{t_-} - Y$ and the distribution of $X_{t_-} - Y$ is $t(0, \gamma_t \sigma^2 I, \alpha)$ with $\alpha > 2$. This results into

$$\mathbb{E}_{\mathbb{P}} \left[\mathbb{E}_{Y|X_{t_-}} [\|Y - X_{t_-}\|^2] \right] = \mathbb{E}_{Z \sim t(0, \gamma_t \sigma^2 I, \alpha)} [\|Z\|^2] = \gamma_t \sigma^2 d \frac{\alpha}{\alpha - 2}.$$

By assumption on the schedule

$$\frac{\lambda_{\kappa t_-}}{\lambda_{\kappa t}} = \mathcal{O}(1 + h_l), \quad \left(\sqrt{\frac{\lambda_{\kappa t}}{\lambda_{\kappa t_-}}} - 1 \right)^2 = \mathcal{O}(h_l^2), \quad \gamma_t \lesssim 1 - \frac{\lambda_{\kappa t_-}}{\lambda_{\kappa t}} = \mathcal{O}(h_l).$$

Given that $X_t = \sqrt{\lambda_t} X + \sqrt{1 - \lambda_t} \sigma^2 Z$ for $X_t \sim \hat{\mu}_t$, we derive the following moment bound

$$\mathbb{E}_{\mathbb{P}} \left[\|X_{t_-}\|^2 \right] = \mathbb{E}_{\mathbb{P}} \left[\left\| \sqrt{\lambda_{\kappa t_-}} X + \sqrt{1 - \lambda_{\kappa t_-}} Z \right\|^2 \right] = \lambda_{\kappa t_-} \mathbb{E}_{\pi_{\text{data}}} [\|X\|^2] + (1 - \lambda_{\kappa t_-}) \sigma^2 \frac{\alpha d}{\alpha - 2} \lesssim \mathbb{E}_{\pi_{\text{data}}} [\|X\|^2] + d.$$

Similarly to the proof of Theorem 3.4, it holds that

$$\mathbb{E}_{\mathbb{P}} \left[\left\| \nabla \log \hat{\mu}_t(X_{t_-}) \right\|^2 \right] \leq L_{\kappa t} d + L_{\kappa t}^2 \mathbb{E}_{\mathbb{P}} \left[\|X_t - X_{t_-}\|^2 \right].$$

This results into

$$\mathbb{E}_{\mathbb{P}} \left[\left\| \nabla \log \frac{\hat{\mu}_t(X_{t_-})}{\hat{\mu}_{t_-}(X_{t_-})} \right\|^2 \right] \lesssim h_l^2 \left(\mathbb{E}_{\pi_{\text{data}}} [\|X\|^2] + d \right) + d h_l^2 L_{\kappa t} + h_l^2 L_{\kappa t}^2 \mathbb{E}_{\mathbb{P}} \left[\|X_t - X_{t_-}\|^2 \right] + h_l L_{\kappa t}^2 \sigma^2 d \frac{\alpha}{\alpha - 2}.$$

Substituting this expression into (32), we have

$$\begin{aligned} \text{KL}(\mathbb{P} || \mathbb{Q}) &\lesssim \sum_{l=1}^M \int_{t_{l-1}}^{t_l} L_{\kappa t}^2 \mathbb{E}_{\mathbb{P}} \left[\|X_t - X_{t_-}\|^2 \right] dt + \int_0^{T/\kappa} \|v_t\|_{L^2(\hat{\mu}_t)}^2 dt \\ &+ \sum_{l=1}^M \int_{t_{l-1}}^{t_l} \left(d h_l^2 L_{\kappa t} + h_l^2 \left(\mathbb{E}_{\pi_{\text{data}}} [\|X\|^2] + d \right) + h_l L_{\kappa t}^2 \sigma^2 d \frac{\alpha}{\alpha - 2} \right) dt. \end{aligned}$$

Using the bound derived in (27), it follows

$$\begin{aligned} \text{KL}(\mathbb{P} \parallel \mathbb{Q}) &\lesssim \sum_{l=1}^M \left(1 + h_l^2 \max_{[t_{l-1}, t_l]} L_t^2 \right) \int_{t_{l-1}}^{t_l} \left| \dot{\hat{\mu}}_t \right|^2 dt + \left(dh_l \int_{t_{l-1}}^{t_l} L_{\kappa t}^2 dt \right) \left(1 + h_l \max_{[t_{l-1}, t_l]} L_t \right) \\ &\quad + \sum_{l=1}^M \int_{t_{l-1}}^{t_l} \left(dh_l^2 L_{\kappa t} + h_l^2 \left(\mathbb{E}_{\pi_{\text{data}}} [\|X\|^2] + d \right) + h_l L_{\kappa t}^2 \sigma^2 d \frac{\alpha}{\alpha-2} \right) dt. \end{aligned}$$

Let $h = \max_{l \in \{1, \dots, M\}} h_l$, we can further simplify the previous expression to obtain

$$\begin{aligned} \text{KL}(\mathbb{P} \parallel \mathbb{Q}) &\lesssim \sum_{l=1}^M (1 + h^2 L_{\max}^2) \int_{t_{l-1}}^{t_l} \left| \dot{\hat{\mu}}_t \right|^2 dt + d h (1 + h L_{\max}) \int_{t_{l-1}}^{t_l} L_{\kappa t}^2 dt \\ &\quad + \frac{T}{\kappa} h^2 (\mathbb{E}_{\pi_{\text{data}}} [\|X\|^2] + d) + d h \frac{\sigma^2 \alpha}{\alpha-2} \int_0^{T/\kappa} L_{\kappa t}^2 dt \\ &\lesssim (1 + h^2 L_{\max}^2) \int_0^{T/\kappa} \left| \dot{\hat{\mu}}_t \right|^2 dt + d h \left(1 + \frac{\sigma^2 \alpha}{\alpha-2} + h L_{\max} \right) \int_0^{T/\kappa} L_{\kappa t}^2 dt + \frac{T}{\kappa} h^2 (\mathbb{E}_{\pi_{\text{data}}} [\|X\|^2] + d) \\ &= (1 + h^2 L_{\max}^2) \kappa \mathcal{A}_\lambda(\mu) + \frac{d h}{\kappa} (1 + h L_{\max}) \int_0^T L_t^2 dt + \frac{T}{\kappa} h^2 (\mathbb{E}_{\pi_{\text{data}}} [\|X\|^2] + d). \end{aligned}$$

The step size h can be expressed in terms of the number of steps M and κ as $h \asymp \frac{1}{M\kappa}$. Therefore, we have

$$\begin{aligned} \text{KL}(\mathbb{P} \parallel \mathbb{Q}) &\lesssim \left(1 + \frac{L_{\max}^2}{M^2 \kappa^2} \right) \kappa \mathcal{A}_\lambda(\mu) + \frac{d}{M \kappa^2} \left(1 + \frac{\sigma^2 \alpha}{\alpha-2} + \frac{L_{\max}}{M \kappa} \right) \int_0^T L_t^2 dt + \frac{1}{M^2 \kappa^3} (\mathbb{E}_{\pi_{\text{data}}} [\|X\|^2] + d) \\ &\lesssim \left(1 + \frac{L_{\max}^2}{M^2 \kappa^2} + \frac{1}{M^2 \kappa^4} \right) \kappa (\mathbb{E}_{\pi_{\text{data}}} [\|X\|^2] + d) + \frac{d}{M \kappa^2} \left(1 + \frac{\sigma^2 \alpha}{\alpha-2} + \frac{L_{\max}}{M \kappa} \right) \int_0^T L_t^2 dt, \end{aligned}$$

where we have used the bound on the action obtained in Lemma 4.2 and $T = \mathcal{O}(1)$. To conclude, note that

$$\begin{aligned} \text{KL}(\mathbb{P} \parallel \mathbb{Q}_\theta) &\lesssim \int_0^{T/\kappa} \mathbb{E}_{\mathbb{P}} \left[\left\| \nabla \log \hat{\mu}_t(X_t) - \nabla \log \hat{\mu}_{t-}(X_{t-}) + v_t(X_t) \right\|^2 \right] dt \\ &\quad + \int_0^{T/\kappa} \mathbb{E}_{\mathbb{P}} \left[\left\| \nabla \log \hat{\mu}_{t-}(X_{t-}) - s_\theta(X_{t-}, t_-) \right\|^2 \right] dt \\ &= \text{KL}(\mathbb{P} \parallel \mathbb{Q}) + \sum_{l=0}^{M-1} h_l \mathbb{E}_{\hat{\mu}_t} \left[\left\| \nabla \log \hat{\mu}_l(X_{t_l}) - s_\theta(X_{t_l}, t_l) \right\|^2 \right] = \text{KL}(\mathbb{P} \parallel \mathbb{Q}) + \varepsilon_{\text{score}}^2 \\ &\lesssim \left(1 + \frac{L_{\max}^2}{M^2 \kappa^2} + \frac{1}{M^2 \kappa^4} \right) \kappa (\mathbb{E}_{\pi_{\text{data}}} [\|X\|^2] + d) + \frac{d L_{\max}^2}{M \kappa^2} \left(1 + \frac{\sigma^2 \alpha}{\alpha-2} + \frac{L_{\max}}{M \kappa} \right) + \varepsilon_{\text{score}}^2 \\ &\lesssim \left(1 + \frac{L_{\max}^2}{M^2 \kappa^2} + \frac{1}{M^2 \kappa^4} \right) \kappa (M_2 \vee d) + \frac{d L_{\max}^2}{M \kappa^2} \left(1 + \frac{\alpha}{\alpha-2} + \frac{L_{\max}}{M \kappa} \right) + \varepsilon_{\text{score}}^2. \end{aligned}$$

We can conclude that by taking

$$\kappa = \mathcal{O} \left(\frac{\varepsilon_{\text{score}}^2}{M_2 \vee d} \right), \quad M = \mathcal{O} \left(\frac{d(M_2 \vee d)^2 L_{\max}^2}{\varepsilon_{\text{score}}^6} \right),$$

we have that $\text{KL}(\mathbb{P} \parallel \mathbb{Q}_\theta) \lesssim \varepsilon_{\text{score}}^2$. Therefore, for any $\varepsilon = \mathcal{O}(\varepsilon_{\text{score}})$, the heavy-tailed DALMC algorithm requires at most

$$M = \mathcal{O} \left(\frac{d(M_2 \vee d)^2 L_{\max}^2}{\varepsilon^6} \right)$$

steps to approximate π_{data} to within ε^2 in KL divergence. \square