An analysis of the noise schedule for score-based generative models

Stanislas Strasman, Antonio Ocello, Claire Boyer, Sylvain Le Corff, Vincent Lemaire

An analysis of the noise schedule for score-based generative models

Stanislas Strasman[1], Antonio Ocello[2], Claire Boyer[1,3], Sylvain Le Corff[1], and Vincent Lemaire[1]

[1]Sorbonne Université and Université Paris Cité, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, F-75005 Paris, France
[2]CMAP, Ecole Polytechnique
[3]Institut Universitaire de France (IUF)

Abstract

Score-based generative models (SGMs) aim at estimating a target data distribution by learning score functions using only noise-perturbed samples from the target. Recent literature has focused extensively on assessing the error between the target and estimated distributions, gauging the generative quality through the Kullback-Leibler (KL) divergence and Wasserstein distances. All existing results have been obtained so far for time-homogeneous speed of the noise schedule. Under mild assumptions on the data distribution, we establish an upper bound for the KL divergence between the target and the estimated distributions, explicitly depending on any time-dependent noise schedule. Assuming that the score is Lipschitz continuous, we provide an improved error bound in Wasserstein distance, taking advantage of favourable underlying contraction mechanisms. We also propose an algorithm to automatically tune the noise schedule using the proposed upper bound. We illustrate empirically the performance of the noise schedule optimization in comparison to standard choices in the literature.

1 Introduction

Recent years have seen impressive advances in machine learning and artificial intelligence, with one of the most notable breakthroughs being the success of diffusion models, introduced by Sohl-Dickstein et al. (2015). Diffusion models in generative modeling refer to a class of algorithms that generate new samples given training samples of an unknown distribution π_{data}. This method is now recognized for its ability to produce high-quality images that appear genuine to human observers (see *e.g.*, Ramesh et al., 2022, for text-to-image generation). Its range of applications is expanding rapidly,

yielding impressive outcomes in areas such as computer vision Li et al. (2022); Lugmayr et al. (2022) or natural language generation Gong et al. (2023), among others, see Yang et al. (2023) for a comprehensive overview of the latest advances in this topic.

Score-based generative models (SGMs). Generative diffusion models aim at creating synthetic instances of a target distribution when only a genuine sample (*e.g.*, a dataset of real-life images) is accessible. It is crucial to note that the complexity of real data prohibits a thorough depiction of the distribution π_{data} through a conventional parametric model, and its estimation via traditional maximum likelihood methods. Standard strategies based on non-parametric density estimation such as kernel smoothing are also generally ruled out due to the high dimensionality of the data in play.

Score-based Generative Models (SGMs) are probabilistic models designed to address this challenge using two main phases. The first phase, the noising phase (also referred to as the forward phase), involves progressively perturbing the empirical distribution by adding noise to the training data until its distribution approximately reaches an easy-to-sample distribution π_∞. The second phase involves learning to reverse this noising dynamics by sequentially removing the noise, which is referred to as the sampling phase (or backward phase). Reversing the dynamics during the backward phase would require in principle knowledge of the score function, *i.e.*, the gradient of the logarithm of the density at each time step of the diffusion. However, knowing the score amounts to knowing the distribution at time $t = 0$, *i.e.*, knowing the distribution π_{data} according to which we wish to simulate new examples. To circumvent this issue, the score function is learned based on the evolution of the noised data samples and using a deep neural network architecture. When applying these learned reverse dynamics to samples from the distribution π_∞, we obtain a generative distribution that approximates π_{data}.

Related works. Significant attention has been paid to understanding the sources of errors that affect the quality of data generation associated with SGMs (Chen et al., 2023a,b; Block et al., 2020; De Bortoli, 2022; Lee et al., 2022, 2023; Benton et al., 2023). In particular, a key area of interest has been the derivation of upper bounds for distances or pseudo-distances between the training and generated sample distributions. Note that all the mathematical theory for diffusion models developed so far covers general time discretizations of time-homogeneous SGMs (see Song and Ermon, 2019, in the variance-preserving case), which means that the strength of the noise is prescribed to be constant during the forward phase. De Bortoli et al. (2021); Chen (2023) provided upper bounds in terms of total variation, by assuming smoothness properties of the score and its derivatives. On the other hand, the upper bounds in total variation and Wasserstein distances provided by Lee et al. (2023) also require smoothness assumptions on the data distribution and involve non-explicit constants. More recently, Conforti et al. (2023) established an upper bound in terms of Kullback–Leibler (KL) divergence avoiding strong assumptions about the score regularity, and relying on mild conditions about the data distribution assumed to be of finite Fisher information w.r.t. the Gaussian distribution. Regarding time-inhomogeneous SGMs, the central role of the noise schedule

has already been exhibited in numerical experiments, see for instance Chen (2023); Nichol and Dhariwal (2021); Anonymous (2023). However, a rigorous theoretical analysis of it is still missing.

Contributions. In this paper, we conduct a thorough mathematical analysis of the role of the noise schedule in score-based generative models.

- We establish an upper bound on the Kullback-Leibler divergence between the data distribution and the law of the SGM. This bound holds under mild assumptions and explicitly depends on the noise schedule used to train the SGM.
- We illustrate, through numerical experiments, the upper bound obtained in practice in regard of the effective empirical KL divergences. These simulations highlights the relevancy of the upper bound, reflecting in practice the effect of the noise schedule on the quality of the generative distribution.
- By making an additional assumption on the Lipschitz property of the score function, we establish a sharper bound of the error due to the mixing time in terms of Wasserstein distance, by leveraging from the contraction of the drift not only of the forward, but also of the backward stochastic diffusion.
- Finally, we propose to exploit the theoretical bound obtained to drive and improve the implementation of SGMs in practice. We indeed suggest a procedure to jointly optimize the score network and the noise schedule using a loss function encompassing the proposed upper bound.

2 A theoretical analysis of the noise schedule in SGMs

In this section, we provide a theoretical analysis of the effect of the noise schedule used when training an SGM. Its impact is theoretically captured through a bound on the KL divergence between the data distribution and the generative one.

2.1 Notation and definitions

Forward process. Denote as $\beta : [0, T] \mapsto \mathbb{R}_{>0}$ the noise schedule, assumed to be continuous and non decreasing. Although originally developed using a finite number of noising steps Sohl-Dickstein et al. (2015); Song and Ermon (2019); Ho et al. (2020); Song et al. (2021b), most recent approaches consider time-continuous noise perturbations through the use of stochastic differential equations (SDEs) Song et al. (2021b). Consider, therefore, a forward process given by

$$\mathrm{d}\overrightarrow{X}_t = -\frac{\beta(t)}{2\sigma^2}\overrightarrow{X}_t \mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}B_t, \quad \overrightarrow{X}_0 \sim \pi_{\mathrm{data}}\,. \tag{1}$$

We denote by p_t the density of $\overrightarrow{X}_t$ at time $t \in (0, T]$. Note that, up to the time change $t \mapsto \int_0^t \beta(s)/2 \mathrm{d}s$, this process corresponds to the standard Ornstein–Uhlenbeck (OU) process, solution to

$$\mathrm{d}\overrightarrow{X}_t = -\frac{1}{\sigma^2}\overrightarrow{X}_t \mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \quad \overrightarrow{X}_0 \sim \pi_{\mathrm{data}}\,,$$

see for instance Karatzas and Shreve (2012, Chapter 3). Due to the linear nature of the drift with respect to $(X_t)_t$, an exact simulation can be performed for this process. The stationary distribution of the forward process is the Gaussian distribution with mean 0 and variance $\sigma^2 \mathrm{I}_d$ and is denoted by π_∞.

Note that in the literature, when $\beta(t)$ is constant equal to 2 (meaning that there is no time change), this diffusion process is referred to as the Variance-Preserving SDE (VPSDE, De Bortoli et al., 2021; Conforti et al., 2023; Chen et al., 2023b), leading to the so-called Denoising Diffusion Probabilistic Models (DDPM, Ho et al., 2020). Understanding the effects of the general diffusion model (1), in particular when reversing the dynamic, remains a challenging problem, to which we devote the rest of our analysis.

Backward process. The corresponding backward process is initialized at the stationary distribution π_∞ and can be written as

$$\mathrm{d}\overleftarrow{X}_t = \eta(t, \overleftarrow{X}_t)\mathrm{d}t + \sqrt{\bar{\beta}(t)}\mathrm{d}B_t, \quad \overleftarrow{X}_0 \sim \pi_\infty\,,$$

where

$$\begin{cases} \bar{\beta}(t) & := \beta(T-t) \\ \eta(t, \overleftarrow{X}_t) & := \bar{\beta}(t)\overleftarrow{X}_t/(2\sigma^2) + \bar{\beta}(t)\nabla \log p_{T-t}\left(\overleftarrow{X}_t\right)\,. \end{cases}$$

We denote by $\mathbb{Q}_T \in \mathcal{P}(C([0,T], \mathbb{R}^d))$ the path measure associated with the backward diffusion. We consider the marginal time distribution of the forward process divided by the density of its stationary distribution, introducing

$$\forall x \in \mathbb{R}^d, \quad \tilde{p}_t(x) = \frac{p_t(x)}{\varphi_{\sigma^2}(x)}\,, \tag{2}$$

where φ_{σ^2} denote the density function of π_∞, a Gaussian distribution with mean 0 and variance $\sigma^2 \mathrm{I}_d$. Thus, the backward process can be rewritten as

$$\mathrm{d}\overleftarrow{X}_t = \bar{\eta}\left(t, \overleftarrow{X}_t\right)\mathrm{d}t + \sqrt{\bar{\beta}(t)}\mathrm{d}B_t, \quad \overleftarrow{X}_0 \sim \pi_\infty\,, \tag{3}$$

where $\bar{\eta}(t, \overleftarrow{X}_t) := -\frac{\bar{\beta}(t)}{2\sigma^2}\overleftarrow{X}_t + \bar{\beta}(t)\nabla \log \tilde{p}_{T-t}(\overleftarrow{X}_t)$. The benefit of using the renormalization $\tilde{p}_t$ in our analysis results in considering the backward equation as a perturbation of an OU process. This trick is crucial to highlight the central role of the relative Fisher information in the performance of the SGM. It has already been used by Conforti et al. (2023).

Score estimation. Simulating the backward process means knowing how to operate the score. However, the (modified) score function $\nabla \log \tilde{p}_t(x) = \nabla \log p_t(x) + x/\sigma^2$ cannot be evaluated directly, because it depends on the unknown data distribution. To work around this problem, the score function $\nabla \log p_t$ needs to be estimated. In Hyvärinen and Dayan (2005), the authors proposed to estimate the score function associated with a distribution by minimizing the expected L^2-squared distance between the true score function and the proposed approximation. In the context of diffusion models, this is typically done with the use of a deep neural network architecture $s_\theta : [0,T] \times \mathbb{R}^d \mapsto \mathbb{R}^d$ parameterized by $\theta \in \Theta$, and trained to minimize:

$$\mathcal{L}_{\text{explicit}}(\theta) = \mathbb{E}\left[\left\| s_\theta\left(\tau, \overrightarrow{X}_\tau\right) - \nabla \log p_\tau\left(\overrightarrow{X}_\tau\right)\right\|^2\right],$$

with $\tau \sim \mathcal{U}(0,T)$ independent of the forward process $\left(\overrightarrow{X}_t\right)_{t\geq 0}$. However, this estimation problem still suffers from the fact that the regression target is not explicitly known. A tractable optimization problem sharing the same optima can be defined though, through the marginalization over π_{data} of p_τ (see Vincent, 2011; Song et al., 2021a):

$$\mathcal{L}_{\text{score}}(\theta) = \mathbb{E}\left[\| s_\theta(\tau, \overrightarrow{X}_\tau) - \nabla \log p_\tau(\overrightarrow{X}_\tau | X_0)\|^2\right],$$

where τ is uniformly distributed on $[0,T]$, and independent of $X_0 \sim \pi_{\text{data}}$ and $\overrightarrow{X}_\tau \sim p_\tau(\cdot|X_0)$. This loss function is appealing as it only requires to know the transition kernel of the forward process. In the classical setting of diffusion models given by (1), this is a Gaussian kernel with explicit mean and variance.

Discretization. Once the score function is learned, it remains that, in most cases, the backward dynamics no longer enjoys a linear drift, which makes its exact simulation challenging. To address this issue, one solution is to discretize the continuous dynamics of the backward process. In this way, Song et al. (2021b) propose an Euler-Maruyama (EM) discretization scheme in which both the drift and the diffusion coefficients are discretized recursively (see 26). In particular, introduce $\tilde{s}_\theta(t,x) := s_\theta(t,x) + x/\sigma^2$ and consider the time discretization $0 =: t_0 \leq t_1 \leq \cdots \leq t_N := T$, the EM scheme corresponds to

$$\mathrm{d}\overleftarrow{X}^{EM}_t = \left(-\frac{\bar{\beta}(t_k)}{2\sigma^2}\overleftarrow{X}^{EM}_{t_k} + \bar{\beta}(t)\tilde{s}_\theta\left(T - t_k, \overleftarrow{X}^{EM}_{t_k}\right)\right)\mathrm{d}t + \sqrt{\bar{\beta}(t_k)}\mathrm{d}B_t\,.$$

The Euler Exponential Integrator (EI) (see Durmus and Moulines, 2015), as already used in Conforti et al. (2023), only requires to discretize the part associated with the modified score function. Let $\left(\overleftarrow{X}^\theta_t\right)_{t\in[0,T]}$ be such that, for $t \in [t_k, t_{k+1}]$,

$$\mathrm{d}\overleftarrow{X}^\theta_t = \bar{\beta}(t)\left(-\frac{1}{2\sigma^2}\overleftarrow{X}^\theta_t + \tilde{s}_\theta\left(T - t_k, \overleftarrow{X}^\theta_{t_k}\right)\right)\mathrm{d}t + \sqrt{\bar{\beta}(t)}\mathrm{d}B_t\,.$$

This scheme can be seen as a refinement of the classical Euler-Maruyama one as it handles the linear drift term by integrating it explicitly. We consider therefore such a scheme in our further theoretical developments.

We denote by $\bar{\mathbb{Q}}_N^{\beta,\theta} \in \mathcal{P}(C([0,T],\mathbb{R}^d))$ the path measure associated with this discretized version of the backward diffusion and by $\widehat{\pi}_N^{(\beta,\theta)}$ the marginal probability density of $\overleftarrow{X}_T^\theta$ under an N-time discretization (recall that $\overleftarrow{X}_0^\theta \sim \pi_\infty$).

2.2 Main result: nonasymptotic Kullback-Leibler upper bound depending on the noise schedule

In this section, we present theoretical guarantees on time-inhomogeneous SGMs with an explicit dependency on the noise schedule $t \mapsto \beta(t)$.

Statement. The data distribution π_{data} is assumed to be absolutely continuous with respect to the Gaussian measure π_∞. Define the relative Fisher information $\mathcal{I}(\pi_{\text{data}}|\pi_\infty)$ by

$$\mathcal{I}(\pi_{\text{data}}|\pi_\infty) := \int \left\| \nabla \log \left(\frac{\mathrm{d}\pi_{\text{data}}}{\mathrm{d}\pi_\infty} \right) \right\|^2 \mathrm{d}\pi_{\text{data}} \,,$$

and consider the following assumptions.

H1 The noise schedule is continuous, non decreasing and such that $\int_0^\infty \beta(t)\mathrm{d}t = \infty$.

H2 The data distribution has finite Fisher information w.r.t. the normal distribution, *i.e.*, $\mathcal{I}(\pi_{\text{data}}|\pi_\infty) < \infty$.

H3 The parameter $\theta \in \Theta$ and the schedule β satisfy

$$\mathbb{E}\left[\exp\left\{ \frac{1}{2} \int_0^T \bar{\beta}(t) \left\| \left(\tilde{s}\left(T-t, \overleftarrow{X}_t\right) - \tilde{s}_\theta\left(T-t_k, \overleftarrow{X}_{t_k}\right) \right) \right\|^2 \mathrm{d}t \right\} \right] < \infty \,,$$

where $\tilde{s}(t,x) := \nabla \log \tilde{p}_t(x)$ corresponds to the score function up to the renormalization (2) by the stationary distribution.

Assumption H1 is necessary to ensure that the forward process converges to the stationary distribution when the diffusion time tends to infinity. Assumption H2 is inherent to the data distribution, as it involves only the L^2-integrability of the score function. Such a kind of hypothesis has already been considered in the literature, see Conforti et al. (2023). We stress that we do not require extra assumptions about the smoothness of the score function. Lastly, Assumption H3 is the guarantor of a good approximation of the score by the neural network $\tilde{s}_\theta$, weighted by the level of noise in play. We are now in position to provide an upper bound for the relative entropy between the distribution $\widehat{\pi}_N^{(\beta,\theta)}$ of samples obtained using the discretized reverse-time process, and the target data distribution π_{data}. This theoretical guarantee on the quality of the generated samples explicitly depends on the noise schedule $t \mapsto \beta(t)$.

Theorem 2.1. *Assume that H1, H2 and H3 hold. Then,*

$$\mathrm{KL}\left(\pi_{\mathrm{data}} \middle|\middle| \widehat{\pi}_N^{(\beta,\theta)}\right) \leq \mathcal{E}_1(\beta) + \mathcal{E}_2(\theta,\beta) + \mathcal{E}_3(\beta)\,,$$

where

$$\mathcal{E}_1(\beta) = \mathrm{KL}\left(\pi_{\mathrm{data}} || \pi_\infty\right) \exp\left\{-\frac{1}{\sigma^2}\int_0^T \beta(s)\mathrm{d}s\right\}\,,$$

$$\mathcal{E}_2(\theta,\beta) = \sum_{k=1}^N \mathbb{E}\left[\left\|\nabla \log \tilde{p}_{T-t_k}\left(\overrightarrow{X}_{T-t_k}\right) - \tilde{s}_\theta\left(T-t_k, \overrightarrow{X}_{T-t_k}\right)\right\|^2\right] \int_{t_k}^{t_{k+1}} \beta(t)\mathrm{d}t\,,$$

$$\mathcal{E}_3(\beta) = 2h\beta(T)\max\left\{\frac{h\beta(T)}{4\sigma^2}; 1\right\} \mathcal{I}(\pi_{\mathrm{data}}|\pi_\infty)\,,$$

with $h := \sup_{k\in\{1,\dots,N\}}(t_k - t_{k-1})$ *and* $t_0 := 0$.

The obtained bound is composed of three terms, all depending on the noise schedule β, through either its integrated version over the diffusion time, or its final value at time T. Note that if the result is derived for the EI discretization scheme, it could be adapted to the Euler one up to minor technicalities. Remark also that relying on Pinsker's inequality, the obtained bound could be transferred in terms of total variation. To understand the origin of each term of the upper bound, we propose to give the main ideas of the proof in what follows. Our approach to establish Theorem 2.1 falls into the category of the Girsanov-based approach as in De Bortoli et al. (2021); Chen et al. (2023b); Conforti et al. (2023), adapted to obtain sharp upper bounds in time-inhomogeneous cases.

Elements of proof. We are interested in the relative entropy of the training data distribution π_{data} with respect to the generated data distribution $\widehat{\pi}_N^{(\beta,\theta)}$. Denoting by $(Q_t)_{t\in[0;T]}$ the semi-group of $\overleftarrow{X}_t$ (we drop the dependence on the noise schedule β in the notation for the ease of readability) and leveraging the time-reverse property we have[1]:

$$\mathrm{KL}\left(\pi_{\mathrm{data}} \middle|\middle| \widehat{\pi}_N^{(\beta,\theta)}\right) = \mathrm{KL}\left(p_T Q_T \middle|\middle| \widehat{\pi}_N^{(\beta,\theta)}\right)\,.$$

By the data processing inequality,

$$\mathrm{KL}\left(p_T Q_T \middle|\middle| \widehat{\pi}_N^{(\beta,\theta)}\right) \leq \mathrm{KL}\left(p_T \mathbb{Q}_T \middle|\middle| \pi_\infty \bar{\mathbb{Q}}_N^{\beta,\theta}\right)\,.$$

[1]For any probability density p and any kernel Q, pQ is the probability density given by $pQ : x \mapsto \int p(u)Q(u,x)\mathrm{d}u$ where $Q(u,\cdot)$ is the probability density of $Q(u,\mathrm{d}x)$.

Writing the backward time $\tau_t = T - t$ and its discretized version $\tau_k = T - t_k$, we have (by Lemma B.2) that

$$\begin{aligned}
&\mathrm{KL}\left(\pi_{\mathrm{data}} \| \widehat{\pi}_N^{(\beta,\theta)}\right) \\
&\leq \mathrm{KL}\left(p_T \| \varphi_{\sigma^2}\right) \\
&\quad + \frac{1}{2}\int_0^T \frac{1}{\bar{\beta}(t)} \mathbb{E}\left[\left\| \frac{-\bar{\beta}(t)}{2\sigma^2} \overleftarrow{X}_t + \bar{\beta}(t)\nabla \log \tilde{p}_{\tau_t}\left(\overleftarrow{X}_t\right) \right.\right. \\
&\quad \left.\left. - \left(-\frac{\bar{\beta}(t)}{2\sigma^2}\overleftarrow{X}_t + \bar{\beta}(t)\tilde{s}_\theta\left(\tau_k, \overleftarrow{X}_{t_k}\right)\right)\right\|^2\right] \mathrm{d}t\,.
\end{aligned}$$

From there, the KL divergence can be split into the theoretical mixing time of the forward OU process and the approximation error for the score function made by the neural network, as follows:

$$\begin{aligned}
&\mathrm{KL}\left(\pi_{\mathrm{data}} \| \widehat{\pi}_N^{(\beta,\theta)}\right) \\
&\leq \mathrm{KL}\left(p_T \| \varphi_{\sigma^2}\right) + \frac{1}{2}\int_0^T \frac{1}{\bar{\beta}(t)} \mathbb{E}\left[\left\| \bar{\beta}(t)\left(\tilde{s}\left(\tau_t, \overleftarrow{X}_t\right) - \tilde{s}_\theta(\tau_k, \overleftarrow{X}_{t_k})\right)\right\|^2\right] \mathrm{d}t\,.
\end{aligned}$$

By discretizing the interval $[0, T]$ using $0 = t_0 < t_1 < \ldots < t_N = T$, one can disentangle the last term as follows:

$$\begin{aligned}
&\mathrm{KL}\left(\pi_{\mathrm{data}} \middle\| \widehat{\pi}_N^{(\beta,\theta)}\right) \\
&\quad \leq \mathrm{KL}\left(p_T \| \varphi_{\sigma^2}\right) + \frac{1}{2}\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} \bar{\beta}(t)\mathbb{E}\left[\left\|\tilde{s}\left(\tau_t, \overleftarrow{X}_t\right) - \tilde{s}_\theta\left(\tau_k, \overleftarrow{X}_{t_k}\right)\right\|^2\right] \mathrm{d}t \\
&\quad \leq E_1(\beta) + E_2(\theta,\beta) + E_3(\beta)\,,
\end{aligned}$$

where

$$E_1(\beta) = \mathrm{KL}\left(p_T \| \varphi_{\sigma^2}\right)\,, \tag{4}$$

$$E_2(\theta,\beta) = \sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} \bar{\beta}(t)\mathbb{E}\left[\left\|\tilde{s}\left(\tau_k, \overleftarrow{X}_{t_k}\right) - \tilde{s}_\theta\left(\tau_k, \overleftarrow{X}_{t_k}\right)\right\|^2\right] \mathrm{d}t\,, \tag{5}$$

$$E_3(\beta) = \sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} \bar{\beta}(t)\mathbb{E}\left[\left\|\tilde{s}\left(\tau_t, \overleftarrow{X}_t\right) - \tilde{s}\left(\tau_k, \overleftarrow{X}_{t_k}\right)\right\|^2\right] \mathrm{d}t\,. \tag{6}$$

Finishing the proof of Theorem 2.1 amounts to obtaining upper bounds for $E_1(\beta)$, $E_2(\theta,\beta)$ and $E_3(\beta)$. This is done in Lemmas A.1,A.2 and A.3, so that $E_1(\beta) \leq \mathcal{E}_1(\beta)$, $E_2(\theta,\beta) \leq \mathcal{E}_2(\theta,\beta)$ and $E_3(\beta) \leq \mathcal{E}_3(\beta)$.

Dissecting the upper bound. The upper bound of Theorem 2.1 involves three different types of error that affect the training of an SGM. The term $\mathcal{E}_1$ (or E_1 in the proof) represents the *mixing time* of the OU forward process, arising from the practical limitation of considering the forward process up to a finite time T. Indeed, $\mathcal{E}_1$ is shrinked to 0 when T grows to infinity. Note that the multiplicative term in $\mathcal{E}_1$ corresponds to the KL divergence between π_{data} and π_∞ which is ensured to be finite by Assumption H2. The second term $\mathcal{E}_2$ (or E_2 in the proof) corresponds to the *approximation error*, which stems from the use of a deep neural network to estimate the score function. Note that if we assume that the error of the score approximation is uniformly (in time) bounded by M_θ (see De Bortoli et al., 2021, Equation (8)), the term $\mathcal{E}_2$ admits as a crude bound $M_\theta \int_0^T \beta(t)\mathrm{d}t$, with the disadvantage of exploding when $T \to +\infty$. Otherwise, by considering Conforti et al. (2023, Assumption H3), one can make this bound finer and finite, by balancing the quality of the score approximation, the discretization grid and the final time T. Finally, $\mathcal{E}_3$ (or E_3 in the proof) is the *discretization error* of the EI discretization scheme. This last term vanishes as the discretisation grid is refined (*i.e.*, $h \to 0$).

Comparison with existing bounds. Under perfect score approximation, *i.e.* (with $\tau_k = T - t_k$),

$$\sum_{k=1}^{N} \mathbb{E}\left[\left\| \nabla \log \tilde{p}_{\tau_k}\left(\overrightarrow{X}_{\tau_k}\right) - \tilde{s}_\theta\left(\tau_k, \overrightarrow{X}_{\tau_k}\right) \right\|^2 \right] = 0\,,$$

and infinitely precise discretization, *i.e.*, $h \to 0$, we recover that the Variance Preserving SDE (VPSDE, De Bortoli et al., 2021; Conforti et al., 2023; Chen et al., 2023b) converge exponentially fast to the target distribution. Beyond this idealized setting, the bound established in Theorem 2.1 recovers that of Conforti et al. (2023, Theorem 1) when choosing a constant noise schedule $\beta(t) = 2$, the stationary variance as $\sigma^2 = 1/2$, fixing the final time T at 1, and using a discretization step size $h \leq 1$.

3 On the tightness of the upper bound

3.1 A refined version

In this section, we focus on the framework of "perfect score approximation" and infinitely precise discretization, *i.e.*, $\mathcal{E}_2(\theta, \beta) = \mathcal{E}_3(\theta, \beta) = 0$. This allows to assess the sharpness of the term $\mathcal{E}_1(\beta)$ in the upper bound of Theorem 2.1.

When restricting the data distribution to be Gaussian $\mathcal{N}(\mu_0, \Sigma_0)$, one can exploit the backward contraction assuming that $\lambda_{\max}(\Sigma_0) \leq \sigma^2$, where $\lambda_{\max}(\Sigma_0)$ denotes the largest eigenvalue of Σ_0. In this specific case, we can obtain a refined version for $\mathcal{E}_1$ (see

Proposition C.1), given by

$$\mathrm{KL}\left(\pi_{\mathrm{data}} \| \varphi_{\sigma^2} Q_T\right) \leq \mathrm{KL}\left(\pi_{\mathrm{data}} \| \varphi_{\sigma^2}\right) \exp\left(-\frac{2}{\sigma^2} \int_0^T \beta(s) \mathrm{d}s\right) .$$

In the literature, much attention is paid to derive upper bounds with other metrics such as the (sliced)-Wasserstein. In Lee et al. (2023), the authors obtain a control for the Wasserstein and total variation distances. However, those results rely on strong smoothness assumptions on the score function as in De Bortoli (2022), and with additional assumptions on π_{data} (assumed to have bounded support).

Hereafter, we also propose a control in Wasserstein distance under the following assumption.

H4 For any t, there exists $C_t \geq 0$ such that $\forall x, y \in \mathbb{R}^d$,

$$\left(\nabla \log \tilde{p}_t(x) - \nabla \log \tilde{p}_t(y)\right)^{\top}(x-y) \leq -C_t \|x-y\|^2 .$$

Assumption H4 includes a notion of smoothness for the score accounting for its sign, which plays a crucial role in terms of contraction of the backward SDE. This is a key element that has already been identified in OU processes for instance for which contraction is indeed well-established and serves as the foundation for the convergence of SGMs.

Proposition 3.1. *Suppose that $x \mapsto \nabla \log \tilde{p}_t(x)$ is L_t-Lipschitz, for $t \in (0, T]$. Then,*

$$\mathcal{W}_2\left(\pi_{\mathrm{data}}, \varphi_{\sigma^2} Q_T\right)^2 \leq \mathcal{W}_2\left(p_T, \varphi_{\sigma^2}\right)^2 \exp\left(-\int_0^T \frac{\beta(t)}{\sigma^2}\left(1 - 2L_t\sigma^2\right) \mathrm{d}t\right) . \tag{7}$$

Moreover, under Assumption H4, we have

$$\mathcal{W}_2\left(\pi_{\mathrm{data}}, \varphi_{\sigma^2} Q_T\right)^2 \leq \mathcal{W}_2\left(p_T, \varphi_{\sigma^2}\right)^2 \exp\left(-\int_0^T \frac{\beta(t)}{\sigma^2}\left(1 + 2C_t\sigma^2\right) \mathrm{d}t\right) . \tag{8}$$

Remark that Assumption H4 is a more restrictive hypothesis compared to the Lipschitz continuity of the score, as the former implies the latter. This stringency is reflected in the upper bound, as the contraction strength is always improved under Assumption H4 by involving the term $1 + 2C_t\sigma^2$ instead of $1 - 2L_t\sigma^2$ when the score is only assumed to be Lipschitz. Note that C_t could take negative values to a certain extent, and still preserving the contraction property.

Note also that Assumption H4 or the Lipschitz property of the score are both satisfied when the target distribution is assumed to be Gaussian and provided some conditions on its covariance structure. Indeed, when π_{data} is a Gaussian distribution $\mathcal{N}(\mu_0, \Sigma_0)$, the score can be expressed by a closed-form formula, leading to a fine evaluation of the constant C_t (and L_t).

Lemma 3.2. *Assume that π_{data} is a Gaussian distribution $\mathcal{N}(\mu_0, \Sigma_0)$, satisfying that $\lambda_{\max}(\Sigma_0) \leq \sigma^2$. Then, the error bound (8) holds with a contraction dictated by the following constant*

$$C_t := \frac{m_t^2 \left(\sigma^2 - \lambda_{\max}(\Sigma_0)\right)}{m_t^2 \lambda_{\max}(\Sigma_0) + \sigma^2 \left(1 - m_t^2\right)} .$$

This result, restricted to the Gaussian case, sets the focus on the importance of calibrating the parameter σ^2 depending on the covariance structure of the data distribution, in order to accelerate the convergence speed of the algorithm.

3.2 Numerical illustration

To illustrate the upper bound, we consider the setting where the true distribution is Gaussian in dimension $d = 50$ with mean $\mathbf{1}_d$ and different choices of covariance structure.

1. (Isotropic) $\Sigma^{(\text{iso})} = 0.5\mathrm{I}_d$.
2. (Heteroscedastic) $\Sigma^{(\text{heterosc})} \in \mathbb{R}^{d\times d}$ is a diagonal matrix such that $\Sigma_{jj}^{(\text{heterosc})} = 10$ for $1 \leq j \leq 5$, and $\Sigma_{jj}^{(\text{heterosc})} = 0.1$ otherwise.
3. (Correlated) $\Sigma^{(\text{corr})} \in \mathbb{R}^{d\times d}$ is a full matrix whose diagonal entries are equal to one and the off-diagonal terms are $\Sigma_{jj'}^{(\text{corr})} = 1/\sqrt{|j - j'|}$ for $1 \leq j \neq j' \leq d$.

The resulting data distributions are respectively denoted by $\pi_{\text{data}}^{(\text{iso})}$, $\pi_{\text{data}}^{(\text{heterosc})}$ and $\pi_{\text{data}}^{(\text{corr})}$. Theorem 2.1 provides a generic Kullback-Leibler upper-bound:

$$\mathcal{L}_{\text{sched}}(\theta, \beta) = \mathcal{E}_1(\beta) + \mathcal{E}_2(\theta, \beta) + \mathcal{E}_3(\beta) . \tag{9}$$

We propose to evaluate (9) for the different data distributions above, and for a noise schedule of the form

$$\beta_a(t) \propto (\mathrm{e}^{at} - 1)/(\mathrm{e}^{aT} - 1) , \tag{10}$$

with $a \in \mathbb{R}$ ranging from -10 to 10, see Figure 1. To do so, for each value of a, and each data distribution, we train with $n = 10000$ Gaussian samples an SGM with 200 discretization steps of the time interval $[0, 1]$. In all our numerical experiments, we use an Euler-Maruyama scheme, as being the most encountered in practice. The score is learned using a dense neural network with 3 hidden layers of width 256 over 100 epochs, see Figure 7. We compare the obtained value of (9) with empirical KL divergence between samples from the data distribution and samples from the trained model $\widehat{\pi}_N^{(\beta_a,\theta)}$. Note that in a Gaussian setting, the evaluation of the bound or the KL divergence relies on closed-form formulas; see Appendix D.1.

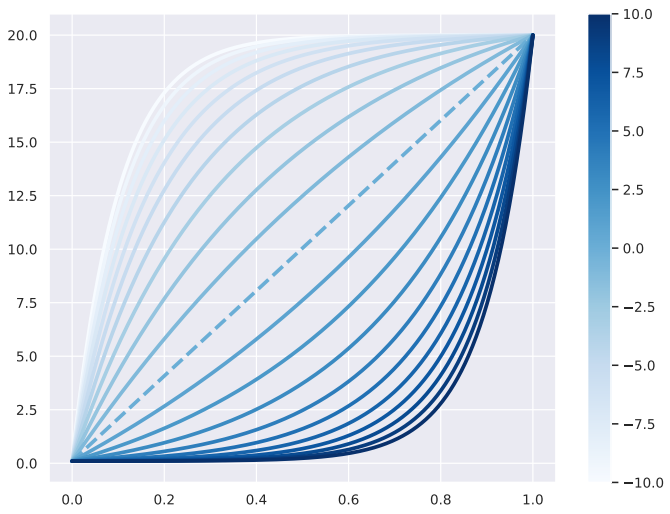


Figure 1: Evolution of noise schedules β_a w.r.t. time, for different values of parameter between -10 to 10. The original choice of noise schedule in the VPSDE case (Ho et al., 2020) is shown as a dashed line, corresponding to a linear noising function.

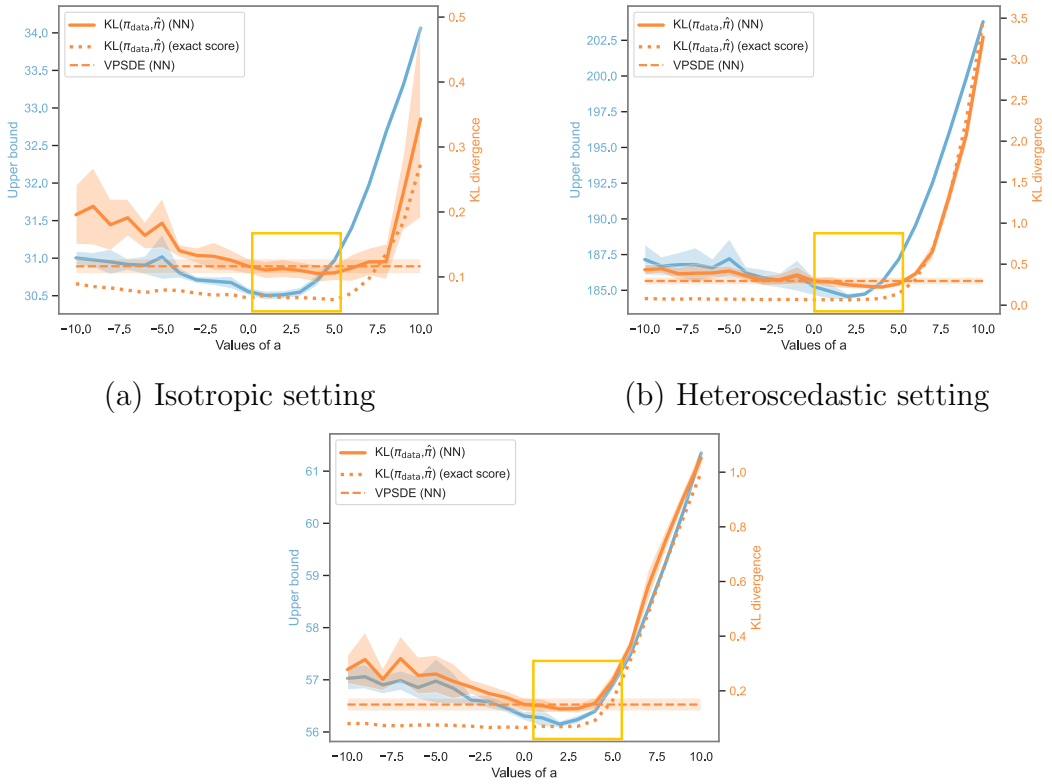
When the parameter a ranges from -10 to 10 with a unit step size, *i.e.*,$a \in \{-10, -9, \ldots, 9, 10\}$, the results are displayed in Figure 2. They highlight in all the scenarios that the noise schedule involved in the SGM impacts the value of $\mathrm{KL}(\pi_{\text{data}} \| \widehat{\pi}_N^{(\beta,\theta)})$, and thereby the quality of the learned distribution.

In the isotropic case (Figure 2 (a)), the behavior of the upper bound of Theorem 2.1 does not exactly match the one of $\mathrm{KL}(\pi_{\text{data}} \| \widehat{\pi}_N^{(\beta_a,\theta)})$ suggesting that the refinement relying on contraction arguments specific to the Gaussian setting (see Lemma 3.2) is indeed more informative in such a case.

When considering data distributions less naive such as $\pi_{\text{data}}^{(\text{heterosc})}$ and $\pi_{\text{data}}^{(\text{corr})}$ (Figure 2 (b) and (c)), the upper bound of Theorem 2.1 remains clearly relevant to assess the efficiency of the noise schedule used during training. Note that in all these experiments (Figure 2 and 3), the generic upper bound provided by Theorem 2.1 indicates a window of possible values for a improving over the classical linear noise schedule. This suggests that optimizing this upper bound with respect to the noise schedule through the parameter a could enable us to lower the discrepancy between π_{data} and the estimated one $\widehat{\pi}_N^{(\beta_a,\theta)}$, and thus improving the quality of the generated samples.

For all the settings (isotropic, heteroscedastic and correlated), we also verify these findings by making the dimension of the inputs vary in $\{5, 10, 25, 50\}$, and we compare the empirical KL obtained by (i) classical VPSDE (Song and Ermon, 2019), with a linear noise schedule (*i.e.*,$a = 0$), (ii) a time-inhomogeneous SGM involving a cosine schedule as in Nichol and Dhariwal (2021), and (iii) the one obtained by our time-inhomogeneous SGM. For the latter, we adopt the noise schedule to be $\beta^\star = \beta_{a^\star}$ where the parameter $a^\star \in \{-10, -9, \ldots, 9, 10\}$ corresponds to the minimizer of the estimated upper bound (9). In Figure 4, we observe that whatever the dimension is, $\widehat{\pi}_N^{\beta^\star\theta}$ always outperforms

the state-of-the-art diffusion models in terms of KL divergence, see Table 1 in the appendix for precise KL values. It appears to produce more stable generative models, as its variance in terms of KL over the different runs is clearly reduced compared to competitors when the dimension increases.



(a) Isotropic setting

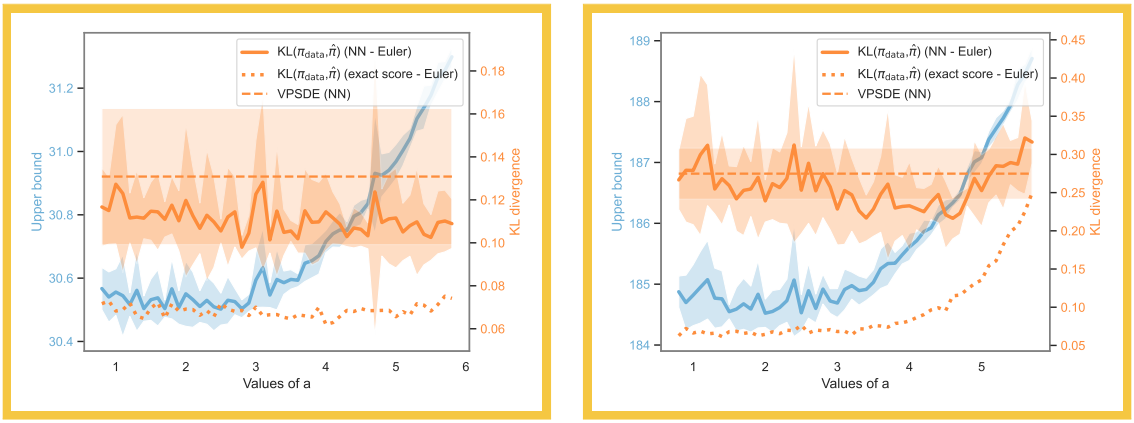
(b) Heteroscedastic setting

(c) Correlated setting

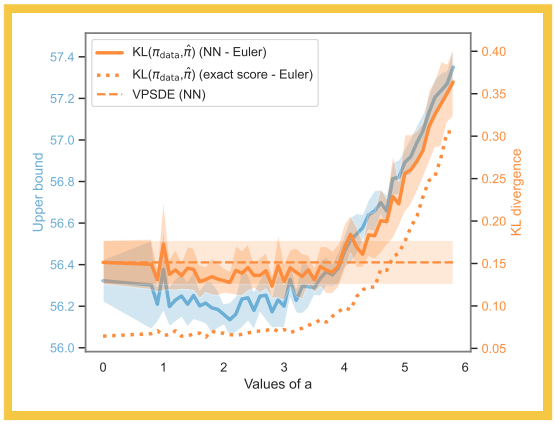
Figure 2: Comparison of the empirical KL divergence (mean value $\pm$ std over 10 runs) between π_{data} and $\widehat{\pi}_N^{(\beta,\theta)}$ (in orange) and the upper bound (9) (in blue) w.r.t. the parameter a used in the definition of the noise schedule β_a, for $d = 50$. We also represent the KL divergence obtained with the VPSDE model (dashed line) and the one obtained with our model (dotted line) when the score is not approximated but exactly evaluated. The data distribution π_{data} is chosen Gaussian, corresponding to (a) $\pi_{\text{data}}^{(\text{iso})}$, (b) $\pi_{\text{data}}^{(\text{heterosc})}$ and (c) $\pi_{\text{data}}^{(\text{corr})}$. The parameter a ranges from -10 to 10 by a unit step size.

4 Noise schedule optimization

Algorithm. Building on the previous numerical experiments, we propose to exploit the theoretical upper bound (9) to tune the choice of the noise schedule. To this aim, we design an iterative method to jointly optimize the weights θ of the NN score estimator



(a) Isotropic setting (b) Heteroscedastic setting

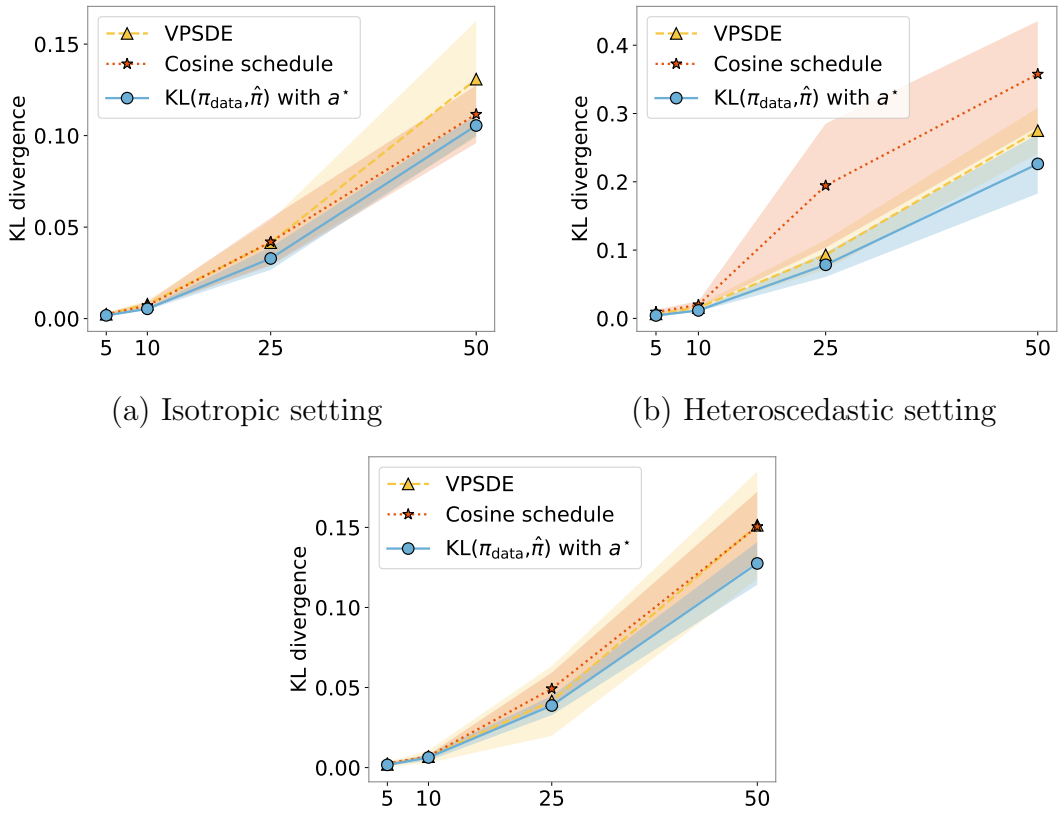


(c) Correlated setting

Figure 3: Zoom of Figure 2 by refining the grid for a: the parameter a ranges from 0.8 to 5.8 with a step size of 0.1.

and the noise schedule β, see Algorithm 1. The admissible functions β_a for the noise schedule are given in (10). For fair comparisons, we train both the VPSDE network and the adaptive scheduling network with 10000 samples over 100 epochs using the same learning rate. For the latter, the noise schedule, through the parameter a, is initialized at $a = 0$ and updated every 10 epochs.

Results. We assess the performance of Algorithm 1 by considering a Gaussian data distribution $\pi_{\text{data}}^{(\text{corr})}$. On Figure 5, along the epochs, we display the empirical KL divergences w.r.t. the generated distribution via Algorithm 1, vs. the regular VPSDE generator. From the very first epochs, Algorithm 1 produces better samples than the standard VPSDE model. As expected, the value of a selected by Algorithm 1 tends to be shifted to positive values with some stabilization around optimal values already observed in Figure 2.



(a) Isotropic setting

(b) Heteroscedastic setting

(c) Correlated setting

Figure 4: Comparison of the empirical KL divergence (mean value $\pm$ std over 10 runs) between π_{data} and the generative distribution $\hat{\pi}$ for different values of the dimension. The generative distributions considered are $\widehat{\pi}_N^{(\beta,\theta)}$ (blue plain) obtained by the time-inhomogeneous SGM for $\beta_{a^\star}$, the one obtained by a standard VPSDE model (yellow dashed), and the one obtained using a cosine schedule (orange dotted). The data distribution π_{data} is chosen Gaussian, corresponding to (a) $\pi_{\text{data}}^{(\text{iso})}$, (b) $\pi_{\text{data}}^{(\text{heterosc})}$ and (c) $\pi_{\text{data}}^{(\text{corr})}$.

Algorithm 1 Iterative optimization of the noise schedule and the score function

Input: N training samples, initial schedule β_a with $a = a^{(0)}$, initial parameter $\theta^{(0)}$.
Set $a^\star = a^{(0)}$
for $e = 0$ **to** number of epochs **do**
  Compute $\theta^{(e+1)}$ using score matching with noise schedule $\beta_{a^\star}$ and initial estimate $\theta^{(e)}$.
  if e mod10 $= 0$ **then**
    Update $a^\star \in \text{argmin}_a \mathcal{L}_{\text{sched}}(\theta^{(e+1)}, \beta_a)$.
  end if
end for

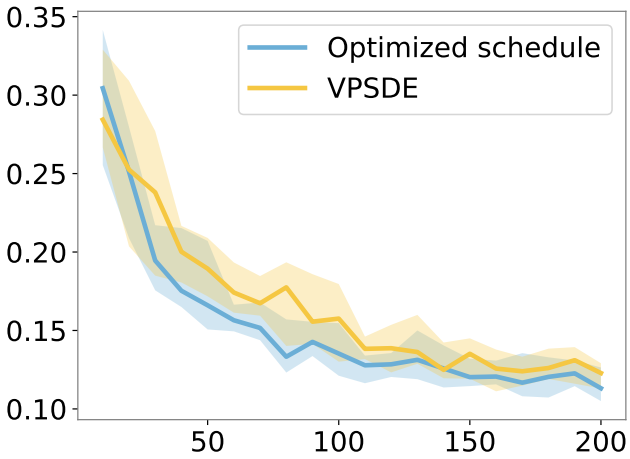


Figure 5: Empirical KL divergences (median and quartiles over 10 runs) between π_{data} and the distributions obtained by Algorithm 1 (blue) and the VPSDE model (yellow).

5 Discussion

In this paper, under mild assumptions, we have established an upper bound on the Kullback-Leibler divergence between the data distribution and that of score-based generative models with an explicit dependency on the noise schedule. We have also proposed a new procedure to jointly optimize the score network and the noise schedule. The tightness of the upper bound as long as the performance of the optimization procedure were illustrated empirically in Gaussian settings to allow fair comparisons with existing approaches and sampling methods based on exact score functions. Many extensions can be considered to exploit such upper bounds in order to improve the sampling performance of these models. Obtaining explicit and generic upper bounds for (sliced)-Wasserstein distances, when the data distribution is assumed to have only finite Fisher information would be useful as these metrics are highly valuable in practice. Extending our theoretical results to multi-dimensional noise schedules would also be of

particular interest to be able to deal with target distribution with complex covariance structures. Last but not least, establishing lower bounds either for Kullback-Leibler divergences or Wasserstein distances remains an exciting open problem, which would shed light on the performances and limitations of score-based generative models.

Acknowledgements

Antonio Ocello was supported by the European Union (ERC-2022-SYG-OCEAN-101071601). Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

Anonymous. Rethinking the noise schedule of diffusion-based generative models. In Submitted to The Twelfth International Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=ylHLVq0psd. under review.

Dominique Bakry, Ivan Gentil, Michel Ledoux, et al. Analysis and geometry of Markov diffusion operators, volume 103. Springer, 2014.

Paolo Baldi. Stochastic Calculus. Springer International Publishing AG, 1 edition, 2017. ISBN 978-3319622255.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Linear convergence bounds for diffusion models via stochastic localization, 2023.

Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative modeling with denoising auto-encoders and langevin sampling. arXiv preprint arXiv:2002.00107, 2020.

Djalil Chafai. Entropies, convexity, and functional inequalities. Kyoto Journal of Mathematics, 44(2), 2004. ISSN 2156-2261. doi: 10.1215/kjm/1250283556. URL http://arxiv.org/abs/math/0211103.

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, pages 4735–4763. PMLR, 2023a.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions, 2023b.

Ting Chen. On the importance of noise scheduling for diffusion models. arXiv preprint arXiv:2301.10972, 2023.

Jean-François Collet and Florent Malrieu. Logarithmic sobolev inequalities for inhomogeneous markov semigroups. European Series in Applied and Industrial Mathematics (ESAIM): Probability and Statistics, 12:492–504, 2008. ISSN 1292-8100. doi: 10.1051/ps:2007042.

Giovanni Conforti, Alain Durmus, and Marta Gentiloni Silveri. Score diffusion models without early stopping: finite fisher information is all you need, 2023.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. arXiv preprint arXiv:2208.05314, 2022.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. Advances in Neural Information Processing Systems, 34:17695–17709, 2021.

P. Del Moral, M. Ledoux, and L. Miclo. On contraction properties of markov kernels. Probability Theory and Related Fields, 126(3):395–420, 2003. ISSN 0178-8051. doi: 10.1007/s00440-003-0270-6.

Alain Durmus and Éric Moulines. Quantitative bounds of convergence for geometrically ergodic markov chain in the wasserstein distance with application to the metropolis adjusted langevin algorithm. Statistics and Computing, 25:5–19, 2015.

G. Franzese, S. Rossi, L. Yang, A. Finamore, D. Rossi, M. Filippone, and P. Michiardi. How much is enough? a study on diffusion times in score-based generative models. Entropy, 25:633, 2023. doi: 10.3390/e25040633.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. In Proceedings of International Conference on Learning Representations, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems, 2020.

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

Ioannis Karatzas and Steven Shreve. Brownian motion and stochastic calculus, volume 113. Springer Science & Business Media, 2012.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. Advances in Neural Information Processing Systems, 35:22870–22882, 2022.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In International Conference on Algorithmic Learning Theory, pages 946–985. PMLR, 2023.

Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. Neurocomputing, 479:47–59, 2022.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11461–11471, 2022.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In Marina Meila and Tong Zhang, editors, Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 8162–8171. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/nichol21a.html.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 1(2):3, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, Proceedings of the 32nd International Conference on Machine Learning, volume 37 of Proceedings of Machine Learning Research, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

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

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. International Conference on Learning Representations (ICLR), 2021a.

Yee Whye Song, Jascha Sohl-Dickstein, Durk P Kingma, Avinash Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In International Conference on Learning Representations, 2021b.

Michel Talagrand. Transportation cost for gaussian and other product measures. Geometric & Functional Analysis GAFA, 6(3):587–600, 1996.

Cédric Villani. Topics in optimal transportation, volume 58. American Mathematical Soc., 2021.

Pascal Vincent. A connection between score matching and denoising autoencoders. Neural Computation, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a_00142.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. ACM Computing Surveys, 56(4):1–39, 2023.

Notations and assumptions. Consider the following notations, used throughout the appendices. For all $d \geq 1$, $\mu \in \mathbb{R}^d$ and definite positive matrices $\Sigma \in \mathbb{R}^{d\times d}$, let $\varphi_{\mu,\Sigma}$ be the probability density function of a Gaussian random variable with mean μ and variance Σ. We also use the notation $\varphi_{\sigma^2} = \varphi_{0,\sigma^2 \mathrm{I}_d}$. For all twice-differentiable real-valued function f, let Δf be the Laplacian of f. For all matrix $A \in \mathbb{R}^{m\times n}$, $\|A\|_{\mathrm{Fr}}$ is the Frobenius norm of A, *i.e.*, $\|A\|_{\mathrm{Fr}} = (\sum_{i=1}^m \sum_{j=1}^n |A_{i,j}|^2)^{1/2}$.

Let π_0 be a probability density function with respect to the Lebesgue measure on $\mathbb{R}^d$ and $\alpha : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be two continuous and increasing functions. Consider the general forward process

$$\mathrm{d}\overrightarrow{X}_t = -\alpha(t)\overrightarrow{X}_t \mathrm{d}t + g(t)\mathrm{d}B_t, \quad \overrightarrow{X}_0 \sim \pi_0\,, \tag{11}$$

and introduce $\tilde{p}_t : x \mapsto p_t(x)/\varphi_{\sigma^2}(x)$, where p_t is the probability density function of $\overrightarrow{X}_t$. The backward process associated with (11) is referred to as $(\overleftarrow{X}_t)_{t\in[0,T]}$ and given by

$$\mathrm{d}\overleftarrow{X}_t = \left\{ \left(\bar{\alpha}(t) - \frac{\bar{g}^2(t)}{\sigma^2} \right) - \overleftarrow{X}_t + \bar{g}^2(t) \nabla \log \tilde{p}_{T-t}\left(\overleftarrow{X}_t\right) \right\} \mathrm{d}t + \bar{g}(t)\mathrm{d}\bar{B}_t \quad \overleftarrow{X}_0 \sim p_T\,, \tag{12}$$

with $\bar{\alpha}(t) := \alpha(T-t)$ and $\bar{g}(t) := g(T-t)$ and $\bar{B}$ a standard Brownian motion in $\mathbb{R}^d$. Moreover, consider

$$\sigma_t^2 := \exp\left(-2\int_0^t \alpha(s)\mathrm{d}s\right) \int_0^t g^2(s) \exp\left(2\int_0^s \alpha(u)du\right) \mathrm{d}s\,. \tag{13}$$

A Proof of Theorem 2.1

Lemma A.1. *For any noise schedule β,*

$$E_1(\beta) = \mathrm{KL}\left(p_T \| \varphi_{\sigma^2}\right) \leq \mathrm{KL}\left(\pi_{\mathrm{data}} \| \varphi_{\sigma^2}\right) \exp\left(-\frac{1}{\sigma^2}\int_0^T \beta(s)\mathrm{d}s\right)\,.$$

Proof. The proof follows the same lines as Franzese et al. (2023, Lemma 1). The Fokker-Planck equation associated with (1) is

$$\partial_t p_t(x) = \frac{\beta(t)}{2\sigma^2}\mathrm{div}\left(x p_t(x)\right) + \frac{\beta(t)}{2}\Delta p_t(x) = \frac{\beta(t)}{2}\mathrm{div}\left(\frac{1}{\sigma^2} x p_t(x) + \nabla p_t(x)\right)\,,$$

for $t \in [0,T], x \in \mathbb{R}^d$. Combing this with the derivation under the integral theorem, we

get

$$\begin{aligned}
\frac{\partial}{\partial t}\mathrm{KL}\left(p_t\|\varphi_{\sigma^2}\right) &= \frac{\partial}{\partial t}\int_{\mathbb{R}^d}\log\frac{p_t(x)}{\varphi_{\sigma^2}(x)}p_t(x)\mathrm{d}x \\
&= \int_{\mathbb{R}^d}\frac{\partial}{\partial t}p_t(x)\log\frac{p_t(x)}{\varphi_{\sigma^2}(x)}\mathrm{d}x + \frac{p_t(x)\partial_t p_t(x)}{p_t(x)}\mathrm{d}x \\
&= \int_{\mathbb{R}^d}\frac{\partial}{\partial t}p_t(x)\log\frac{p_t(x)}{\varphi_{\sigma^2}(x)}\mathrm{d}x + \int\frac{\partial}{\partial t}p_t(x)\mathrm{d}x \\
&= \int_{\mathbb{R}^d}\frac{\beta(t)}{2}\mathrm{div}\left(\frac{x}{\sigma^2}p_t(x)+\nabla p_t(x)\right)\log\frac{p_t(x)}{\varphi_{\sigma^2}(x)}\mathrm{d}x \\
&= \frac{\beta(t)}{2}\int_{\mathbb{R}^d}\mathrm{div}\left(-\nabla\log\varphi_{\sigma^2}(x)\;p_t(x)+\nabla p_t(x)\right)\log\frac{p_t(x)}{\varphi_{\sigma^2}(x)}\mathrm{d}x \\
&= -\frac{\beta(t)}{2}\int_{\mathbb{R}^d}\left(-\nabla\log\varphi_{\sigma^2}(x)\;p_t(x)+\nabla p_t(x)\right)^\top\nabla\log\frac{p_t(x)}{\varphi_{\sigma^2}(x)}\mathrm{d}x \\
&= -\frac{\beta(t)}{2}\int_{\mathbb{R}^d}p_t(x)\left(-\nabla\log\varphi_{\sigma^2}(x)+\nabla\log p_t(x)\right)^\top\nabla\log\frac{p_t(x)}{\varphi_{\sigma^2}(x)}s\mathrm{d}x \\
&= -\frac{\beta(t)}{2}\int_{\mathbb{R}^d}p_t(x)\left\|\nabla\log\frac{p_t(x)}{\varphi_{\sigma^2}(x)}\right\|^2\mathrm{d}x\,.
\end{aligned}$$

Using the Stam-Gross logarithmic Sobolev inequality given in Proposition B.3, we get

$$\frac{\partial}{\partial t}\mathrm{KL}\left(p_t\|\varphi_{\sigma^2}\right) \le -\frac{\beta(t)}{\sigma^2}\mathrm{KL}\left(p_t\|\varphi_{\sigma^2}\right)\,.$$

Applying Grönwall's inequality, we obtain

$$\mathrm{KL}\left(p_T\|\varphi_{\sigma^2}\right) \le \mathrm{KL}\left(p_0\|\varphi_{\sigma^2}\right)\exp\left\{-\frac{1}{\sigma^2}\int_0^T\beta(s)\mathrm{d}s\right\}\,,$$

which concludes the proof. □

Lemma A.2. *For all θ and all β,*

$$E_2(\theta,\beta) = \sum_{k=1}^{N}\mathbb{E}\left[\left\|\nabla\log\tilde{p}_{t_k}\left(\overrightarrow{X}_{t_k}\right)-\tilde{s}_\theta\left(t_k,\overrightarrow{X}_{t_k}\right)\right\|^2\right]\int_{t_k}^{t_{k+1}}\beta(t)\mathrm{d}t\,,$$

where $E_2(\theta,\beta)$ is defined by (5).

Proof. By definition of $E_2(\theta,\beta)$,

$$\begin{aligned}
E_2(\theta,\beta) &= \sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}\bar{\beta}(t)\mathbb{E}\left[\left\|\nabla\log\tilde{p}_{T-t_k}\left(\overleftarrow{X}_{t_k}\right)-\tilde{s}_\theta\left(T-t_k,\overleftarrow{X}_{t_k}\right)\right\|^2\right]\mathrm{d}t \\
&= \sum_{k=0}^{N-1}\mathbb{E}\left[\left\|\nabla\log\tilde{p}_{T-t_k}\left(\overleftarrow{X}_{t_k}\right)-\tilde{s}_\theta\left(T-t_k,\overleftarrow{X}_{t_k}\right)\right\|^2\right]\int_{t_k}^{t_{k+1}}\bar{\beta}(t)\mathrm{d}t \\
&= \sum_{k=1}^{N}\mathbb{E}\left[\left\|\nabla\log\tilde{p}_{t_k}\left(\overrightarrow{X}_{t_k}\right)-\tilde{s}_\theta\left(t_k,\overrightarrow{X}_{t_k}\right)\right\|^2\right]\int_{t_k}^{t_{k+1}}\beta(t)\mathrm{d}t\,,
\end{aligned}$$

where the last equality comes from the fact that the forward and backward processes have same marginals since $\overrightarrow{X}_T \sim p_T$. □

Lemma A.3. *Assume that H1 holds. For all $T, \sigma > 0$, θ and all β,*

$$E_3(\beta) \leq 2h\beta(T) \max\left\{\frac{h\beta(T)}{4\sigma^2}; 1\right\} \mathcal{I}(p_{\text{data}}|\pi_\infty)\,,$$

where $E_3(\beta)$ is defined by (6).

Proof. By Lemma B.4,

$$\mathrm{d}Y_t = \frac{\bar{\beta}(t)}{2\sigma^2} Y_t \mathrm{d}t + \sqrt{\bar{\beta}(t)} Z_t \mathrm{d}B_t\,.$$

By applying Itô's lemma to the function $x \mapsto \|x\|^2$, we obtain

$$\mathrm{d}\|Y_t\|^2 = \left(\frac{\bar{\beta}(t)}{\sigma^2}\|Y_t\|^2 + \bar{\beta}(t)\|Z_t\|_{\text{Fr}}^2\right)\mathrm{d}t + \sqrt{\bar{\beta}(t)} Y_t^\top Z_t \mathrm{d}B_t\,.$$

Fix $\delta > 0$. From Baldi (2017, Theorem 7.3), we have that $(\int_0^t g(s) Y_s^T Z_s \mathrm{d}B_s)_{t\in[0,T-\delta]}$ is a square integrable martingale if

$$\mathbb{E}\left[\int_0^{T-\delta} g^2(s) \left\|Y_s^T Z_s\right\|^2 \mathrm{d}s\right] < \infty\,.$$

From Cauchy-Schwarz inequality, we get that

$$\mathbb{E}\left[\left\|Y_s^T Z_s\right\|_2^2\right] \leq \mathbb{E}\left[\|Y_s\|_2^2 \|Z_s\|_{\text{Fr}}^2\right] \leq \mathbb{E}\left[\|Y_s\|_2^4\right]^{1/2} \mathbb{E}\left[\|Z_s\|_{\text{Fr}}^4\right]^{1/2}\,.$$

Applying Lemma B.5 and B.6, we get that both $\mathbb{E}[\|Y_s\|_2^4]$ and $\mathbb{E}[\|Z_s\|_2^4]$ are bounded by a quantity depending on σ_{T-t}^{-8}. As the term σ_{T-t}^{-8} is uniformly bounded in $[0, T-\delta]$ and by Fubini's theorem, we get

$$\mathbb{E}\left[\int_0^T g^2(s) \left\|Y_s^T Z_s\right\|^2 \mathrm{d}s\right] = \int_0^T g^2(s) \mathbb{E}\left[\left\|Y_s^T Z_s\right\|^2\right] \mathrm{d}s < \infty\,.$$

Therefore, $\left(\int_0^t g(s) Y_s^T Z_s \mathrm{d}B_s\right)_{t\in[0,T-\delta]}$ is a square integrable martingale. This means that, on one hand, we have

$$\mathbb{E}\left[\|Y_t\|^2\right] - \mathbb{E}\left[\|Y_{t_k}\|^2\right] = \mathbb{E}\left[\int_{t_k}^t \frac{\bar{\beta}(s)}{\sigma^2}\|Y_s\|^2 \mathrm{d}s + \int_{t_k}^t \bar{\beta}(s)\|Z_s\|_{\text{Fr}}^2 \mathrm{d}s\right]\,,$$

and, on the other hand,

$$\begin{aligned}
\mathbb{E}\left[\|Y_t - Y_{t_k}\|^2\right] &= \mathbb{E}\left[\left\|\int_{t_k}^t \frac{\bar{\beta}(s)}{2\sigma^2} Y_s \mathrm{d}s + \int_{t_k}^t \sqrt{\bar{\beta}(s)} Z_s \mathrm{d}B_s\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\int_{t_k}^t \frac{\bar{\beta}(s)}{2\sigma^2} Y_s \mathrm{d}s\right\|^2\right] + 2\mathbb{E}\left[\int_{t_k}^t \left\|\sqrt{\bar{\beta}(s)} Z_s \mathrm{d}B_s\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\frac{1}{2\sigma}\int_{t_k}^t \sqrt{\bar{\beta}(s)}\frac{\sqrt{\bar{\beta}(s)}}{\sigma} Y_s \mathrm{d}s\right\|^2\right] + 2\mathbb{E}\left[\int_{t_k}^t \left\|\sqrt{\bar{\beta}(s)} Z_s \mathrm{d}B_s\right\|^2\right] \\
&\leq \frac{1}{2\sigma^2}\int_{t_k}^{t_{k+1}} \bar{\beta}(s)\mathrm{d}s \mathbb{E}\left[\int_{t_k}^{t_{k+1}} \frac{\bar{\beta}(s)}{\sigma^2}\left\|Y_s\right\|^2 ds\right] \\
&\quad + 2\mathbb{E}\left[\int_{t_k}^{t_{k+1}} \bar{\beta}(s)\left\|Z_s\right\|_{\mathrm{Fr}}^2 \mathrm{d}s\right] \\
&\leq 2\max\left\{\frac{\int_{t_k}^{t_{k+1}} \bar{\beta}(s)\mathrm{d}s}{4\sigma^2}, 1\right\}\left(\mathbb{E}\left[\|Y_{t_{k+1}}\|^2\right] - \mathbb{E}\left[\|Y_{t_k}\|^2\right]\right) .
\end{aligned}$$

Without loss of generality, we have that $t_{N-1} = T_\delta$. Then, the discretization error can be bounded as follows

$$\begin{aligned}
&\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} \bar{\beta}(t)\mathbb{E}\left[\left\|\nabla \log \tilde{p}_{T-t}\left(\overleftarrow{X}_t\right) - \nabla \log \tilde{p}_{T-t_k}\left(\overleftarrow{X}_{t_k}\right)\right\|^2\right]\mathrm{d}t \\
&\quad = \sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} \bar{\beta}(t)\mathbb{E}\left[\|Y_t - Y_{t_k}\|^2\right]\mathrm{d}t \\
&\quad \leq 2\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} \bar{\beta}(t)\max\left\{\frac{\int_{t_k}^{t_{k+1}} \bar{\beta}(s)\mathrm{d}s}{4\sigma^2}, 1\right\}\left(\mathbb{E}\left[\|Y_{t_{k+1}}\|^2\right] - \mathbb{E}\left[\|Y_{t_k}\|^2\right]\right)\mathrm{d}t \\
&\quad \leq 2\sum_{k=0}^{N-1}\max\left\{\frac{\int_{t_k}^{t_{k+1}} \bar{\beta}(s)\mathrm{d}s}{4\sigma^2}, 1\right\}\left(\mathbb{E}\left[\|Y_{t_{k+1}}\|^2\right] - \mathbb{E}\left[\|Y_{t_k}\|^2\right]\right)\int_{t_k}^{t_{k+1}} \bar{\beta}(t)\mathrm{d}t \\
&\quad \leq 2\sum_{k=0}^{N-1}\max\left\{\frac{\left(\int_{t_k}^{t_{k+1}} \bar{\beta}(s)\mathrm{d}s\right)^2}{4\sigma^2}, \int_{t_k}^{t_{k+1}} \bar{\beta}(s)\mathrm{d}s\right\}\left(\mathbb{E}\left[\|Y_{t_{k+1}}\|^2\right] - \mathbb{E}\left[\|Y_{t_k}\|^2\right]\right) \\
&\quad \leq 2\max_{0\leq k\leq N-1}\left\{\max\left\{\frac{\left(\int_{t_k}^{t_{k+1}} \bar{\beta}(s)\mathrm{d}s\right)^2}{4\sigma^2}, \int_{t_k}^{t_{k+1}} \bar{\beta}(s)\mathrm{d}s\right\}\right\} \\
&\qquad \mathbb{E}\left[\left\|\nabla \log \tilde{p}_{T-t_{N-1}}\left(\overleftarrow{X}_{t_{N-1}}\right)\right\|^2\right] .
\end{aligned}$$

By H1, $t \mapsto \beta(t)$ is increasing, so that $t \mapsto \bar{\beta}(t)$ is decreasing. Therefore, defining

$\delta_k := t_{k+1} - t_k$,

$$
\begin{aligned}
&\sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \bar{\beta}(t) \mathbb{E}\left[\left\| \nabla \log \tilde{p}_{T-t}\left(\overleftarrow{X}_t\right) - \nabla \log \tilde{p}_{T-t_k}\left(\overleftarrow{X}_{t_k}\right) \right\|^2 \right] \mathrm{d}t \\
&\quad \leq 2 \max_{0 \leq k \leq N-1} \left\{ \max \left\{ \frac{\left(\delta_k \bar{\beta}(t_k)\right)^2}{4\sigma^2}, \delta_k \bar{\beta}(t_k) \right\} \right\} \mathbb{E}\left[\left\| \nabla \log \tilde{p}_{T-t_{N-1}}\left(\overleftarrow{X}_{t_{N-1}}\right) \right\|^2 \right] \\
&\quad \leq 2 \max_{0 \leq k \leq N-1} \left\{ \max \left\{ \frac{h^2 \bar{\beta}^2(t_k)}{4\sigma^2}, h\bar{\beta}(t_k) \right\} \right\} \mathcal{I}(p_{\mathrm{data}} Q_{T-\delta} | \pi_\infty) \\
&\quad \leq 2h\bar{\beta}(0) \max \left\{ \frac{h\bar{\beta}(0)}{4\sigma^2}, 1 \right\} \mathcal{I}(p_{\mathrm{data}} Q_{T-\delta} | \pi_\infty) \\
&\quad \leq 2h\beta(T) \max \left\{ \frac{h\beta(T)}{4\sigma^2}, 1 \right\} \mathcal{I}(p_{\mathrm{data}} Q_{T-\delta} | \pi_\infty) \,.
\end{aligned}
$$

Finally, following the steps of the proof of Conforti et al. (2023, Lemma 2), we can take the limit for δ that goes to zero, under Assumption H2, concluding the proof. □

B Technical results

Lemma B.1. *Assume that H1 and H2 hold. Let* $\left(\overrightarrow{X}_t\right)_{t \geq 0}$ *be a weak solution to the forward process (1). Then, the stationary distribution of* $\left(\overrightarrow{X}_t\right)_{t \geq 0}$ *is Gaussian with mean 0 and variance* $\sigma^2 \mathrm{I}_d$.

Proof. Consider the process

$$
\bar{X}_t = \exp\left(\frac{1}{2\sigma^2} \int_0^t \beta(s) \mathrm{d}s \right) \overrightarrow{X}_t \,.
$$

Itô's formula yields

$$
\overrightarrow{X}_t = \exp\left(-\frac{1}{2\sigma^2} \int_0^t \beta(s) \mathrm{d}s \right) \left(\overrightarrow{X}_0 + \int_0^t \sqrt{\beta(s)} \exp\left(\int_0^s \beta(u)/(2\sigma^2) \mathrm{d}u \right) \mathrm{d}B_s \right) \,. \tag{14}
$$

First, we have that

$$
\lim_{t \to \infty} \exp\left(-\frac{1}{2\sigma^2} \int_0^t \beta(s) \mathrm{d}s \right) \overrightarrow{X}_0 = 0 \,.
$$

Secondly, we have that the second term in the r.h.s. of (14), by property of the Wiener integral, is Gaussian with mean 0 and variance $\sigma_t^2 \mathrm{I}_d$, where

$$
\sigma_t^2 = \exp\left(-\frac{1}{\sigma^2} \int_0^t \beta(s) \mathrm{d}s \right) \int_0^t \beta(s) \mathrm{e}^{\int_0^s \beta(u)/\sigma^2 \mathrm{d}u} \mathrm{d}s = \sigma^2 \left(1 - \exp\left(-\frac{1}{\sigma^2} \int_0^t \beta(s) \mathrm{d}s \right) \right) \,.
$$

By H1, $\lim_{t \to \infty} \sigma_t^2 = \sigma^2$, which concludes the proof. □

Lemma B.2. *Let $T > 0$ and $b_1, b_2 : [0,T] \times C([0,T], \mathbb{R}^d) \to \mathbb{R}^d$ be measurable functions such that for $i \in \{1,2\}$,*

$$\mathrm{d}X_t^{(i)} = b_i\left(t, (X_s^{(i)})_{s\in[0,T]}\right)\mathrm{d}t + \sqrt{\beta(T-t)}\mathrm{d}B_t \tag{15}$$

admits a unique strong solution with $X_0^{(i)} \sim \pi_0^{(i)}$. Suppose that $(b_i(t, (X_s^{(i)})_{s\in[0,t]}))_{t\in[0,T]}$ is progressively measurable, with Markov semi-group $(P_t^{(i)})_{t\geq 0}$. In addition, assume that

$$\mathbb{E}\left[\exp\left\{\frac{1}{2}\int_0^T \frac{1}{\beta(T-s)}\left\|b_1\left(s, \left(X_u^{(1)}\right)_{u\in[0,s]}\right) - b_2\left(s, \left(X_u^{(1)}\right)_{u\in[0,s]}\right)\right\|^2 \mathrm{d}s\right\}\right] < \infty\,. \tag{16}$$

Then,

$$\begin{aligned}\mathrm{KL}\left(\pi_0^{(1)}P_T^{(1)} \| \pi_0^{(2)}P_T^{(2)}\right) &\leq \mathrm{KL}\left(\pi_0^{(1)}\|\pi_0^{(2)}\right)\\ &+ \frac{1}{2}\int_0^T \frac{1}{\beta(T-t)}\mathbb{E}\left[\left\|b_1\left(s, \left(X_u^{(1)}\right)_{u\in[0,s]}\right) - b_2\left(s, \left(X_u^{(1)}\right)_{u\in[0,s]}\right)\right\|^2\right]\mathrm{d}t\,.\end{aligned}$$

Proof. For $i \in \{1,2\}$, let $\mu^{(i)}$ be the distribution of $(X_t^{(i)})_{t\in[0,T]}$ on the Wiener space $(C([0,T];\mathbb{R}^d), \mathcal{B}(C([0,T];\mathbb{R}^d)))$ with $X_0^{(i)} \sim \pi_0^{(i)}$. Define $u(t,\omega)$ as

$$u(t,\omega) := \beta(T-t)^{-1/2}\left(b_1\left(t, \left(X_u^{(1)}\right)_{u\in[0,t]}\right) - b_2\left(t, \left(X_u^{(1)}\right)_{u\in[0,t]}\right)\right)\,,$$

and define $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P}(\omega) = M_T(\omega)$ where, for $t \in [0,T]$,

$$M_t(\omega) = \exp\left\{-\int_0^t u(s,\omega)^\top \mathrm{d}B_s - \frac{1}{2}\int_0^t \|u(s,\omega)\|^2 \mathrm{d}s\right\}\,.$$

From (16), the Novikov's condition is satisfied (Karatzas and Shreve, 2012, Chapter 3.5.D), thus the process M is a martingale. We can then define an equivalent probability measure, denoted by $\mathbb{Q}$, such that $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P} := M_T$. Applying Girsanov theorem, $\mathrm{d}\bar{B}_t = \mathrm{d}B_t + u(t, (X_s^{(1)})_{s\in[0,t]})\mathrm{d}t$ is a Brownian motion under the measure $\mathbb{Q}$. Therefore,

$$\begin{aligned}\mathrm{d}X_t^{(1)} &= b_1\left(t, \left(X_u^{(1)}\right)_{u\in[0,t]}\right)\mathrm{d}t + \sqrt{\beta(T-t)}\mathrm{d}B_t\\ &= b_2\left(t, \left(X_u^{(1)}\right)_{u\in[0,t]}\right)\mathrm{d}t + \sqrt{\beta(T-t)}\mathrm{d}\bar{B}_t\,.\end{aligned}$$

Using the uniqueness in law of (15), the law of $X^{(1)}$ under $\mathbb{P}$ is the same as the one of $\bar{X}^{(2)}$ under $\mathbb{Q}$, with $\bar{X}^{(2)}$ solution of (15) with $i = 2$ and $\bar{X}_0^{(2)} = \pi_0^{(1)}$. Denote by $\bar{\mu}^{(2)}$ the law of $\bar{X}^{(2)}$. Therefore,

$$\mu^{(1)}(A) = \mathbb{P}(X^{(1)} \in A) = \mathbb{Q}(\bar{X}^{(2)} \in A) = \int \mathbb{1}_A(\bar{X}^{(2)}(\omega))\mathbb{Q}(\mathrm{d}\omega) = \int_A M_t \mu^{(2)}(\mathrm{d}y)\,,$$

which implies that

$$\frac{\mathrm{d}\mu^{(2)}}{\mathrm{d}\bar{\mu}^{(1)}} = M_T \,.$$

Hence, we obtain that

$$\begin{aligned}
\mathrm{KL}\left(\mu^{(1)} \middle\| \mu^{(2)}\right) &= \mathrm{KL}\left(\pi_0^{(1)} \middle\| \pi_0^{(2)}\right) + \mathbb{E}\left[\log\left(\frac{\mathrm{d}\mu_\pi^{(1)}}{\mathrm{d}\bar{\mu}_\pi^{(2)}}\right)\right] \\
&= \mathrm{KL}\left(\pi_0^{(1)} \middle\| \pi_0^{(2)}\right) + \frac{1}{2}\mathbb{E}\left[\int_0^t u(s,\omega)^\top \mathrm{d}B_s + \frac{1}{2}\int_0^t \|u(s,\omega)\|^2 \mathrm{d}s\right] \\
&= \mathrm{KL}\left(\pi_0^{(1)} \middle\| \pi_0^{(2)}\right) \\
&\quad + \frac{1}{2}\int_0^T \frac{1}{\beta(T-t)}\mathbb{E}\left[\left\|b_1(t,(X_s^{(1)})_{s\in[0,t]}) - b_2(t,(X_s^{(1)})_{s\in[0,t]})\right\|^2\right]\mathrm{d}t \,,
\end{aligned}$$

which concludes the proof. □

Lemma B.3. *Let p be a probability density function on $\mathbb{R}^d$. For all $\sigma^2 > 0$,*

$$\mathrm{KL}\left(p(x) \| \varphi_{\sigma^2}(x)\right) = \int p(x) \log \frac{p(x)}{\varphi_{\sigma^2}(x)} \,\mathrm{d}x \leq \frac{\sigma^2}{2}\int \left\|\nabla \log \frac{p(x)}{\varphi_{\sigma^2}(x)}\right\|^2 p(x)\,\mathrm{d}x \,.$$

Proof. Define $f_{\sigma^2} : x \mapsto p(x)/\varphi_{\sigma^2}(x)$. Since $\nabla^2 \log \varphi_{\sigma^2}(x) = -\sigma^{-2}\mathrm{I}_d$, the Bakry-Emery criterion is satisfied with constant $\sigma^{2^{-1}}$, see Bakry et al. (2014); Villani (2021); Talagrand (1996). By the classical logarithmic Sobolev inequality,

$$\int f_{\sigma^2}(x) \log f_{\sigma^2}(x) \varphi_{\sigma^2}(x) \mathrm{d}x \leq \frac{\sigma^2}{2}\int \frac{\|\nabla f_{\sigma^2}(x)\|^2}{f_{\sigma^2}(x)} \varphi_{\sigma^2}(x) \mathrm{d}x \,,$$

which concludes the proof. □

Lemma B.4. *Define $Y_t := \nabla \log \tilde{p}_{T-t}(\overleftarrow{X}_t)$ and $Z_t := \nabla^2 \log \tilde{p}_{T-t}(\overleftarrow{X}_t)$, where $\{\overleftarrow{X}_t\}_{t\geq 0}$ is a weak solution to (11). Then,*

$$\mathrm{d}Y_t = \left(\frac{\bar{g}^2(t)}{\sigma^2} - \bar{\alpha}(t)\right) Y_t \mathrm{d}t - \frac{2}{\sigma^2}\left(\frac{\bar{g}^2(t)}{2\sigma^2} - \bar{\alpha}(t)\right)\overleftarrow{X}_t \mathrm{d}t + \bar{g}(t) Z_t \mathrm{d}\bar{B}_t \,. \tag{17}$$

Proof. The Fokker-Planck equation associated with the forward process (11) is

$$\partial_t p_t(x) = \alpha(t)\mathrm{div}\left(x p_t(x)\right) + \frac{g^2(t)}{2}\Delta p_t(x) \,, \tag{18}$$

for $x \in \mathbb{R}^d$. First, we prove that $\tilde{p}_t$ satisfies the following PDE

$$\begin{aligned}
\partial_t \log \tilde{p}_t(x) = d\left(\bar{\alpha}(t) - \frac{\bar{g}^2(t)}{2\sigma^2}\right) + \langle \nabla \log \tilde{p}_t(x), x\rangle \left(\bar{\alpha}(t) - \frac{\bar{g}^2(t)}{\sigma^2}\right) \\
+ \frac{\|x\|^2}{\sigma^2}\left(\frac{\bar{g}^2(t)}{2\sigma^2} - \bar{\alpha}(t)\right) + \frac{\bar{g}^2(t)}{2}\frac{\Delta \tilde{p}_t(x)}{\tilde{p}_t(x)} \,.
\end{aligned} \tag{19}$$

Using that $\nabla \log \varphi_{\sigma^2}(x) = -x/\sigma^2$, we have

$$\begin{aligned}\operatorname{div}(x p_t(x)) &= d\, p_t(x) + p_t(x)\, x^\top \nabla \log p_t(x) \\ &= \varphi_{\sigma^2}(x) \left(d\, \tilde{p}_t(x) + \tilde{p}_t(x) \nabla \log \tilde{p}_t(x)^\top x - \frac{\|x\|}{\sigma^2} \right) \\ &= \varphi_{\sigma^2}(x) \left(d\, \tilde{p}_t(x) + \nabla \tilde{p}_t(x)^\top x - \frac{\|x\|}{\sigma^2} \tilde{p}_t(x) \right) .\end{aligned}$$

Then, since $\Delta \varphi_{\sigma^2}(x) = (\varphi_{\sigma^2}(x)/\sigma^2)\,(\|x\|^2/\sigma^2 - d)$, we get

$$\begin{aligned}\Delta p_t(x) &= \tilde{p}_t(x) \Delta \varphi_{\sigma^2}(x) + 2 \nabla \tilde{p}_t(x)^\top \nabla \varphi_{\sigma^2}(x) + \varphi_{\sigma^2}(x) \Delta \tilde{p}_t(x) \\ &= \varphi_{\sigma^2}(x) \left(\frac{\tilde{p}_t(x)}{\sigma^2} \left(\frac{\|x\|^2}{\sigma^2} - d \right) - \frac{2}{\sigma^2} \nabla \tilde{p}_t(x)^\top x + \Delta \tilde{p}_t(x) \right) .\end{aligned}$$

Combining these results with (18), we obtain

$$\begin{aligned}\partial_t \tilde{p}_t(x) = d\, \tilde{p}_t(x) \left(\alpha(t) - \frac{g^2(t)}{2\sigma^2} \right) + \nabla \tilde{p}_t(x)^\top x \left(\alpha(t) - \frac{g^2(t)}{\sigma^2} \right) \\ + \tilde{p}_t(x) \frac{\|x\|^2}{\sigma^2} \left(\frac{g^2(t)}{2\sigma^2} - \alpha(t) \right) + \frac{g^2(t)}{2} \Delta \tilde{p}_t(x) .\end{aligned}$$

Hence, diving by $\tilde{p}_t$ yields (19).

The previous computation, together with the fact that $\Delta \tilde{p}_t / \tilde{p}_t = \Delta \log \tilde{p}_t + \|\nabla \log \tilde{p}_t\|^2$, yields that the function $\phi_t(x) := \log \tilde{p}_{T-t}(x)$ is a solution to the following PDE

$$\partial_t \phi_t(x) = -d \left(\bar{\alpha}(t) - \frac{\bar{g}^2(t)}{2\sigma^2} \right) - \nabla \phi_t(x)^\top x \left(\bar{\alpha}(t) - \frac{\bar{g}^2(t)}{\sigma^2} \right) \tag{20}$$

$$- \frac{\|x\|^2}{\sigma^2} \left(\frac{\bar{g}^2(t)}{2\sigma^2} - \bar{\alpha}(t) \right) - \frac{\bar{g}^2(t)}{2} \left(\Delta \phi_t(x) + \|\nabla \phi_t(x)\|^2 \right) . \tag{21}$$

Following the lines of the Conforti et al. (2023, Proposition 1), we get that, since α and g are continuous and non-increasing, the map p_t, solution to (18), belongs to $C^{1,2}((0,T] \times \mathbb{R}^d)$. Moreover, (12) can be rewritten as follows, with respect to $\tilde{p}_t$

$$\mathrm{d}\overleftarrow{X}_t = \left\{ \left(\bar{\alpha}(t) - \frac{\bar{g}^2(t)}{\sigma^2} \right) \overleftarrow{X}_t + \bar{g}^2(t) \nabla \log p_{T-t} \left(\overleftarrow{X}_t \right) \right\} \mathrm{d}t + \bar{g}(t) \mathrm{d}\bar{B}_t , \quad \overleftarrow{X}_0 \sim p_T ,$$

This means that, as $Y_t = \nabla \phi_t \left(\overleftarrow{X}_t \right)$, we can apply Itô's formula and obtain

$$\begin{aligned}\mathrm{d}Y_t = \Bigg[\partial_t \nabla \phi_t \left(\overleftarrow{X}_t \right) + \nabla^2 \phi_t \left(\overleftarrow{X}_t \right) \left(\left(\bar{\alpha}(t) - \frac{\bar{g}^2(t)}{\sigma^2} \right) \overleftarrow{X}_t + \bar{g}^2(t) \nabla \phi_t \left(\overleftarrow{X}_t \right) \right) \\ + \frac{\bar{g}^2(t)}{2} \Delta \nabla \phi_t \left(\overleftarrow{X}_t \right) \Bigg] \mathrm{d}t + \bar{g}(t) \nabla^2 \phi_t \left(\overleftarrow{X}_t \right) \mathrm{d}\bar{B}_t \\ = \Bigg[\nabla \left(\partial_t \phi_t \left(\overleftarrow{X}_t \right) + \frac{\bar{g}^2(t)}{2} \left(\Delta \phi_t \left(\overleftarrow{X}_t \right) + \left\| \nabla \phi_t \left(\overleftarrow{X}_t \right) \right\|^2 \right) \right) \\ + \left(\bar{\alpha}(t) - \frac{\bar{g}^2(t)}{\sigma^2} \right) \nabla^2 \phi_t \left(\overleftarrow{X}_t \right) \overleftarrow{X}_t \Bigg] \mathrm{d}t + \bar{g}(t) \nabla^2 \phi_t \left(\overleftarrow{X}_t \right) \mathrm{d}\bar{B}_t ,\end{aligned}$$

using that $2\nabla^2\phi_t(x)\nabla\phi_t(x) = \nabla\|\nabla\phi_t(x)\|^2$. Using (20), we get

$$\begin{aligned}\mathrm{d}Y_t = \Bigg[&-\left(\bar{\alpha}(t) - \frac{\bar{g}^2(t)}{\sigma^2}\right)\nabla\psi_t\left(\overleftarrow{X}_t\right) + \frac{2}{\sigma^2}\left(\bar{\alpha}(t) - \frac{\bar{g}^2(t)}{2\sigma^2}\right)\overleftarrow{X}_t \\ &+ \left(\bar{\alpha}(t) - \frac{\bar{g}^2(t)}{\sigma^2}\right)\nabla^2\phi_t\left(\overleftarrow{X}_t\right)\overleftarrow{X}_t\Bigg]\mathrm{d}t + \bar{g}(t)\nabla^2\phi_t\left(\overleftarrow{X}_t\right)\mathrm{d}\bar{B}_t\,,\end{aligned}$$

with $\psi_t(x) := \nabla\phi_t(x)^\top x$. With the identity $\nabla\left(x^\top\nabla\phi_t(x)\right) = \nabla\phi_t(x) + \nabla^2\phi_t(x)x$, we have

$$\begin{aligned}\mathrm{d}Y_t &= \left[\left(\frac{\bar{g}^2(t)}{\sigma^2} - \bar{\alpha}(t)\right)\nabla\phi_t\left(\overleftarrow{X}_t\right) + \frac{2}{\sigma^2}\left(\bar{\alpha}(t) - \frac{\bar{g}^2(t)}{2\sigma^2}\right)\overleftarrow{X}_t\right]\mathrm{d}t + \bar{g}(t)\nabla^2\phi_t\left(\overleftarrow{X}_t\right)\mathrm{d}\bar{B}_t \\ &= \left[\left(\frac{\bar{g}^2(t)}{\sigma^2} - \bar{\alpha}(t)\right)Y_t + \frac{2}{\sigma^2}\left(\bar{\alpha}(t) - \frac{\bar{g}^2(t)}{2\sigma^2}\right)\overleftarrow{X}_t\right]\mathrm{d}t + \bar{g}(t)Z_t\mathrm{d}\bar{B}_t\,,\end{aligned}$$

which concludes the proof. □

Lemma B.5. *Let $Y_t := \nabla\log\tilde{p}_{T-t}(\overleftarrow{X}_t)$, with $\overleftarrow{X}$ satisfying (12). There exists a constant $C > 0$ such that*

$$\mathbb{E}\left[\|Y_t\|^4\right] \le C\left(\sigma_{T-t}^{-4}\mathbb{E}\left[\|N\|^4\right] + \sigma^{-8}\mathbb{E}\left[\left\|\overrightarrow{X}_0\right\|^4\right]\right)\,, \tag{22}$$

with $N \sim \mathcal{N}(0, \mathrm{I}_d)$ and σ_t^2 as in (13).

Proof. The transition density $q_t(y, x)$ associated with the semi-group of the process (11) is given by

$$q_t(y,x) = \left(2\pi\sigma_t^2\right)^{-d/2}\exp\left(\frac{-\left\|x - y\exp\left(-\int_0^t\alpha(s)\mathrm{d}s\right)\right\|^2}{2\sigma_t^2}\right)\,.$$

Therefore, we have

$$\begin{aligned}\nabla\log p_{T-t}(x) &= \frac{1}{p_{T-t}(x)}\int p_0(y)\nabla_x q_{T-t}(y,x)\mathrm{d}y \\ &= \frac{1}{p_{T-t}(x)}\int p_0(y)\frac{y\exp\left(-\int_0^{T-t}\alpha(u)\mathrm{d}u\right) - x}{\sigma_{T-t}^2}q_{T-t}(y,x)\mathrm{d}y\,.\end{aligned}$$

This, together with the definition of $\tilde{p}$, yields

$$\nabla\log\tilde{p}_{T-t}\left(\overrightarrow{X}_{T-t}\right) = \sigma_{T-t}^{-2}\mathbb{E}\left[\overrightarrow{X}_0\mathrm{e}^{-\int_0^{T-t}\alpha(u)\mathrm{d}u} - \overrightarrow{X}_{T-t}\middle|\overrightarrow{X}_{T-t}\right] + \sigma^{-2}\overrightarrow{X}_{T-t}\,.$$

Using Jensen's inequality for conditional expectation, there exists a constant $C > 0$ (which may change from line to line) such that

$$\begin{aligned}
&\left\|\nabla \log \tilde{p}_{T-t}\left(\overrightarrow{X}_{T-t}\right)\right\|^4 \\
&\quad \leq C\left(\sigma_{T-t}^{-8}\left\|\mathbb{E}\left[\overrightarrow{X}_0 \mathrm{e}^{-\int_0^{T-t}\alpha(s)\mathrm{d}s} - \overrightarrow{X}_{T-t}\middle|\overrightarrow{X}_{T-t}\right]\right\|^4 + \sigma^{-8}\left\|\overrightarrow{X}_{T-t}\right\|^4\right) \\
&\quad \leq C\left(\sigma_{T-t}^{-8}\mathbb{E}\left[\left\|\overrightarrow{X}_0 \mathrm{e}^{-\int_0^{T-t}\alpha(s)\mathrm{d}s} - \overrightarrow{X}_{T-t}\right\|^4\middle|\overrightarrow{X}_{T-t}\right] + \sigma^{-8}\left\|\overrightarrow{X}_{T-t}\right\|^4\right).
\end{aligned}$$

Note that $\overrightarrow{X}_t$ has the same law as $\exp(-\int_0^t \alpha(s)\mathrm{d}s)\overrightarrow{X}_0 + \sigma_t N$, with $N \sim \mathcal{N}(0, \mathrm{I}_d)$. This means that we have that

$$\mathbb{E}\left[\left\|\nabla \log p_{T-t}\left(\overrightarrow{X}_{T-t}\right)\right\|^4\right] \leq C\sigma_{T-t}^{-4}\left(\mathbb{E}\left[\|N\|^4\right] + \mathbb{E}\left[\left\|\overrightarrow{X}_0\right\|^4\right]\right).$$

Finally,

$$\begin{aligned}
\mathbb{E}\left[\|Y_t\|^4\right] &= \mathbb{E}\left[\left\|\nabla \log \tilde{p}_{T-t}\left(\overleftarrow{X}_t\right)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\nabla \log \tilde{p}_{T-t}\left(\overrightarrow{X}_{T-t}\right)\right\|^4\right] \leq \sigma_{T-t}^{-4}\mathbb{E}\left[\|N\|^4\right] \\
&\leq C\left(\sigma_{T-t}^{-4}\mathbb{E}\left[\|N\|^4\right] + \sigma^{-8}\mathbb{E}\left[\left\|\overrightarrow{X}_0\right\|^4\right]\right),
\end{aligned}$$

which concludes the proof. □

Lemma B.6. *Let $Z_t := \nabla^2 \log \tilde{p}_{T-t}(\overleftarrow{X}_t)$, where $\{\overleftarrow{X}_t\}_{t\geq 0}$ is a weak solution to (12). There exists a constant $C > 0$ such that*

$$\mathbb{E}\left[\|Z_t\|^4\right] \leq C\left(\sigma_{T-t}^{-8} + \sigma^{-8}\right)\left(\mathbb{E}\left[\|N\|_2^8\right] + d^4\right), \tag{23}$$

with $N \sim \mathcal{N}(0, \mathrm{I}_d)$ and σ_t^2 as in (13).

Proof. Let $q_t(y, x)$ be the transition density associated to the semi-group of the process (11). Write

$$\begin{aligned}
&\nabla^2 \log p_{T-t}(x) \\
&\quad = \nabla\left(\frac{1}{p_{T-t}(x)}\int p_0(y)\frac{y\mathrm{e}^{-\int_0^{T-t}\alpha(s)\mathrm{d}s} - x}{\sigma_{T-t}^2}q_{T-t}(y, x)\mathrm{d}y\right) \\
&\quad = -\frac{\nabla p_{T-t}(x)}{p_{T-t}^2(x)}\left(\int p_0(y)\frac{y\mathrm{e}^{-\int_0^{T-t}\alpha(s)\mathrm{d}s} - x}{\sigma_{T-t}^2}q_{T-t}(y, x)\mathrm{d}y\right)^\top \\
&\qquad + \frac{1}{p_{T-t}(x)}\nabla\int p_0(y)\frac{y\mathrm{e}^{-\int_0^{T-t}\alpha(s)\mathrm{d}s} - x}{\sigma_{T-t}^2}q_{T-t}(y, x)\mathrm{d}y \\
&\quad = \frac{1}{\sigma_{T-t}^2\, p_{T-t}(x)}\left(-\int\left(\frac{\nabla p_{T-t}(x)}{p_{T-t}(x)}\right)\left(\frac{y\mathrm{e}^{-\int_0^{T-t}\alpha(s)\mathrm{d}s} - x}{\sigma_{T-t}^2}\right)^\top q_{T-t}(y, x)p_0(y)\mathrm{d}y\right.
\end{aligned}$$

$$- \mathrm{I}_d + \int \frac{1}{\sigma_{T-t}^2} \left(y e^{-\int_0^{T-t} \alpha(s) \mathrm{d}s} - x \right) \left(y e^{-\int_0^{T-t} \alpha(s) \mathrm{d}s} - x \right)^\top q_{T-t}(y, x) p_0(y) \mathrm{d}y \Bigg).$$

Therefore,

$$\begin{aligned}
&\nabla^2 \log \tilde{p}_{T-t} \left(\overrightarrow{X}_{T-t} \right) \\
&= -\frac{1}{\sigma_{T-t}^2} \left(\mathbb{E} \left[\left(\frac{\nabla p_{T-t} \left(\overrightarrow{X}_{T-t} \right)}{p_{T-t} \left(\overrightarrow{X}_{T-t} \right)} \right) \left(\overrightarrow{X}_0 e^{-\int_0^{T-t} \alpha(s) \mathrm{d}s} - \overrightarrow{X}_{T-t} \right)^\top \middle| \overrightarrow{X}_{T-t} \right] + \mathrm{I}_d \right) \\
&\quad + \sigma_{T-t}^{-4} \mathbb{E} \left[\left(\overrightarrow{X}_0 e^{-\int_0^{T-t} \alpha(s) \mathrm{d}s} - \overrightarrow{X}_{T-t} \right) \left(\overrightarrow{X}_0 e^{-\int_0^{T-t} \alpha(s) \mathrm{d}s} - \overrightarrow{X}_{T-t} \right)^\top \middle| \overrightarrow{X}_{T-t} \right] + \sigma^{-2} \mathrm{I}_d \\
&= -\sigma_{T-t}^{-4} \left(\mathbb{E} \left[\overrightarrow{X}_0 e^{-\int_0^{T-t} \alpha(s) \mathrm{d}s} - \overrightarrow{X}_{T-t} \middle| \overrightarrow{X}_{T-t} \right] \right) \left(\mathbb{E} \left[\overrightarrow{X}_0 e^{-\int_0^{T-t} \alpha(s) \mathrm{d}s} - \overrightarrow{X}_{T-t} \middle| \overrightarrow{X}_{T-t} \right] \right)^\top \\
&\quad + \left(\sigma^{-2} - \sigma_{T-t}^{-2} \right) \mathrm{I}_d \\
&\quad + \sigma_{T-t}^{-4} \mathbb{E} \left[\left(\overrightarrow{X}_0 e^{-\int_0^{T-t} \alpha(s) \mathrm{d}s} - \overrightarrow{X}_{T-t} \right) \left(\overrightarrow{X}_0 e^{-\int_0^{T-t} \alpha(s) \mathrm{d}s} - \overrightarrow{X}_{T-t} \right)^\top \middle| \overrightarrow{X}_{T-t} \right].
\end{aligned}$$

There exists a constant $C > 0$ (which may change from line to line) such that

$$\begin{aligned}
&\mathbb{E} \left[\left\| \nabla^2 \log p_{T-t} \left(\overrightarrow{X}_{T-t} \right) \right\|_{\mathrm{Fr}}^4 \right] \\
&\leq C \sigma_{T-t}^{-16} \mathbb{E} \left[\left\| \mathbb{E} \left[\overrightarrow{X}_0 e^{-\int_0^{T-t} \alpha(s) \mathrm{d}s} - \overrightarrow{X}_{T-t} \middle| \overrightarrow{X}_{T-t} \right] \mathbb{E} \left[\overrightarrow{X}_0 e^{-\int_0^{T-t} \alpha(s) \mathrm{d}s} - \overrightarrow{X}_{T-t} \middle| \overrightarrow{X}_{T-t} \right]^\top \right\|_{\mathrm{Fr}}^4 \right] \\
&\quad + C \left(\sigma_{T-t}^{-8} + \sigma^{-8} \right) d^4 \\
&\quad + C \sigma_{T-t}^{-16} \mathbb{E} \left[\left\| \mathbb{E} \left[\left(\overrightarrow{X}_0 e^{-\int_0^{T-t} \alpha(s) \mathrm{d}s} - \overrightarrow{X}_{T-t} \right) \left(\overrightarrow{X}_0 e^{-\int_0^{T-t} \alpha(s) \mathrm{d}s} - \overrightarrow{X}_{T-t} \right)^\top \middle| \overrightarrow{X}_{T-t} \right] \right\|_{\mathrm{Fr}}^4 \right].
\end{aligned}$$

As in the previous proof, we note that $\overrightarrow{X}_t$ has the same law as $\mathrm{e}^{-\int_0^t \alpha(s) \mathrm{d}s} \overrightarrow{X}_0 + \sigma_t N$, with $N \sim \mathcal{N}(0, \mathrm{I}_d)$. Therefore, using Jensen's inequality,

$$\begin{aligned}
&\mathbb{E} \left[\left\| \mathbb{E} \left[\overrightarrow{X}_0 e^{-\int_0^{T-t} \alpha(s) \mathrm{d}s} - \overrightarrow{X}_{T-t} \middle| \overrightarrow{X}_{T-t} \right] \mathbb{E} \left[\overrightarrow{X}_0 e^{-\int_0^{T-t} \alpha(s) \mathrm{d}s} - \overrightarrow{X}_{T-t} \middle| \overrightarrow{X}_{T-t} \right]^\top \right\|_{\mathrm{Fr}}^4 \right] \\
&\leq \mathbb{E} \left[\left\| \mathbb{E} \left[\overrightarrow{X}_0 e^{-\int_0^{T-t} \alpha(s) \mathrm{d}s} - \overrightarrow{X}_{T-t} \middle| \overrightarrow{X}_{T-t} \right] \right\|_2^4 \left\| \mathbb{E} \left[\overrightarrow{X}_0 e^{-\int_0^{T-t} \alpha(s) \mathrm{d}s} - \overrightarrow{X}_{T-t} \middle| \overrightarrow{X}_{T-t} \right] \right\|_2^4 \right] \\
&\leq \mathbb{E} \left[\mathbb{E} \left[\left\| \overrightarrow{X}_0 e^{-\int_0^{T-t} \alpha(s) \mathrm{d}s} - \overrightarrow{X}_{T-t} \right\|_2^8 \middle| \overrightarrow{X}_{T-t} \right] \right] \\
&\leq \sigma_t^8 \mathbb{E} \left[\| N \|_2^8 \right]
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbb{E}\left[\left(\overrightarrow{X}_0 e^{-\int_0^{T-t}\alpha(s)\mathrm{d}s}-\overrightarrow{X}_{T-t}\right)\left(\overrightarrow{X}_0 e^{-\int_0^{T-t}\alpha(s)\mathrm{d}s}-\overrightarrow{X}_{T-t}\right)^\top\middle|\overrightarrow{X}_{T-t}\right]\right\|_{\mathrm{Fr}}^4\right]\\
&\leq \mathbb{E}\left[\mathbb{E}\left[\left\|\left(\overrightarrow{X}_0 e^{-\int_0^{T-t}\alpha(s)\mathrm{d}s}-\overrightarrow{X}_{T-t}\right)\left(\overrightarrow{X}_0 e^{-\int_0^{T-t}\alpha(s)\mathrm{d}s}-\overrightarrow{X}_{T-t}\right)^\top\right\|_{\mathrm{Fr}}^4\middle|\overrightarrow{X}_{T-t}\right]\right]\\
&= \mathbb{E}\left[\left\|\left(\overrightarrow{X}_0 e^{-\int_0^{T-t}\alpha(s)\mathrm{d}s}-\overrightarrow{X}_{T-t}\right)\right\|_2^8\right]\\
&\leq \sigma_t^8\mathbb{E}\left[\|N\|_2^8\right]\,.
\end{aligned}$$

Hence, we can conclude that

$$\mathbb{E}\left[\|Z_t\|_{\mathrm{Fr}}^4\right]=\mathbb{E}\left[\left\|\nabla^2\log\tilde{p}_{T-t}\left(\overrightarrow{X}_{T-t}\right)\right\|_{\mathrm{Fr}}^4\right]\leq C\left(\sigma_{T-t}^{-8}+\sigma^{-8}\right)\left(\mathbb{E}\left[\|N\|_2^8\right]+d^4\right)\,.$$

□

C Proof of Proposition 3.1

To establish (7), let $x\in\mathbb{R}^d$ (resp. $y\in\mathbb{R}^d$) and denote by $\overleftarrow{X}^x$ (resp. $\overleftarrow{X}^y$) the solution of (3), with initial condition $\overleftarrow{X}_0^x=x$ (resp. $\overleftarrow{X}_0^x=y$). Applying Itô's formula and using Cauchy-Schwarz inequality, we get

$$\begin{aligned}
\left\|\overleftarrow{X}_t^x-\overleftarrow{X}_t^y\right\|^2&=\|x-y\|^2+2\int_0^t-\frac{\bar{\beta}(s)}{2\sigma^2}\left\|\overleftarrow{X}_s^x-\overleftarrow{X}_s^y\right\|^2\mathrm{d}s\\
&\quad+2\int_0^t\bar{\beta}(s)\left(\nabla\log\tilde{p}_{T-s}\left(\overleftarrow{X}_s^x\right)-\nabla\log\tilde{p}_{T-s}\left(\overleftarrow{X}_s^y\right)\right)^\top\left(\overleftarrow{X}_s^x-\overleftarrow{X}_s^y\right)\mathrm{d}s\\
&\leq\|x-y\|^2-\int_0^t\frac{\bar{\beta}(s)}{\sigma^2}\left(1-2L_s\sigma^2\right)\left\|\overleftarrow{X}_s^x-\overleftarrow{X}_s^y\right\|^2\mathrm{d}s\,.
\end{aligned}$$

Therefore, applying Grönwall's lemma, we obtain

$$\mathbb{E}\left[\sup_{t\in[0,T]}\left\|\overleftarrow{X}_t^x-\overleftarrow{X}_t^y\right\|^2\right]\leq\exp\left(-\int_0^T\frac{\bar{\beta}(t)}{\sigma^2}\left(1-2L_s\sigma^2\right)\mathrm{d}t\right)\|x-y\|^2\,.$$

From this, we can show contraction (7) in the 2–Wasserstein distance by taking the infimum over all couplings.

To establish (8) note that, under Assumption H4, we have

$$\begin{aligned}
\left\|\overleftarrow{X}_t^x-\overleftarrow{X}_t^y\right\|^2&=\|x-y\|^2+2\int_0^t-\frac{\bar{\beta}(s)}{2\sigma^2}\left\|\overleftarrow{X}_s^x-\overleftarrow{X}_s^y\right\|^2\mathrm{d}s\\
&\quad+2\int_0^t\bar{\beta}(s)\left(\nabla\log\tilde{p}_{T-s}\left(\overleftarrow{X}_s^x\right)-\nabla\log\tilde{p}_{T-s}\left(\overleftarrow{X}_s^y\right)\right)^\top\left(\overleftarrow{X}_s^x-\overleftarrow{X}_s^y\right)\mathrm{d}s\\
&\leq\|x-y\|^2-\int_0^t\frac{\bar{\beta}(s)}{\sigma^2}\left(1+2C_s\sigma^2\right)\left\|\overleftarrow{X}_s^x-\overleftarrow{X}_s^y\right\|^2\mathrm{d}s\,.
\end{aligned}$$

Therefore, applying Grönwall's lemma, we obtain

$$\mathbb{E}\left[\sup_{t\in[0,T]}\left\|\overleftarrow{X}_t^x-\overleftarrow{X}_t^y\right\|^2\right]\leq\exp\left(-\int_0^T\frac{\bar{\beta}(t)}{\sigma^2}\left(1+2C_s\sigma^2\right)\mathrm{d}t\right)\|x-y\|^2\,.$$

From this, we can show contraction (8) in the 2–Wasserstein distance by taking the infimum over all couplings.

Note that a similar assumption as Assumption H4 is used in De Bortoli et al. (2021, Proposition 10,11,12), in particular to bound the conditional moments of $\overleftarrow{X}_0$ given $\overleftarrow{X}_t$ for $t>0$. However, in this paper the authors also require additional assumptions, in particular that the score of π_{data} has a linear growth.

C.1 Gaussian case: proof of Lemma 3.2

In the case where π_{data} is the Gaussian probability density with mean μ_0 and variance Σ_0, we have

$$\nabla\log\tilde{p}_t(x)=-\left(m_t^2\Sigma_0+\sigma_t^2\mathrm{I}_d\right)^{-1}(x-m_t\mu_0)+\sigma^{-2}x\,,$$

with $m_t=\exp\left(-\int_0^t\beta(s)\mathrm{d}s/(2\sigma^2)\right)$ and $\sigma_t=\sigma^2(1-m_t^2)$. Let $\overrightarrow{\Sigma}_t=m_t^2\Sigma_0+\sigma_t^2\mathrm{I}_d$ be the covariance of the forward process $\overrightarrow{X}_t$ and $b_t=\overrightarrow{\Sigma}_t^{-1}m_t\mu_0$ so that

$$\nabla\log\tilde{p}_t(x)=A_tx+b_t\quad\text{with}\quad A_t=-\left(\overrightarrow{\Sigma}_t^{-1}-\sigma^{-2}\mathrm{I}_d\right)\,.\tag{24}$$

Note that, if we denote by $\lambda_0^1\leq\cdots\leq\lambda_0^d$ the eigenvalues of Σ_0, which are positive as Σ_0 is positive definite, we have that the eigenvalues of A_t are

$$\lambda_t^i:=-\frac{1}{m_t^2\lambda_0^i+\sigma_t^2}+\frac{1}{\sigma^2}\,.$$

It is straightforward to see that $\lambda_t^1\leq\cdots\leq\lambda_t^d$. Moreover, we always have that in this case

$$\begin{aligned}\left(\nabla\log\tilde{p}_t(x)-\nabla\log\tilde{p}_t(y)\right)^\top(x-y)&\leq\lambda_t^d\|x-y\|^2\,,\\ \|\nabla\log\tilde{p}_t(x)-\nabla\log\tilde{p}_t(y)\|&\leq\max\left\{\left|\lambda_t^1\right|,\left|\lambda_t^d\right|\right\}\|x-y\|\,,\end{aligned}$$

which entails that we can define

$$L_t:=\max\left\{\left|\lambda_t^1\right|,\left|\lambda_t^d\right|\right\}\,,\qquad C_t:=-\lambda_t^d\,,$$

and apply Proposition 3.1.

The condition $\lambda_t^d\leq 0$, or equivalently $\sigma^2\geq\lambda_{\max}(\Sigma_0)$, yields a contraction in 2–Wasserstein distance in the backward process as well in the forward process from Proposition 3.1. This shows that, in specific cases, with an appropriate calibration of

the variance of the stationary law with respect to the initial law, we have a contraction both in the forward and in the backward flows.

As a consequence, note that

$$\mathcal{W}_2\left(\pi_{\text{data}}, \varphi_{\sigma^2} Q_T\right)^2 \leq \mathcal{W}_2\left(p_T, \varphi_{\sigma^2}\right)^2 \exp\left(-\frac{1}{\sigma^2}\int_0^T \beta(t)(1+2C_t\sigma^2)\mathrm{d}t\right).$$

Using Talagrand's T_2 inequality for the Gaussian measure $\mathcal{W}_2\left(\mu, \varphi_{\sigma^2}\right)^2 \leq 2\sigma^2 \mathrm{KL}(\mu \| \varphi_{\sigma^2})$ and Lemma A.1 we get

$$\mathcal{W}_2\left(\pi_{\text{data}}, \varphi_{\sigma^2} Q_T\right)^2 \leq 2\sigma^2 \mathrm{KL}\left(\pi_{\text{data}} \| \varphi_{\sigma^2}\right) \exp\left(-\frac{2}{\sigma^2}\int_0^T \beta(t)(1+2C_t\sigma^2)\mathrm{d}t\right).$$

Proposition C.1. *Assume that π_{data} is a Gaussian distribution $\mathcal{N}(\mu_0, \Sigma_0)$ such that $\lambda_{\max}(\Sigma_0) \leq \sigma^2$ where $\lambda_{\max}(\Sigma_0)$ denotes the largest eigenvalue of Σ_0. Then,*

$$\mathrm{KL}\left(\pi_{\text{data}} \| \varphi_{\sigma^2} Q_T\right) \leq \mathrm{KL}\left(\pi_{\text{data}} \| \varphi_{\sigma^2}\right) \exp\left(-\frac{2}{\sigma^2}\int_0^T \beta(s)\mathrm{d}s\right).$$

Proof. In this Gaussian case, the backward process is linear (see (24)) and the associated infinitesimal generator writes, for $g \in \mathcal{C}^2$,

$$\overleftarrow{\mathcal{L}}_t g(x) = \nabla g(x)^\top \left(-\frac{\bar{\beta}(t)}{2\sigma^2} + \bar{\beta}(t)(\bar{A}_t x + \bar{b}_t)\right) + \frac{1}{2}\bar{\beta}(t)\Delta g(x),$$

where $\bar{A}_t = A_{T-t}$ and $\bar{b}_t = b_{T-t}$.

Our objective is to monitor the evolution of the KL divergence, $\mathrm{KL}(p_T Q_t \| \varphi_{\sigma^2} Q_t)$, for $t \in [0, T]$. We follow Del Moral et al. (2003, Section 6) (see also Collet and Malrieu, 2008). Let $q_t = p_T Q_t$ and $\phi_t = \varphi_{\sigma^2} Q_t$ two densities that satisfy the Fokker-Planck equation, involving the dual operator $\overleftarrow{\mathcal{L}}_t^*$ of the infinitesimal generator $\overleftarrow{\mathcal{L}}$

$$\begin{aligned}
\partial_t q_t &= \overleftarrow{\mathcal{L}}_t^* q_t, \qquad & q_0(x) &= p_T(x) \\
\partial_t \phi_t &= \overleftarrow{\mathcal{L}}_t^* \phi_t, \qquad & \phi_0(x) &= \varphi_{\sigma^2}(x).
\end{aligned}$$

Let $f_t = q_t/\phi_t$. By definition of $\mathrm{KL}(q_t \| \phi_t) = \int \ln\left(f_t(x)\right) q_t(x)\mathrm{d}x$ we have

$$\begin{aligned}
\partial_t \mathrm{KL}\left(q_t \| \phi_t\right) &= \int \ln\left(f_t(x)\right) \partial_t q_t(x)\mathrm{d}x + \int \partial_t \ln\left(f_t(x)\right) q_t(x)\mathrm{d}x \\
&= \int \ln\left(f_t(x)\right) \partial_t q_t(x)\mathrm{d}x - \int f_t(x)\partial_t \phi_t(x)\mathrm{d}x.
\end{aligned}$$

By employing the Fokker-Planck equation and the adjoint relation, which states that $\int f(x)\overleftarrow{\mathcal{L}}_t^*(g)(x)\mathrm{d}x = \int \overleftarrow{\mathcal{L}}_t f(x) g(x)\mathrm{d}x$ we obtain

$$\partial_t \mathrm{KL}\left(q_t \| \phi_t\right) = \int \overleftarrow{\mathcal{L}} \ln\left(f_t\right)(x) q_t(x)\mathrm{d}x - \int \overleftarrow{\mathcal{L}} f_t(x)\phi_t(x)\mathrm{d}x.$$

The infinitesimal generator $\overleftarrow{\mathcal{L}}$ satisfies the change of variables formula (see Bakry et al., 2014) so that

$$\overleftarrow{\mathcal{L}}_t(\ln(f)) = \frac{1}{f}\overleftarrow{\mathcal{L}}_t f - \frac{1}{2f^2}\overleftarrow{\Gamma}_t(f,f)\,,$$

where $\overleftarrow{\Gamma}_t$ is the "carré du champ" operator associated with $\overleftarrow{\mathcal{L}}_t$ defined by $\overleftarrow{\Gamma}_t(f,f)(x) = \beta(t)|\nabla f(x)|^2$. We then obtain

$$\begin{aligned}\partial_t \mathrm{KL}\left(q_t \| \phi_t\right) &= \int \overleftarrow{\mathcal{L}} f_t(x) \frac{q_t(x)}{f_t(x)} \mathrm{d}x - \int \frac{\beta(t)}{2} \frac{|\nabla f_t(x)|^2}{f_t^2(x)} q_t(x) \mathrm{d}x - \int \overleftarrow{\mathcal{L}} f_t(x) \phi_t(x) \mathrm{d}x \\ &= -\frac{\beta(t)}{2} \int \frac{|\nabla f_t(x)|^2}{f_t(x)} \phi_t(x) \mathrm{d}x\,. \end{aligned} \tag{25}$$

To obtain a control of the Kullback-Leibler divergence we need a logarithmic Sobolev inequality for the distribution of density $\phi_t = \varphi_{\sigma^2} Q_t$. In this Gaussian case, if $\overleftarrow{X}_0 \sim \mathcal{N}(0, \sigma^2)$ then for all $t \in [0,T]$ the law of $\overleftarrow{X}_t$ is a centered Gaussian with covariance matrix $\overleftarrow{\Sigma}_t$ given by

$$\overleftarrow{\Sigma}_t = \sigma^2 \exp\left(\int_0^t -\frac{\bar{\beta}(s)}{\sigma^2} + 2\bar{\beta}_s \bar{A}_s \mathrm{d}s\right) + \int_0^t \beta(s) \exp\left(\int_s^t -\frac{\bar{\beta}(u)}{\sigma^2} + 2\bar{\beta}(u)\bar{A}_u \mathrm{d}u\right) \mathrm{d}s\,,$$

where we use the matrix exponential. As mentioned before, if $\lambda_{\max}(\Sigma_0) \leq \sigma^2$, the eigenvalues of A_s, for $s \in [0,T]$, are negative. We can easily deduce that $\lambda_{\max}(\overleftarrow{\Sigma}_t) \leq \sigma^2$. We recall the logarithmic Sobolev inequality for a normal distribution (see Chafai, 2004, Corollary 9)

$$\mathrm{KL}(q_t \| \phi_t) \leq \frac{1}{2} \int \frac{1}{f_t(x)} \nabla f_t(x)^\top \overleftarrow{\Sigma}_t \nabla f_t(x) \phi_t(x) \mathrm{d}x \leq \frac{\lambda_{\max}(\overleftarrow{\Sigma}_t)}{2} \int \frac{|\nabla f_t(x)|^2}{f_t(x)} \phi_t(x) \mathrm{d}x\,.$$

Plugging this into (25) we get

$$\partial_t \mathrm{KL}(q_t \| \phi_t) \leq -\frac{\beta(t)}{\sigma^2} \mathrm{KL}(q_t \| \phi_t)\,.$$

Therefore, recalling that $q_0 = p_T$ and $\phi_0 = \varphi_{\sigma^2}$

$$\mathrm{KL}\left(q_T \| \varphi_{\sigma^2} Q_T\right) \leq \mathrm{KL}(p_T \| \varphi_{\sigma^2}) \exp\left(-\int_0^T \frac{\beta(s)}{\sigma^2} \mathrm{d}s\right)\,.$$

We conclude using Lemma A.1. □

D Additional experiments

D.1 Exact score and metrics in the Gaussian case

Lemma D.1. *Assume that the forward process defined in (1) :*

$$\mathrm{d}\overrightarrow{X}_t = -\frac{\beta(t)}{2\sigma^2}\overrightarrow{X}_t\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}B_t, \quad \overrightarrow{X}_0 \sim \pi_0\,,$$

is initialised with π_0 the Gaussian probability density function with mean μ_0 and variance Σ_0. Then, the score function of (1) is:

$$\nabla \log p_t(x) = -(m_t^2\Sigma_0 + \sigma_t^2 \mathrm{I}_d)^{-1}(x - m_t\mu_0)\,,$$

where p_t is the probability density function of $\overrightarrow{X}_t$, $m_t = \exp\{-\int_0^t \beta(s)\mathrm{d}s/(2\sigma^2)\}$ and $\sigma_t^2 = \sigma^2(1-m_t^2)$.

Proof. Note the following equality in law

$$\overrightarrow{X}_t = m_t X_0 + \sigma_t N\,,$$

for $N \sim \mathcal{N}(0, \mathrm{I}_d)$ independent of X_0. Therefore $\overrightarrow{X}_t \sim \mathcal{N}(m_t\mu_0, \overrightarrow{\Sigma}_t)$ with $\overrightarrow{\Sigma}_t = m_t^2\Sigma_0 + \sigma_t^2\mathrm{I}_d$ which concludes the proof.

□

Lemma D.2. *The relative Fisher information between $X_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$ and $X_\infty \sim \mathcal{N}(0, \sigma^2\mathrm{I}_d)$ is given by:*

$$\mathcal{I}\left(\varphi_{\mu_0,\Sigma_0} \| \varphi_{\sigma^2}\right) = \frac{1}{\sigma^4}\left(\mathrm{Tr}\left(\Sigma_0\right) + \|\mu_0\|^2\right) - \frac{2d}{\sigma^2} + \mathrm{Tr}\left(\Sigma_0^{-1}\right)\,.$$

Proof. The relative Fisher information between X_0 and X_∞ is given by

$$\mathcal{I}\left(\varphi_{\mu_0,\Sigma_0} \| \varphi_{\sigma^2}\right) = \int \left\| \nabla \log\left(\frac{\varphi_{\mu_0,\Sigma_0}(x)}{\varphi_{\sigma^2}(x)}\right)\right\|^2 \varphi_{\mu_0,\Sigma_0}(x)\mathrm{d}x\,.$$

Write

$$\nabla \log \frac{\varphi_{\mu_0,\Sigma_0}(x)}{\varphi_{\sigma^2}(x)} = \frac{x}{\sigma^2} - \Sigma_0^{-1}(x-\mu_0)\,,$$

so that,

$$\begin{aligned}
\left\| \nabla \log \frac{\varphi_{\mu_0,\Sigma_0}(x)}{\varphi_{\sigma^2}(x)}\right\|^2 &= \left\| \frac{x}{\sigma^2} - \Sigma_0^{-1}(x-\mu_0)\right\|^2 \\
&= \left(\frac{x}{\sigma^2} - \Sigma_0^{-1}(x-\mu_0)\right)^\top \left(\frac{x}{\sigma^2} - \Sigma_0^{-1}(x-\mu_0)\right) \\
&= \frac{\|x\|^2}{\sigma^4} - \frac{2}{\sigma^2}x^\top\Sigma_0^{-1}(x-\mu_0) + (x-\mu_0)^\top\Sigma_0^{-2}(x-\mu_0)\,.
\end{aligned}$$

First,

$$\mathbb{E}\left[\frac{\|X_0\|^2}{\sigma^4}\right] = \frac{1}{\sigma^4}\left(\mathrm{Tr}\left(\Sigma_0\right) + \|\mu_0\|^2\right) .$$

Then,

$$\mathbb{E}\left[\frac{2}{\sigma^2} X_0^T \Sigma_0^{-1}(X_0 - \mu_0)\right] = \frac{2}{\sigma^2}\left(\mathrm{Tr}\left(\Sigma_0^{-1}\mathbb{E}\left[X_0 X_0^\top\right]\right) - \mu_0^\top \Sigma_0^{-1}\mu_0\right) .$$

Using that $\mathbb{E}\left[X_0 X_0^\top\right] = \Sigma_0 + \mu_0\mu_0^\top$ yields

$$\begin{aligned}
\mathbb{E}\left[\frac{2}{\sigma^2} X_0^\top \Sigma_0^{-1}(X_0 - \mu_0)\right] &= \frac{2}{\sigma^2}\left(\mathrm{Tr}\left(\Sigma_0^{-1}\left(\Sigma_0 + \mu_0\mu_0^\top\right)\right) - \mu_0^\top\Sigma_0^{-1}\mu_0\right) \\
&= \frac{2}{\sigma^2}\left(d + \mathrm{Tr}\left(\Sigma_0^{-1}\mu_0\mu_0^\top\right) - \mu_0^\top\Sigma_0^{-1}\mu_0\right) \\
&= \frac{2d}{\sigma^2} .
\end{aligned}$$

Finally,

$$\begin{aligned}
\mathbb{E}\left[(X_0 - \mu_0)^\top \Sigma_0^{-2}(X_0 - \mu_0)\right] &= \mathbb{E}\left[\mathrm{Tr}\left((X_0 - \mu_0)^\top \Sigma_0^{-2}(X_0 - \mu_0)\right)\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\left(\Sigma_0^{-2}(X_0 - \mu_0)(X_0 - \mu_0)^\top\right)\right] \\
&= \mathrm{Tr}\left(\Sigma_0^{-2}\mathbb{E}\left[(X_0 - \mu_0)(X_0 - \mu_0)^\top\right]\right) \\
&= \mathrm{Tr}\left(\Sigma_0^{-2}\Sigma_0\right) \\
&= \mathrm{Tr}\left(\Sigma_0^{-1}\right) ,
\end{aligned}$$

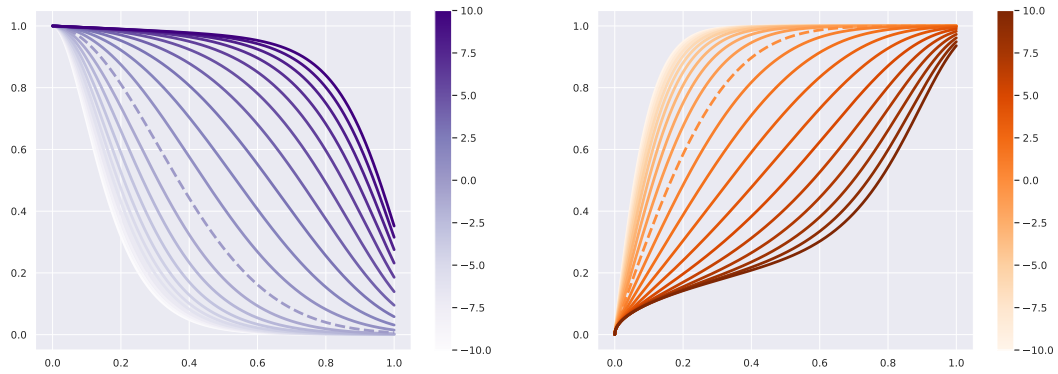
which concludes the proof. □

D.2 Stochastic differential equation exact simulation

In certain cases, exact simulation of stochastic differential equations is possible. In particular, due to the linear nature of the drift the forward process (1) can be simulated exactly. Indeed, the marginal distribution of (1) at time t writes as

$$\overrightarrow{X}_t = m_t X_0 + \sigma_t Z ,$$

with $Z \sim \mathcal{N}(0, \mathrm{I}_d)$ independent of X_0, $X_0 \sim \pi_0$, $m_t = \exp\{-\int_0^t \beta(s)\mathrm{d}s/(2\sigma^2)\}$ and $\sigma_t^2 = \sigma^2(1 - \exp\{-\int_0^t \beta(s)/\sigma^2\mathrm{d}s\})$. Therefore, sampling from the forward process only necessitates access to samples from π_0 and $\mathcal{N}(0, \mathrm{I}_d)$.



m_t σ_t

Figure 6: Evolution of m_t and σ_t over time, depending on the noise schedule β_a used (see Section 3.2 for the definition of β_a). The values for a range from -10 to 10. The dashed line corresponds to the case of a diffusion with a linear schedule, as proposed in VPSDE models (Song et al., 2021b).

D.3 Discretization details of the diffusion SDE

In contrast to the forward process, described in Equation (1), which is simulated exactly, the backward process needs to be discretized. Recall that the backward process of (1) is given by:

$$\mathrm{d}\overleftarrow{X}_t = -\frac{\bar{\beta}(t)}{2\sigma^2}\overleftarrow{X}_t + \bar{\beta}(t)\nabla \log p_{T-t}(\overleftarrow{X}_t)\mathrm{d}t + \sqrt{\bar{\beta}(t)}\mathrm{d}B_t, \quad \overleftarrow{X}_0 \sim \pi_\infty\,.$$

Our numerical study explores two distinct discretization approaches for the backward process, as detailed below. Consider time intervals $0 \le t_k \le t \le t_{k+1} \le T$, with $t_k = \sum_{\ell=1}^{k} \gamma_\ell$ and $T = \sum_{k=1}^{N} \gamma_k$.

- The Euler-Maruyama discretization is defined recursively for $t \in [t_k, t_{k+1}]$ by

$$\mathrm{d}\overleftarrow{X}^{EM}_t = -\frac{\bar{\beta}(t_k)}{2\sigma^2}\overleftarrow{X}^{EM}_{t_k} + \bar{\beta}(t_k)\nabla \log p_{T-t_k}(\overleftarrow{X}^{EM}_{t_k})\mathrm{d}t + \sqrt{\bar{\beta}(t_k)}\mathrm{d}B_t, \quad \overleftarrow{X}^{EM}_0 \sim \pi_\infty\,.$$

- The Exponential Integrator discretization is defined recursively for $t \in [t_k, t_{k+1}]$ by

$$\mathrm{d}\overleftarrow{X}^{EI}_t = \bar{\beta}(t)\left(-\frac{1}{2\sigma^2}\overleftarrow{X}^{EI}_t + \nabla \log p_{T-t_k}\left(T-t_k, \overleftarrow{X}^{EI}_{t_k}\right)\right)\mathrm{d}t + \sqrt{\bar{\beta}(t)}\mathrm{d}B_t\,, \quad (26)$$

  for $\overleftarrow{X}^{EI}_0 \sim \pi_\infty$.

D.4 Implementation of the score approximation

Although the score function is explicit when π_{data} is Gaussian, see Lemma D.1, we implement SGMs as done in applications, *i.e.*, we train a deep neural network to witness

the effect of the noising function on the approximation error. We train a neural network architecture $s_\theta(t,x) \in [0,T] \times \mathbb{R}^d \mapsto \mathbb{R}^d$ using the actual score function as a target:

$$\begin{aligned}\mathcal{L}_{\text{explicit}}(\theta) &= \mathbb{E}\left[\left\| s_\theta\left(\tau, \overrightarrow{X}_\tau\right) - \nabla \log p_\tau\left(\overrightarrow{X}_\tau\right)\right\|^2\right] \\ &= \mathbb{E}\left[\left\| s_\theta\left(\tau, \overrightarrow{X}_\tau\right) - (m_\tau^2 \Sigma_0 + \sigma_\tau^2 \mathrm{I}_d)^{-1}(\overrightarrow{X}_\tau + m_\tau \mu_0)\right\|^2\right],\end{aligned}$$

where $t \to m_t$ and $t \to \sigma_t$ are defined in Lemma D.1 and $\tau \sim \mathcal{U}(0,T)$ is independent of $\overrightarrow{X}$. The neural network architecture chosen for this task is described in Figure 7. The width of each dense layer `mid_features` is set to 256 throughout the experiments.

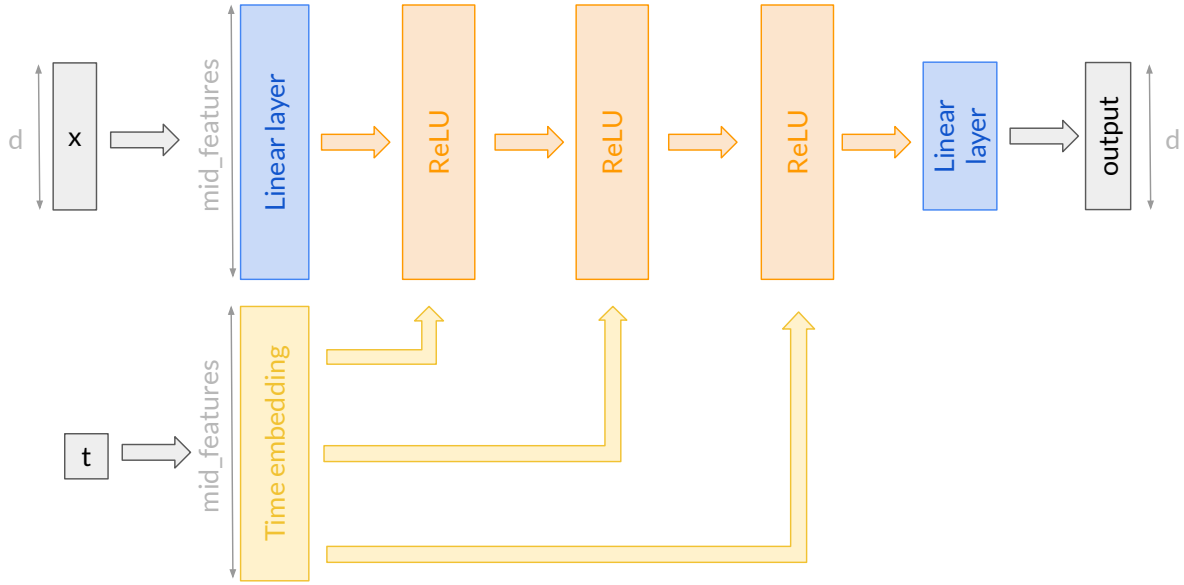


Figure 7: Neural network architecture. The input layer is composed of a vector x in dimension d and the time t. Both are respectively embedded using a linear transformation or a sine/cosine transformation Nichol and Dhariwal (2021) of width `mid_features`. Then, 3 dense layers of constant width `mid_features` followed by ReLu activations and skip connections regarding the time embedding. The output layer is linear resulting in a vector of dimension d.

D.5 Additional numerical results

We investigate the expressivity of the upperbound from Theorem (2.1) in the Gaussian setting. We use as a training sample 10^4 samples with distribution $\mathcal{N}(\mathbf{1}_d, \Sigma)$ for $d \in \{5, 10, 25, 50\}$ with different choices of covariance structure.

1. (Isotropic) $\Sigma^{\text{iso}} = 0.5\mathrm{I}_d$.

2. (Heteroscedastic) $\Sigma^{\text{heterosc}} \in \mathbb{R}^{d\times d}$ is a diagonal matrix such that $\Sigma^{\text{heterosc}}_{jj} = 10$ for $1 \leq j \leq d$, and $\Sigma^{\text{heterosc}}_{jj} = 0.1$ otherwise.

3. (Correlated) $\Sigma^{\text{corr}} \in \mathbb{R}^{d\times d}$ is a full matrix whose diagonal entries are equal to one and the off-diagonal terms are given by $\Sigma^{\text{corr}}_{jj'} = 1/\sqrt{|j-j'|}$ for $1 \leq j \neq j' \leq d$.

When first introduced, Song et al. (2021b) originally proposed a linear schedule $t \to \beta^{\text{lin}}(t)$, for $t \in [0,1]$ setting $\beta^{\text{lin}}(0) = 0.1$ and $\beta^{\text{lin}}(1) = 20$. We study a parametric family of schedules of the form $\beta_a(t) \propto (\mathrm{e}^{at} - 1)/(\mathrm{e}^{aT} - 1)$ sharing the same starting and ending values as the linear schedule (see Figure 1). Our goal is to assess the impact of the noising function on both the data distribution generation and the upper bound.

For the upper bound, we leverage the Gaussianity of the target distribution to compute explicitly both the relative entropy and the Fisher information in the upper bound. On the one hand, the relative entropy in $\mathcal{E}_1$, $\mathrm{KL}\left(\pi_{\text{data}} \| \pi_\infty\right)$ is computed using the analytical formula for KL-divergence between two random Gaussian variable. On the other, the relative Fisher information in $\mathcal{E}_3$, $\mathcal{I}(\pi_{\text{data}}|\pi_\infty)$, is computed using Lemma (D.2). Moreover, as the noising function and its primitive are analytically known, every occurrences of either of them are explicitly computed. Finally, it remains to estimate the expectations in $\mathcal{E}_2(\theta, \beta)$. This is done via Monte Carlo estimation on 500 samples from the forward process for every discretization step.

For the data generation, we either use the exact score function from Lemma D.1 or use the deep neural network architecture discussed in section D.4 to generate 10 000 samples. The batch size is set to 64 and Adam otpimizer was used for the learning phase. In Figures 2 and 3 we represent on the same graph, for different values of a:

- in blue the upper bound from Theorem 2.1.
- in orange (plain line) the KL divergence between the target data π_{data} and the empirical mean and covariance of the data generated using the neural network architecture described above to approximate the score function.
- in orange (dotted line) the KL divergence between the target data π_{data} and the empirical mean and covariance of the data generated using the true score function.
- in orange (dashed line) the KL divergence between the target data π_{data} and the empirical mean and covariance of the data generated by the VPSDE presented in Song and Ermon (2019).

Due to the stochastic nature of our experiments, each was repeated ten times to improve statistical reliability. In our graphs, we have plotted the mean value of these results and we employed a 'fill-between' plot to illustrate the range between the mean plus or minus the standard deviation over the ten runs.

	Dimension	5	10	25	50
Isotropic	Upper bound min $a^\star$	1.3	1.6	2.2	2.5
	Generation value in $a^\star$	0.00177 ± 0.00067	0.00535 ± 0.00067	0.03286 ± 0.00624	0.10551 ± 0.00563
	VPSDE	0.00239 ± 0.00061	0.00785 ± 0.00206	0.04150 ± 0.01206	0.13084 ± 0.03173
	Cosine schedule	0.00226 ± 0.00099	0.00709 ± 0.00191	0.04193 ± 0.01305	0.11165 ± 0.01577
	% gain (vs VPSDE)	+25.95 %	+31.83 %	+20.81 %	+19.36 %
	% gain (vs Cosine)	+21.68 %	+24.54 %	+21.63 %	+5.50 %
Heterosc.	Upper bound min $a^\star$	1.5	1.2	1.3	2.0
	Generation value in $a^\star$	0.00424 ± 0.00163	0.01162 ± 0.00154	0.07845 ± 0.01778	0.22621 ± 0.04355
	VPSDE	0.00660 ± 0.00164	0.01577 ± 0.00394	0.09295 ± 0.02180	0.27483 ± 0.03355
	Cosine schedule	0.00931 ± 0.00331	0.01983 ± 0.00539	0.19442 ± 0.09075	0.35763 ± 0.07769
	% gain (vs VPSDE)	+35.76 %	+26.32 %	+15.60 %	+17.69 %
	% gain (vs Cosine)	+54.46%	+41.40 %	+59.65 %	+36.75 %
Correlated	Upper bound min $a^\star$	1.5	1.2	2.1	2.1
	Generation value in $a^\star$	0.00171 ± 0.00056	0.00632 ± 0.00201	0.03877 ± 0.00602	0.12750 ± 0.013361
	VPSDE	0.00198 ± 0.00069	0.00684 ± 0.00233	0.04163 ± 0.01055	0.15132 ± 0.02597
	Cosine schedule	0.00261 ± 0.00079	0.00701 ± 0.00109	0.04926 ± 0.01002	0.15051 ± 0.02178
	% gain (vs VPSDE)	+13.63 %	+7.60 %	+6.87 %	+15.74 %
	% gain (vs Cosine)	+34.48 %	+9.84 %	+21.29 %	+15.29 %
Parameters	Learning rate	1e-4	1e-4	1e-3	1e-3
	Epochs	20	30	50	100

Table 1: Comparison of the KL divergence between the target value and the generated value at $a^\star$ (the minimum value of the upper bound (9)) with the KL divergence between the generated value by VPSDE and the target distribution. The target distributions are chosen to be Gaussian with different covariance structures: isotropic ($\pi_{\text{data}}^{(\text{iso})}$), heteroscedastic ($\pi_{\text{data}}^{(\text{heterosc})}$) and correlated ($\pi_{\text{data}}^{(\text{corr})}$).

D.6 Analysis of the discretization error

While Theorem 2.1 does not explicit exhibit dependency on the choice of the noising function β_a, our numerical experiments suggest otherwise. Indeed, our derivation of $\mathcal{E}_3(\beta_a)$ depends on β_a only through its terminal value $\beta_a(T)$, which we set in all experiments to $\beta_a(T) = 20$. Figure 8 shows the KL divergence between π_{data} and sample from the data generated using the exact score function for different numbers discretization steps. It clearly appears that, the KL divergence is non constant with respect to the noising function tested.

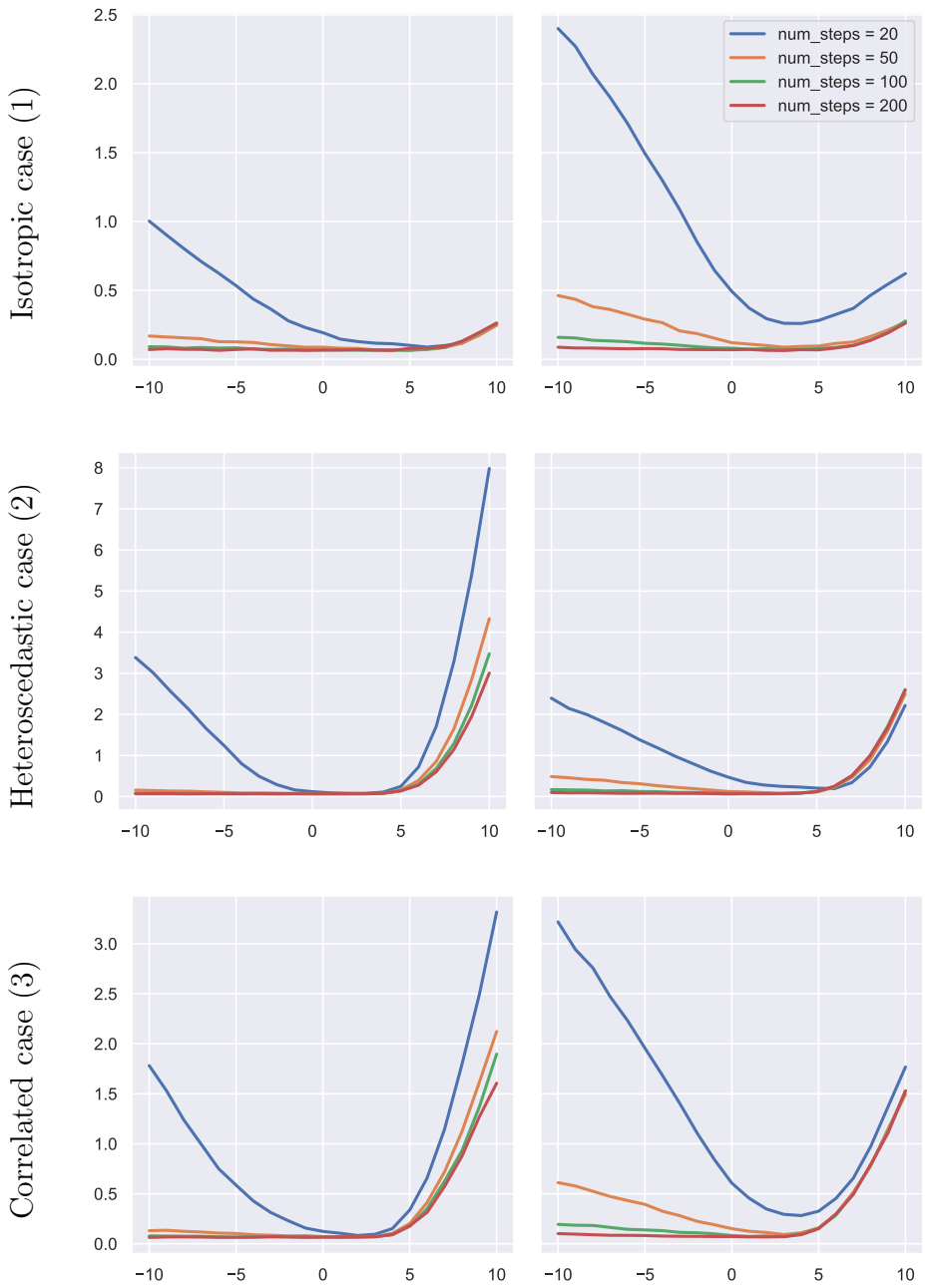


Figure 8: Empirical KL divergence between π_{data} and the generated distribution using the exact score function for different values of a in the noising function β_a, with either Euler-Maruyama discretization (left) or Exponential Integration discretization (right) and `num_steps` discretization steps.

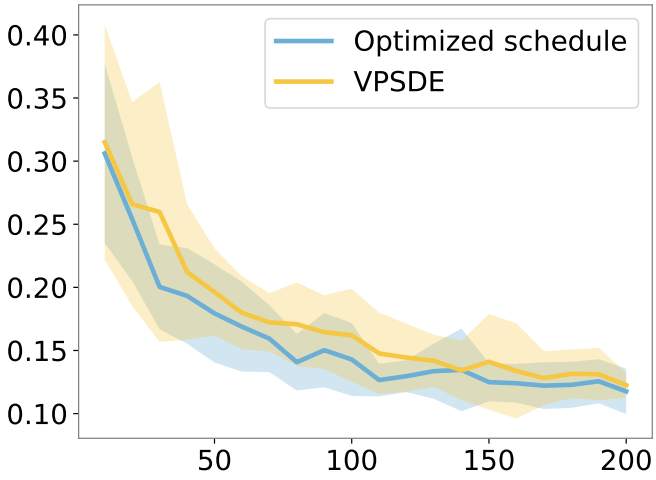


Figure 9: Empirical KL divergences (mean $\pm$ std over 10 runs) between π_{data} and the distributions obtained by Algorithm 1 (blue) and the VPSDE model (yellow).