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

Stanislas Strasman[†], Antonio Ocello[*], Claire Boyer[+,°], Sylvain Le Corff[†], and Vincent Lemaire[†]

[†]LPSM, Sorbonne Université, UMR CNRS 8001, Paris, France.
[*]CMAP, École Polytechnique, Institut Polytechnique de Paris, France.
[+]LMO, Université Paris-Saclay, UMR CNRS 8628, Orsay, France.
[°]IUF, Institut Universitaire de France.

Reviewed on OpenReview: https://openreview.net/forum?id=BlYIPa0Fx1&noteId=BlYIPa0Fx1

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. 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. Under additional regularity assumptions, taking advantage of favorable underlying contraction mechanisms, we provide a tighter error bound in Wasserstein distance compared to state-of-the-art results. In addition to being tractable, this upper bound jointly incorporates properties of the target distribution and SGM hyperparameters that need to be tuned during training. Finally, we illustrate these bounds through numerical experiments using simulated and CIFAR-10 datasets, identifying an optimal range of noise schedules within a parametric family.

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 di usion 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 $_{\text{data}}$ through standard non-parametric density estimation strategies. Score-based Generative Models (SGMs) are probabilistic models designed to address this challenge using two main phases. The rst 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 $_1$. 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 di usion. 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 $_1$, we obtain a generative distribution that approximates $_{\text{data}}$.

Related works. Signi cant attention has been paid to understanding the sources of errors that a ect the quality of data generation associated with SGMs (Block et al., 2020; De Bortoli, 2022; Lee et al., 2022; 2023; Chen et al., 2023a;b). 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 di usion 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); Gao et al. (2023) also require smoothness assumptions on the data distribution, either involving non-explicit constants, or focusing on iteration complexity sharpness. More recently, Conforti et al. (2023); Benton et al. (2024) 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 (e.g., assumed to be of nite 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); Guo et al. (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 propose a uni ed framework for time-inhomogeneous SGMs, to conduct joint theoretical analyses in KL and Wasserstein metrics, with state-of-the-art set of assumptions, using exponential integration of the backward process. In our opinion, these upper-bounds provide numerical insights into proper SGM training.

- We establish an upper bound on the Kullback-Leibler divergence between the data distribution and the law of the SGM. This bound holds under the mildest assumptions used in the SGM literature and explicitly depends on the noise schedule used to train the SGM. The proof follows the same steps as Conforti et al. (2023). However, it requires to establish a Kullback-Leibler upper bound for an inhomogeneous forward di usion which involves

determining a non-asymptotic rate of convergence for the mixing time using Fokker-Planck equations and a log-Sobolev inequality that depends on the noise schedule, and not only on the di usion time horizon, see Lemma B.1. In addition, taking into account the backward contraction for the di usion process (Proposition C.1) provides state-of-the-art results on mixing time convergence for SGM under the Ornstein-Uhlenbeck forward process, whether inhomogeneous or not.

- By making additional assumptions on the Lipschitz and strong log-concavity properties of the score function, we establish a bound in terms of Wasserstein distance explicitly depending on the noise schedule. This extends the similar result for the KL in the Gaussian setting. These results are in the same line of work as Bruno et al. (2023); Gao et al. (2023), incorporating to the time inhomogeneous setting a re nement of the mixing time error based on an analysis of the modi ed score function.
- We illustrate, through numerical experiments, the upper bounds obtained in practice in regard of the e ective empirical KL divergences and Wassertein metrics. These simulations highlights the relevancy of the upper bound, re ecting in practice the e ect of the noise schedule on the quality of the generative distribution. Additionally, the simulations conducted provide theoretically-inspired guidelines for improving SGM training. For reproducibility purposes, the code for the numerical experiments is available at https://github.com/StanislasStrasman/Noise_Schedule_for_Score-based_Generative_Models .

2 Mathematical framework for SGMs

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

$$dX_t = -\frac{(t)}{2} X_t dt + \sqrt{(t)}\, dB_t; \quad X_0 \sim q_{data}: \tag{1}$$

We denote by p_t the density of X_t at time $t \in (0;T]$. Note that, up to the time change $t \to R_0^t (s)=2ds$, this process corresponds to the standard Ornstein Uhlenbeck (OU) process, solution to

$$dX_t = -\frac{1}{2} X_t dt + \sqrt{2}\, dB_t; \quad X_0 \sim q_{data};$$

see, e.g., Karatzas and Shreve (2012, Chapter 3). Due to the linear nature of the drift with respect to $(X_t)_t$, it is well-known that an exact simulation can be performed for this process (Section E.1.2). The stationary distribution $_1$ of the forward process is the Gaussian distribution with mean 0 and variance $^2 I_d$. In the literature, when (t) is constant equal to 2 (meaning that there is no time change), this di usion 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 Di usion Probabilistic Models (DDPM, Ho et al., 2020). Understanding the e ects of the general di usion 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 given by

$$\begin{cases} dX_t = (t; X_t)dt + \sqrt{(t)}dB_t; \\ X_0 \quad _1; \end{cases} \quad \text{with} \quad \begin{cases} (t) := (T \quad t) \\ (t; X_t) := \frac{(t)}{2^{\ 2}} X_t + (t) r \log p_{T \quad t}(X_t) : \end{cases}$$

We consider the marginal time distribution of the forward process divided by the density of its stationary distribution, introducing

$$8x \ 2 \ R^d; \quad \tilde{p}_t(x) := p_t(x)='_{\ 2}(x); \tag{2}$$

where $'_{\ 2}$ denote the density function of $_1$, a Gaussian distribution with mean 0 and variance $^2 I_d$. Thus, the backward process can be rewritten as

$$dX_t = \quad t; X_t \quad dt + \sqrt{(t)}dB_t; \quad X_0 \quad _1; \tag{3}$$

where $(t; X_t) := \frac{(t)}{2^{\ 2}} X_t + (t) r \log \tilde{p}_{T \quad t}(X_t)$. The bene t 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 (modi ed) score function $r \log \tilde{p}_t(x) = r \log p_t(x) + x= \ ^2$ cannot be evaluated directly, because it depends on the unknown data distribution. To work around this problem, the score function $r \log p_t$ 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 di usion models, this is typically done with the use of a deep neural network architecture $s : [0; T] \ R^d \ ! \ R^d$ parameterized by 2 , and trained to minimize:

$$L_{\text{explicit}}( ) = E \quad s \quad ; X \quad r \log p \quad X \quad ^2; \tag{4}$$

with $U(0; T)$ independent of the forward process $(X_t)_{t \ 0}$. However, this estimation problem still su ers from the fact that the regression target is not explicitly known. A tractable optimization problem sharing the same optima can be de ned though, through the marginalization over $_{\text{data}}$ of p (see Vincent, 2011; Song et al., 2021):

$$L_{\text{score}}( ) = E \quad s \quad ; X \quad r \log p \quad X \ jX_0 \quad ^2; \tag{5}$$

where is uniformly distributed on $[0; T]$, and independent of $X_0 \quad _{\text{data}}$ and $X \quad p(\ jX_0)$. This loss function is appealing as it only requires to know the transition kernel of the forward process. In (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. (2021) propose an Euler-Maruyama (EM) discretization scheme in which both the drift and the di usion coe cients are discretized recursively (see (50)). 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 modi ed score function. Introduce $\bar{s}(t,x) := s(t,x) + x/2$ and consider the regular time discretization $0 = t_0 < t_1 < \cdots < t_N = T$. Then, $(X_t)_{t\in[0,T]}$ is such that, for $t \in [t_k, t_{k+1}]$,

$$dX_t = \gamma(t)\left(\frac{1}{2}X_t + \bar{s}(T - t_k, X_{t_k})\right)dt + \sqrt{\gamma(t)}dB_t, \quad X_0 \sim \pi_1 . \tag{6}$$

This scheme can be seen as a re nement of the classical EM one as it handles the linear drift term by integrating it explicitly. In addition, $(X_t)_{t\in\{t_0,\dots,t_N\}}$ can be sampled exactly, see Appendix A. We consider therefore such a scheme in our further theoretical developments.

3 Non-asymptotic Kullback-Leibler bound

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

Statement. The data distribution π_{data} is assumed to be absolutely continuous with respect to the Gaussian measure π_1. De ne the relative Fisher information $I(\pi_{\text{data}} | \pi_1)$ by

$$I(\pi_{\text{data}} | \pi_1) := \int \left\| \nabla \log \frac{d\pi_{\text{data}}}{d\pi_1} \right\|^2 d\pi_{\text{data}} ;$$

and consider the following assumptions.

- H1 The noise schedule γ is continuous, positive, non decreasing and such that $\int_0^{R_1} \gamma(t)dt = +\infty$.
- H2 The data distribution is such that $I(\pi_{\text{data}} | \pi_1) < +\infty$.
- H3 The NN parameter $\hat{s}$ and the schedule γ satisfy

$$\mathbb{E}\left[\exp\left(\frac{1}{2}\int_0^T \gamma(t)\left\|\bar{s}(T-t, X_t) - \hat{s}(T-t, X_t)\right\|^2 dt\right)\right] < +\infty ;$$

with $\bar{s}(t,x) := \nabla \log p_t(x)$ and p_t de ned in (2).

Assumption H1 is necessary to ensure that the forward process converges to the stationary distribution when the di usion time tends to in nity. 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, in this section, we do not require extra assumptions about the smoothness of the score function. Lastly, Assumption H3 is a condition on the approximation of the score by the neural network $\hat{s}$, 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 $\pi_N^{(\cdot;\cdot)}$ of samples obtained from (6), and the target data distribution π_{data}.

Theorem 3.1. Assume that H1, H2 and H3 hold. Then,

$$\mathrm{KL}\left(\pi_{\mathrm{data}} \,\middle|\, \hat{\pi}_N^{(\hat{\gamma},\hat{\theta})}\right) \le E_1^{\mathrm{KL}}(\gamma) + E_2^{\mathrm{KL}}(\hat{\gamma},\hat{\theta}) + E_3^{\mathrm{KL}}(\gamma),$$

where

$$E_1^{\mathrm{KL}}(\gamma) = \mathrm{KL}\left(\pi_{\mathrm{data}} \,\middle|\, \pi_{\gamma}\right) \exp\left(-\frac{1}{2}\int_0^T \alpha(s)\,\mathrm{d}s\right),$$

$$E_2^{\mathrm{KL}}(\hat{\gamma},\hat{\theta}) = \sum_{k=0}^{N-1} \mathbb{E}\left[r\left(\log p_{T-t_k}\left(\cdot \,\middle|\, X_{T-t_k}\right) - \log p_{T-t_k}^{\hat{\gamma},\hat{\theta}}\left(\cdot \,\middle|\, X_{T-t_k}\right)\right)\right] \int_{T-t_{k+1}}^{T-t_k} \alpha(t)\,\mathrm{d}t,$$

$$E_3^{\mathrm{KL}}(\gamma) = 2\, h\, \alpha(T)\, I\left(\pi_{\mathrm{data}} \,\middle|\, \pi_{\gamma}\right),$$

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

The obtained bound is composed of three terms, all depending on the noise schedule, through either its integrated version over the diffusion time, or its final value at time T. The result was derived for the EI discretization scheme, but it could be adapted to the Euler scheme up to minor technicalities. Remark also that using Pinsker's inequality, the obtained bound could be transferred in terms of total variation.

Dissecting the upper bound. The upper bound of Theorem 3.1 involves three different types of error that affect the training of an SGM. The term E_1^{KL} 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, E_1^{KL} is shrinked to 0 when T grows to infinity. Note that the multiplicative term in E_1^{KL} corresponds to the KL divergence between π_{data} and π_{γ} which is ensured to be finite by Assumption H2. The second term E_2^{KL} 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 (see De Bortoli et al., 2021, Equation (8)), the term E_2^{KL} admits as a crude bound $M \int_0^T \alpha(t)\,\mathrm{d}t$, 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, E_3^{KL} is the *discretization error* of the EI discretization scheme. This last term vanishes as the discretization grid is refined (i.e., $h \to 0$).

Comparison with existing bounds. Under perfect score approximation,

and infinitely precise discretization (i.e., when $E_2^{\mathrm{KL}}(\hat{\gamma},\hat{\theta}) = E_3^{\mathrm{KL}}(\gamma) = 0$), we recover that the Variance Preserving SDE (VPSDE, De Bortoli et al., 2021; Conforti et al., 2023; Chen et al., 2023b) converges exponentially fast to the target distribution. Beyond this idealized setting, the bound established in Theorem 3.1 recovers that of Conforti et al. (2023, Theorem 1) when choosing $\alpha(t) = 2$, $\sigma^2 = 1$, $T = 1$, and using a discretization step size $h \to 1$.

Refined analysis of the mixing time error. Still assuming perfect score approximation and infinitely precise discretization (i.e., $E_2^{\mathrm{KL}}(\hat{\gamma},\hat{\theta}) = E_3^{\mathrm{KL}}(\gamma) = 0$), one can assess the sharpness of the term $E_1^{\mathrm{KL}}(\gamma)$ in the upper bound of Theorem 3.1. In particular, when restricting the data distribution to be Gaussian $\mathcal{N}(m_0, \sigma_0)$, one can exploit the backward contraction assuming that

$_{\max}(\ _0)\quad ^2$, where $_{\max}(\ _0)$ denotes the largest eigenvalue of $_0$. In this speci c case, we can obtain a re ned version for E_1^{KL} (see Proposition C.1), given by

$$KL(\ _{data}\,k\ _1\,Q_T)\quad KL(\ _{data}\,k\ _1)\exp\quad \frac{2}{^2}\int_0^T\ (s)ds\ ; \tag{7}$$

where $(Q_t)_{0\ t\ T}$ is the Markov semi-group associated with the backward SDE. This idea is exploited in Section 4 to establish Wasserstein bounds for more general data distributions than Gaussian, but requiring extra regularity of the score.

4 Non-asymptotic Wasserstein bound

In the literature, much attention is paid to derive upper bounds with other metrics such as the W_2 distance, which has the advantage to be a distance and to have easier-to-handle and implementable estimators. In Lee et al. (2023), the authors obtain a control for the 2-Wasserstein and total variation distances. However, those results rely on additional assumptions on$_{data}$ (which is assumed to have bounded support for instance in De Bortoli (2022)).

Regularity assumptions. We consider extra regularity assumptions about the modi ed marginal density p_t at any time of the di usion.

H4 (i) For all $t\quad 0$, there exists $C_t\quad 0$ such that for all $x;y\ 2\ R^d$,

$$(r\ \log p_t(y)\ r\quad \log p_t(x))^{>}(x\quad y)\quad C_t\,kx\quad yk^2\ :$$

(ii) For all $t\quad 0$, there exists $L_t\quad 0$ such that $r\ \log p_t$ is L_t-Lipschitz continuous.

The strong log-concavity (i) (see, e.g., Saumard and Wellner, 2014) plays a crucial role in terms of contraction of the backward SDE. Classical distributions satisfying H4(i) include logistic densities restricted to a compact set, or Gaussian laws with a positive de nite covariance matrix, see Saumard and Wellner (2014) for other examples. We observe, notably, that when the density of the data distribution is log-concave, this property propagates within the probability ow $(p_t)_{0\ t\ T}$ (see Proposition D.1). Similar conclusions can be drawn regarding the Lipschitz continuity of the score (Proposition D.2). This property is formalized in the Lemma 4.1 for Gaussian distributions.

Lemma 4.1. Assume that $_{data}$ is a Gaussian distribution $N(\ _0;\ _0)$, such that $_0$ is invertible and $_{\max}(\ _0)\ <\ ^2$. Let $m_t\ :=\ \exp(\ \int_0^t\ (s)=2\ ^2ds)$. Then, the probability ow p_t given by (1) initialized at $_{data}$ is C_t-strongly log concave, with

$$C_t\ :=\ \frac{m_t^2\quad ^2\quad _{\max}(\ _0)}{m_t^2\quad _{\max}(\ _0)+\ ^2(1\quad m_t^2)}\ :$$

In addition, the associated score$r\ \log p_t$ is L_t-Lipschitz continuous with

$$L_t\ :=\ \min\quad \frac{1}{^2(1\quad m_t^2)};\frac{1}{_{\min}(\ _0)m_t^2}\quad +\ \frac{1}{^2}\ :$$

This result, restricted to the Gaussian case, sets the focus on the importance of calibrating the parameter ^{2} depending on the covariance structure of the data distribution, in order to enhance strong log concavity of the probability ow through the di usion.

Error bound. To establish a 2-Wasserstein bound explicitly depending on the noise schedule, we consider the following additional assumptions, respectively about uniform approximation of the score, and Lipschitz continuity in time of the renormalized score.

H5 There exists $\varepsilon \geq 0$ such that $\sup_{k \in \{0,\dots,N-1\}} \| s(T - t_k, X_{t_k}) - s_\theta(T - t_k, X_{t_k}) \|_{L_2} \leq \varepsilon$.

H6 For a regular discretization $\{t_k, 0 \leq k \leq N\}$ of $[0,T]$ of constant step size h, there exists $M \geq 0$ such that

$$\sup_{k \in \{0,\dots,N-1\}} \sup_{t \in [t_k, t_{k+1}]} \| s(T - t, x) - s(T - t_k, x) \|_{L_2} \leq M h (1 + \|x\|).$$

We now have all the ingredients to present our theoretical guarantee in terms of Wasserstein distance.

Theorem 4.2. Assuming H4, H5 and H6 and that the time step h is small enough, it holds that

$$W_2(q_{\text{data}}, q_N^{(\gamma,\theta)}) \leq \mathcal{E}_1^{W_2}(\Gamma) + \mathcal{E}_2^{W_2}(\gamma, \Gamma) \tag{8}$$

with $$\mathcal{E}_1^{W_2}(\Gamma) = W_2(\gamma_{\text{data}}, \gamma_1) \exp\left(\int_0^T \frac{\Gamma(t)}{2} (1 + C_t^2)\, dt \right);$$

$$\mathcal{E}_2^{W_2}(\gamma, \Gamma) = \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} L_t \Gamma(t) dt \left(\sqrt{2h\Gamma(T)} + \frac{h\Gamma(T)}{2^2} + \int_{t_k}^{t_{k+1}} 2L_t \Gamma(t) dt \right) B + \varepsilon T \Gamma(T) + MhT\Gamma(T)(1 + 2B);$$

$$B = (\mathrm{E}[\|X_0\|^2] + \gamma^2 d)^{1/2}; \text{ and for all } t \in [0,T], L_t = L_{T-t}.$$

In Theorem 4.2, we exploit the contraction entailed by Assumption H4(i) of the backward diffusion processes on top of that of the forward phase. This idea leads to an improvement of all the existing bounds in Wasserstein metrics, by refining their mixing time term. The previous result can be established when the target distribution has a Lipschitz continuous score and is strongly log-concave: by propagating these properties, the constants L_t and C_t can be characterized as a function of L_0 and C_0 (see Propositions D.1 and D.2). The propagation of the log-concave property was also established in Saremi et al. (2023).

Corollary 4.3. Assume that $r \log p_{\text{data}}$ is L_0-Lipschitz, that $\log p_{\text{data}}$ is C_0-strongly concave such that $C_0 > 1/\gamma^2$. Under Assumption H5 and H6, with a time step h small enough,

$$W_2(q_{\text{data}}, q_N^{(\gamma,\theta)}) \leq W_2(\gamma_{\text{data}}, \gamma_1) \exp\left(\int_0^T \frac{\Gamma(t)}{2} (1 + C_t^2)\, dt \right) + c_1 \sqrt{h} + c_2 h + \varepsilon T \Gamma(T);$$

with $c_1 = L_0 \Gamma(T) T \sqrt{2\Gamma(T)}$ and $c_2 = \Gamma(T) T L_0 (1/(2\gamma^2) + 2) L_0 \Gamma(T) B + M(1 + 2B)$.

This provides an easy-to-handle upper bound in Wasserstein distance, encompassing the three types of error (e.g., mixing time, score approximation and discretization error), for Lipschitz scores and strongly-log concave distributions. We remark that it also exhibits an extra term in $\sqrt{h}$ compared to the more general KL bound obtained under milder assumptions. Note however that this term is in line with what can be found in the literature for Wasserstein bounds for SDE approximation (see Alfonsi et al., 2015).

Discussion and comparison with other works. Theorem 4.2 requires more stringent assumptions on the regularity of the score function than Theorem 3.1. However, these assumptions are not specific to our setting. In particular, the strong log-concavity assumption has proven to be a key property for the fast convergence of sampling algorithms (see Dalalyan, 2017; Durmus and Moulines, 2017; Dwivedi et al., 2019). While this is a strong assumption to require on the data density, this can be mitigated.

In Benton et al. (2024), the authors propose quantitative bounds for the Kullback-Leibler divergence only assuming a finite second moments of the data distribution and do not use any smoothness assumption. The authors use early stopping and stop the backward sampling at small time > 0 to avoid the score explosion in the neighborhood of 0. In another line of work, Chen et al. (2023a) used a high-probability bound on the Hessian matrix of p_t to avoid additional assumptions (such as a bounded support) on the data distribution to obtain Kullback-Leibler upper bounds. These works offer promising perspective to obtain Wasserstein bounds under weaker assumptions.

The analysis of the modified score functions in the Gaussian case reveals that by properly adjusting the variance of the stationary distribution of the forward process and rescaling the target distribution, we can attain the desired properties for the score function. This observation is specific to the Gaussian case as one can easily derive values for Γ_0 and L_0 with the eigenvalues of covariance matrix of the target distribution. However, we would like to strengthen the fact that a similar preprocessing has been applied to more complex distributions in Section 5.

5 Evaluation of the theoretical upper bounds

The goal of this section is to numerically illustrate the validity of the theoretical bounds obtained in Theorem 3.1 and Theorem 4.2. More precisely, we aim at unraveling the contributions of each error term of the upper bounds. We consider a simulation design where the target distribution is known, and the associated constants of interest (i.e., the strong log concavity parameter, the Lipschitz constant, $W_2(\pi_{\text{data}}, \pi_1)$, $I(\pi_{\text{data}} | \pi_1)$ or $\mathrm{KL}(\pi_{\text{data}} | \pi_1)$) can be evaluated. The error bounds are assessed for different choices of noise schedules of the form

$$a(t) \propto (e^{at} - 1)/(e^{aT} - 1), \tag{9}$$

with $a \in \mathbb{R}$ ranging from -10 to 10 with a unit step size. We set $T = 1$ and adjust schedules so that they all start at $(0) = 0.1$ and end at $(1) = 20$ (see Figure 1). This choice has been made so that when $a = 0$ the schedule is linear and matches exactly the classical VPSDE implementation (Song and Ermon, 2019; Song et al., 2021).

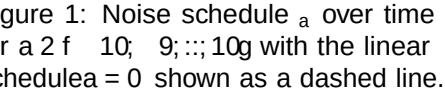

Figure 1: Noise schedule a over time for $a \in \{-10, -9, \ldots, 10\}$ with the linear schedule $a = 0$ shown as a dashed line.

5.1 Gaussian setting

Target distributions. We consider the setting where the true distribution π_{data} is Gaussian in dimension $d = 50$ with mean 1_d and different choices of covariance structure:

1. (Isotropic, denoted by $\pi_{\text{data}}^{(\text{iso})}$) $\Sigma^{(\text{iso})} = 0.5 I_d$.

2. (Heteroscedastic, denoted by $\Sigma_{\text{data}}^{(\text{heterosc})}$) $\Sigma^{(\text{heterosc})} \in \mathbb{R}^{d \times d}$ is a diagonal matrix such that $\Sigma_{jj}^{(\text{heterosc})} = 1$ for $1 \leq j \leq 5$, and $\Sigma_{jj}^{(\text{heterosc})} = 0.01$ otherwise.

3. (Correlated, denoted by $\Sigma_{\text{data}}^{(\text{corr})}$) $\Sigma^{(\text{corr})} \in \mathbb{R}^{d \times d}$ 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 $1 \leq j \neq j' \leq d$.

SGM simulations. We simulate $b_N^{(a,\tau)}$ from SGM using the forward process defined in (1) with $t \mapsto \alpha_a(t)$ for the noise schedule. The score is learned via a dense neural network with 3 hidden layers of width 256 over 150 epochs (see Figure 11) trained to optimize $\mathcal{L}_{\text{explicit}}$ (4). This is feasible because the score is analytically derived when ρ_{data} is Gaussian (Lemma E.1). Numerical experiments have also been run with the commonly used conditional loss $\mathcal{L}_{\text{score}}$, without changing the nature of the conclusions, see Appendix F. For backward process simulation, we use an Euler-Maruyama scheme with 500 steps, as being the most encountered discretization in practice (with these discretization steps the difference with Exponential Integrator scheme is minimal as highlighted in the appendix). For each value of a, and each data distribution, we train the SGM using $n = 10000$ training samples.

KL bound. In Figure 2 (top), we compare the empirical KL divergence between ρ_{data} and samples from $b_N^{(a,\tau)}$ to the upper bound from Theorem 3.1. We refer the reader to Appendix E.2.1 for implementation details. For Gaussian distributions, both the bound and KL divergence can be computed using closed-form expressions (see Lemma E.2 and E.4). In all scenarios the noise schedule significantly impacts the value of $\mathrm{KL}(\rho_{\text{data}} \| b_N^{(a,\tau)})$, and thereby the quality of the learned distribution. Moreover, in all three cases taking into acccount the contraction argument (7) is key to properly align the upper bound trend with the generation results. In all these experiments, the KL upper bound indicates possible values for a improving over the classical linear noise schedule.

2-Wasserstein bound. In Figure 2 (bottom), we compare the empirical W_2 distance between ρ_{data} and samples from $b_N^{(a,\tau)}$ to the upper bound from Theorem 4.2. For Gaussian distributions, both the bound and the W_2 distance can be computed using closed-form expressions (see Lemma 4.1, E.3, and E.5). For the isotropic case, the proposed W_2 upper bound reflects the SGM performances, as already highlighted by the KL bound. However, in non-isotropic cases, the raw distributions $\Sigma_{\text{data}}^{(\text{heterosc})}$ and $\Sigma_{\text{data}}^{(\text{corr})}$ do not directly satisfy Assumption 4 (i) when the variance of the stationary distribution is set to 1. Therefore, scaling the distributions in play becomes crucial for the theoretical W_2 upper bound to hold. That is why we propose the following preprocessing: train an SGM with centered and standardized samples of covariance $\Sigma^{(\text{stand})}$ rescaled in turn by a factor $1/(2\lambda_{\max}(\Sigma^{(\text{stand})}))^{1/2}$. This choice ensures that $\lambda_{\max}(\Sigma^{(\text{scaled})}) < \sigma^2 = 1$, for $\Sigma^{(\text{scaled})}$ the resulting covariance matrix, and thus the strong log-concavity of $p_0 = \rho_{\text{data}} * \mathcal{N}(0, \sigma^2 I_d)$. We call $b_{N;\,\text{scaled}}^{(a,\tau)}$ the resulting generative distribution, and the evaluated metrics is adjusted (see (51)) to ensure a fair numerical comparison. After this preprocessing, not only the W_2 upper bound of Theorem 4.2 aligns with the empirical performances but the SGM performances can be also boosted (see degraded empirical performances on raw distributions in Appendix E.2.2). This highlights the importance of properly calibrating the training sample to the stationary distribution of the SGM. Note that data normalization does not only enforce the strong log-concavity of the modified score at time 0, but can lower the ratio L_0/C_0. To see this, consider the heteroscedastic case, for which $\lambda_{\min}(\Sigma^{(\text{heterosc})})/\lambda_{\max}(\Sigma^{(\text{heterosc})}) = 100$, whereas $\lambda_{\min}(\Sigma^{(\text{scaled})})/\lambda_{\max}(\Sigma^{(\text{scaled})}) = 1$ after scaling. This Gaussian set-up reveals that data

(a) Isotropic setting (b) Heteroscedastic setting (c) Correlated setting

Figure 2: Comparison of the empirical KL divergence (mean± std over 30 runs) (top) and W_2 distance (mean± std over 10 runs)(bottom) between $_{data}$ and $b_N^{(\ ;\ )}$ (orange) and the related upper bounds (blue) from Theorem 3.1 and Theorem 4.2 across parameter for noise schedule $_a$, $d = 50$. In the KL case, the upper bounds in lighter blue are the theoretical upper bounds without taking into the contraction argument (7). We also show the metrics for the linear VPSDE model (dashed line) and our model (dotted line) with exact score evaluation.

renormalization improves the conditioning of the covariance matrix, and thereby the conditioning of SGM training. In particular, this is captured in the upper bound of Theorem 4.2 by limiting the growth of L_t and inducing a more balanced second term.

We now consider a varying dimension in$f5; 10; 25; 50g$, and we compare the empiricalW_2 distance obtained by (i) $_0$ the classical VPSDE (Song et al., 2021), with a linear noise schedule (i.e.$a, = 0$), (ii) $_{cos}$ the SGM with cosine schedule (Nichol and Dhariwal, 2021), and (iii) $_{a^?}$ the SGM with parametric schedule with$a = a^?$ approximately minimizing the upper bound from Theorem 3.1. In Figure 3, we observe that the SGMs run with $_{a^?}$ consistently outperforms those run with linear schedule $_0$ slightly improving the data generation quality. It displays lower average W_2 distances between $_{data}$ and the generated sample distribution, but also reduces the standard deviation of the resulting W_2 distances yielding more stable generation (see Table 2). These performances are comparable to, and often surpass, those achieved with state-of-the-art schedules like the cosine schedule, particularly in higher dimensions.

5.2 More general target distributions

Beyond Gaussian distributions, numerical analysis in terms of KL divergence is not tractable as standard estimators of the KL terms do not scale well with dimen-

(a) Isotropic setting (b) Rescaled heterosc. setting (c) Rescaled correlated setting

Figure 3: Comparison of the empirical W_2 distance (mean value std over 10 runs) between $_{\text{data}}$ and the generative distribution $b_N^{(\ ;\ )}$ across various dimensions. The distributions compared include SGMs with different noise schedules: $_{a^?}$ (blue solid), $_0$ (yellow dashed), and $_{\cos}$ (orange dotted).

sion. On the contrary, there exist computationally-efficient estimators of Wasserstein distances, as for instance the sliced W_2 estimate (Flamary et al., 2021). We use the latter to assess the relevancy of Theorem 4.2 when the target distribution corresponds to a 50-dimensional Funnel distribution defined as: $_{\text{data}}(x) = \ ' _{a^2}(x_1) \ Q_{j=2}^{d} \ ' _{\exp(2\ bx_1)}(x_j)$, with $a = 1$ and $b = 0{:}5$ (see Section E.2.3 for more details and additional experiments on a Gaussian mixture model). As previously, the samples are standardized and rescaled. In Figure 4, empirical results demonstrate that the minimum of the upper bound closely aligns with that of the empirical sliced 2-Wasserstein distance between the simulated and training data. Moreover, implementing SGM with the optimal parameter a yields consistent improvements of the data generation quality across different metrics w.r.t. to classical noise schedule competitors (linear or cosine). These experiments not only support the relevance of the theoretical upper bound beyond the assumptions required in Section 4, but also the validity of theoretically-inspired data preprocessing for improving SGM training with arbitrary target distributions.

Figure 4: Upper bound and sliced 2-Wasserstein distance on a Funnel dataset in dimension 50.

When dealing with high-dimensional real-world datasets, directly evaluating our theoretical upper bounds (Theorems 3.1 and 4.2) becomes more challenging because relevant quantities (distances and constants) are either poorly estimated or unavailable. As a first step toward real data, we evaluate the impact of the noise schedule on the sampling quality of models pre-trained using CIFAR-10 dataset. In Figure 5, we display the FID score with 50,000 generated samples using Euler-Maruyama discretization scheme for various noise schedules drawn from the parametric family in Equation (9). Additional implementation details are available in Appendix E.3. Although the assumptions underpinning our results cannot be verified in this setting, the empirical performance trends mirror closely those observed in the simulated settings.

This consistency highlights that analyzing and optimizing noise schedules could be a promising direction for improving SGM-based generation in more complex scenarios.

6 Discussion

In this paper, we propose a uni ed framework to analyze the impact of the noise schedule for time-inhomogeneous SGMs, providing upper bounds in KL and Wasserstein metrics. The KL upper bound follows the steps of recent works using the mildest assumptions used in the SGM literature. We also provide an improved upper bound in the Gaussian setting with numerical experiments highlighting the impact of the backward contraction of the forward noise process. Following Bruno et al. (2023); Gao et al. (2023), under additional assumptions on the Lipschitz and strong log-concavity properties of the score function, we establish upper bounds for the Wasserstein distance. This bound highlights the role of the noise schedule and provides a detailed analysis based on the modi ed score function. Our results are supported by numerical experiments in simple settings to highlight the several terms of the upper bounds and the role of the noise schedule. There are many perspectives to this work. Studying multi-dimensional noise schedules is of particular interest. Indeed, they could be useful to understand how to deal with target distributions with complex covariance structures, and thereby an alternative solution to data normalization issues. Establishing upper bounds for Wasserstein distances under milder assumptions remains an exciting open problem, which would shed light on the performances and limitations of score-based generative models. A speci c perspective would be to adapt our result using early-stopping to avoid the explosion of error terms in the neighborhood of 0 and to provide other assumptions to control the corresponding error close to 0.

Figure 5: FID Scores using 50,000 generated samples for the parametric and cosine schedules (CIFAR-10 dataset).

Acknowledgements

We would like to thank Gabriel Victorino Cardoso for his valuable insights and thoughtful help on the numerical experiments involving real-world datasets.

Antonio Ocello was funded by the European Union (ERC-2022-SYG-OCEAN-101071601). Views and opinions expressed are however those of the author only and do not necessarily re ect 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

Aurélien Alfonsi, Benjamin Jourdain, and Arturo Kohatsu-Higa. Optimal transport bounds between the time-marginals of a multidimensional di usion and its euler scheme. Electronic Journal of Probability, 2015.

Dominique Bakry, Ivan Gentil, Michel Ledoux, et al. Analysis and geometry of Markov di usion 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. Nearly d-linear convergence bounds for di usion models via stochastic localization. In The Twelfth International Conferenceon Learning Representations, 2024.

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

Stefano Bruno, Ying Zhang, Dong-Young Lim, Ömer Deniz Akyildiz, and Sotirios Sabanis. On di usion-based generative models and their error bounds: The log-concave case with full convergence estimates. arXiv preprint arXiv:2311.13584, 2023.

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

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In International Conferenceon 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 di usion models with minimal data assumptions, 2023b.

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

Jean-François Collet and Florent Malrieu. Logarithmic sobolev inequalities for inhomogeneous markov semigroups. European Seriesin 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 di usion models without early stopping: nite sher information is all you need, 2023.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. Journal of the Royal Statistical Society Series B: Statistical Methodology, 79(3):651 676, 2017.

Valentin De Bortoli. Convergence of denoising di usion models under the manifold hypothesis. Transactions on Machine Learning Research, 2022.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Di usion schrödinger bridge with applications to score-based generative modeling. Advances in Neural Information ProcessingSystems, 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.

Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted langevin algorithm. The Annals Applied Probability, 27(3):1551 1587, 2017.

Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-hastings algorithms are fast. Journal of Machine Learning Research, 20(183):1 42, 2019.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Coren os, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. Journal of Machine Learning Research, 22(78): 1 8, 2021.

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

Xuefeng Gao, Hoang M. Nguyen, and Lingjiong Zhu. Wasserstein convergence guarantees for a general class of score-based generative models, 2023.

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

Qiushan Guo, Sifei Liu, Yizhou Yu, and Ping Luo. Rethinking the noise schedule of di usion-based generative models. visible on Open Review, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising di usion 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.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of di usion-based generative models. In Advances in Neural Information Processing Systems, volume 35, pages 8595 8607, 2022.

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. Srdi : Single image super-resolution with di usion 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 di usion probabilistic models. In Proceedings of the IEEE/CVF Conferenceon Computer Vision and Pattern Recognition, pages 11461 11471, 2022.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising di usion probabilistic models. In Marina Meila and Tong Zhang, editors, Proceedingsof the 38th International Conferenceon Machine Learning, volume 139 of Proceedingsof Machine Learning Research, pages 8162 8171. PMLR, 18 24 Jul 2021.

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.

Saeed Saremi, Ji Won Park, and Francis Bach. Chain of log-concave markov chains. arXiv preprint arXiv:2305.19473, 2023.

Adrien Saumard and Jon A. Wellner. Log-concavity and strong log-concavity: A review. Statistics Surveys, 8(none):45 114, 2014. doi: 10.1214/14-SS107.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, Proceedingsof the 32nd International Conferenceon Machine Learning, volume 37 of Proceedings of Machine Learning Research, pages 2256 2265, Lille, France, 07 09 Jul 2015. PMLR.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In Advancesin Neural Information ProcessingSystems, 2019.

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

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

Achille Thin, Yazid Janati El Idrissi, Sylvain Le Cor , Charles Ollion, Eric Moulines, Arnaud Doucet, Alain Durmus, and Christian P Robert. NEO: Non equilibrium sampling on the orbits of a deterministic transform. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, Advancesin Neural Information ProcessingSystems, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/76tTYokjtG-abstract.html .

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.

Qing Wang, Sanjeev R. Kulkarni, and Sergio Verdu. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. IEEE Transactions on Information Theory, 55(5): 2392 2405, 2009. doi: 10.1109/TIT.2009.2016060.

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

A Notations and assumptions.

Consider the following notations, used throughout the appendices. For all $d \geq 1$, $\mu \in \mathbb{R}^d$ and de nite 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 $\varphi_{\sigma^2} = \varphi_{0,\sigma^2 I_d}$. When the context is clear, we may indi erently use the measure and the associated density w.r.t. the reference measure. For all twice-di erentiable real-valued function f, let Δf be the Laplacian of f. For all matrix $A \in \mathbb{R}^{m \times n}$, $\|A\|_{Fr}$ is the Frobenius norm of A, i.e., $\|A\|_{Fr} = (\sum_{i=1}^{m} \sum_{j=1}^{n} |A_{i,j}|^2)^{1/2}$. For all time-dependent real-valued functions $h : t \mapsto h_t$ or $f : t \mapsto f(t)$, we write $\overleftarrow{h}_t = h_{T-t}$ and $\overleftarrow{f}(t) = f(T-t)$ for all $t \in [0,T]$.

Let p_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

$$d\vec{X}_t = \alpha(t)\vec{X}_t dt + g(t) dB_t, \quad \vec{X}_0 \sim p_0, \tag{10}$$

and introduce $p_t : x \mapsto p_t(x) * \varphi_{\sigma^2}(x)$, where p_t is the probability density function of $\vec{X}_t$. The backward process associated with (10) is referred to as $(\overleftarrow{X}_t)_{t \in [0,T]}$ and given by

$$d\overleftarrow{X}_t = -\alpha(t)\overleftarrow{X}_t + \frac{g^2(t)}{2} \overleftarrow{X}_t + g^2(t) \nabla \log p_{T-t}(\overleftarrow{X}_t) dt + g(t) dB_t, \quad \overleftarrow{X}_0 \sim p_T, \tag{11}$$

with B a standard Brownian motion in $\mathbb{R}^d$. Moreover, consider

$$\sigma_t^2 := \exp\left(2 \int_0^t \alpha(s) ds\right) \int_0^t g^2(s) \exp\left(-2 \int_0^s \alpha(u) du\right) ds. \tag{12}$$

The approximate EI discretization of (11) considered in this paper is, for $t_k \leq t \leq t_{k+1}$, $0 \leq k \leq N-1$,

$$d\overleftarrow{X}_t = \left\{-\alpha(t)\overleftarrow{X}_t + g^2(t) s(T-t_k, \overleftarrow{X}_{t_k})\right\} dt + g(t) dB_t.$$

Sampling from this backward SDE is possible recursively for $k \in \{0, \dots, N-1\}$, with $(Z_k)_{1 \leq k \leq N} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, I_d)$. For $k \in \{0, \dots, N-1\}$, writing $\gamma_k = T - t_k$,

$$\overleftarrow{X}_{t_{k+1}} = e^{-\int_k^{k+1} \alpha(s) ds} \overleftarrow{X}_{t_k} + s(\gamma_k, \overleftarrow{X}_{t_k}) e^{-\int_k^{k+1} \alpha(s) ds} \int_{\gamma_k}^{\gamma_{k+1}} g^2(t) e^{\int_k^t \alpha(v) dv} dt + e^{-\int_k^{k+1} \alpha(s) ds} \left(\int_{\gamma_k}^{\gamma_{k+1}} e^{2\int_k^t \alpha(s) ds} g^2(t) dt\right)^{1/2} Z_{k+1}.$$

We denote by $Q_T \in \mathcal{P}(C([0,T]; \mathbb{R}^d))$ the path measure associated with the backward di usion and by $(Q_t)_{0 \leq t \leq T}$ its Markov semi-group. We also write $\overleftarrow{X}^1_T \sim Q_T$ and, for each time step t_k for $0 \leq k \leq N$, $\overleftarrow{X}^1_{t_k} \sim Q_{t_k}$. For each time step t_k for $0 \leq k \leq N$, the kernel associated with the backward discretization is denoted by $Q^{N;}_{t_k}$, so that we have $\overleftarrow{X}_{t_k} \sim Q^{N;}_{t_k}$.

In Appendix C, these notations are used for the speci c case where $\alpha : t \mapsto \alpha(t) = -(2\sigma^2)$ and $g : t \mapsto \alpha(t)^{1/2}$ and the associated backward discretization is given in (31).

B Proofs of Section 3

B.1 Proof of Theorem 3.1

We are interested in the relative entropy of the training data distribution $_{\text{data}}$ with respect to the generated data distribution $b_N^{(\ ;\ )}$. Leveraging the time-reverse property we have:

$$\mathrm{KL}\quad{}_{\text{data}}\quad b_N^{(\ ;\ )}\ =\ \mathrm{KL}\quad p_T\,Q_T\quad b_N^{(\ ;\ )}\ :$$

By the data processing inequality,

$$\mathrm{KL}\quad p_T\,Q_T\quad b_N^{(\ ;\ )}\qquad \mathrm{KL}\quad p_T\,Q_T\quad{}_1\ Q_T^{N;}\quad :$$

where Q_T and $Q_T^{N;}$ denote the path measures of, respectively, the backward process and the SGM generation. Writing the backward time $_t = T\quad t$ and its discretized version $_k = T\quad t_k$, with $0 = t_0 < t_1 < \cdots < t_N = T$, we have (by Lemma B.5) that

$$\mathrm{KL}\quad{}_{\text{data}}\,\|b_N^{(\ ;\ )}\qquad \mathrm{KL}\,(p_T\,\|\,{}_1\,)+\frac{1}{2}\int_0^T\frac{1}{(t)}E\Big[\quad\frac{(t)}{2\ ^2}X_t+\quad(t)r\ \log p_{\ t}\quad X_t$$
$$\frac{(t)}{2\ ^2}X_t+\quad(t)s\quad{}_k;X_{t_k}\qquad{}_2\Big]\,dt:$$

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:

$$\mathrm{KL}\quad{}_{\text{data}}\,\|b_N^{(\ ;\ )}\qquad \mathrm{KL}\,(p_T\,\|\,{}_1\,)+\frac{1}{2}\int_0^T\frac{1}{(t)}E\Big[\quad(t)\quad s\quad{}_t;X_t\qquad s\,(\ {}_k;X_{t_k})\qquad{}_2\Big]\,dt:$$

By using the regular discretization of the interval $[0;T]$, one can disentangle the last term as follows:

$$\mathrm{KL}\quad{}_{\text{data}}\quad b_N^{(\ ;\ )}\qquad \mathrm{KL}\,(p_T\,\|\,{}_1\,)+\frac{1}{2}\sum_{k=0}^{N\ 1}\int_{t_k}^{t_{k+1}}(t)E\quad s\quad{}_t;X_t\qquad s\quad{}_k;X_{t_k}\qquad^2\,dt$$
$$E_1(\ )+E_2(\ ;\ )+E_3(\ )\,;$$

where

$$E_1(\ )=\mathrm{KL}\,(\ p_T\,\|'\quad{}_2)\ ; \tag{13}$$

$$E_2(\ ;\ )=\sum_{k=0}^{N\ 1}\int_{t_k}^{t_{k+1}}(t)E\quad s\quad{}_k;X_{t_k}\qquad s\quad{}_k;X_{t_k}\qquad^2\,dt\,; \tag{14}$$

$$E_3(\ )=\sum_{k=0}^{N\ 1}\int_{t_k}^{t_{k+1}}(t)E\quad s\quad{}_t;X_t\qquad s\quad{}_k;X_{t_k}\qquad^2\,dt: \tag{15}$$

Finishing the proof of Theorem 3.1 amounts to obtaining upper bounds for $E_1(\ )$, $E_2(\ ;\ )$ and $E_3(\ )$. This is done in Lemmas B.1, B.2 and B.3, so that $E_1(\ ) \quad E_1(\ )$, $E_2(\ ;\ ) \quad E_2(\ ;\ )$ and $E_3(\ ) \quad E_3(\ )$.

Lemma B.1. For any noise schedule ,

$$E_1(\ ) = \mathrm{KL}(\ p_T \,k\, {}_1\ )\quad \mathrm{KL}(\ {}_{\text{data}} \,k\, {}_1\ )\exp\quad \frac{1}{2}\int_0^T \quad (s)ds\quad :$$

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

$$@p_t(x) = \frac{(t)}{2\ {}^2}\mathrm{div}\,(xp_t(x)) + \frac{(t)}{2}\quad p_t(x) = \frac{(t)}{2}\mathrm{div}\quad \frac{1}{2}xp_t(x) + r\ p_t(x)\quad ;$$

for $t \ 2\ [0;T]; x \ 2\ R^d$. Combing this with the derivation under the integral theorem, we get

$$\begin{aligned}
\frac{@}{@t}\mathrm{KL}(\,p_t k'\ {}_2) &= \frac{@}{@t}\int_{R^d} \log ,\frac{p_t(x)}{{}_2(x)}p_t(x)dx \\
&= \int_{R^d} \frac{@}{@t}p_t(x)\log ,\frac{p_t(x)}{{}_2(x)}dx + \int_{R^d} \frac{p_t(x)@p_t(x)}{p_t(x)}dx \\
&= \int_{R^d} \frac{@}{@t}p_t(x)\log ,\frac{p_t(x)}{{}_2(x)}dx + \int_{R^d} \frac{@}{@t}p_t(x)dx \\
&= \int_{R^d} \frac{(t)}{2}\mathrm{div}\quad \frac{x}{2}p_t(x) + r\ p_t(x)\quad \log ,\frac{p_t(x)}{{}_2(x)}dx \\
&= \frac{(t)}{2}\int_{R^d} \mathrm{div}\,(r\quad \log '\ {}_2(x)\ p_t(x) + r\ p_t(x))\log ,\frac{p_t(x)}{{}_2(x)}dx \\
&= \frac{(t)}{2}\int_{R^d} (r\quad \log '\ {}_2(x)\ p_t(x) + r\ p_t(x))^{>}\ r\ \log ,\frac{p_t(x)}{{}_2(x)}dx \\
&= \frac{(t)}{2}\int_{R^d} p_t(x)(r\quad \log '\ {}_2(x) + r\ \log p_t(x))^{>}\ r\ \log ,\frac{p_t(x)}{{}_2(x)}dx \\
&= \frac{(t)}{2}\int_{R^d} p_t(x)\quad r\ \log ,\frac{p_t(x)}{{}_2(x)}\quad ^2 dx :
\end{aligned}$$

Using the Stam-Gross logarithmic Sobolev inequality given in Proposition B.6, we get

$$\frac{@}{@t}\mathrm{KL}(\,p_t k'\ {}_2)\quad \frac{(t)}{2}\mathrm{KL}(\,p_t k'\ {}_2) :$$

Applying Grönwall's inequality, we obtain

$$\mathrm{KL}(\,p_T k'\ {}_2)\quad \mathrm{KL}(\,p_0 k'\ {}_2)\exp\quad \left(\frac{1}{2}\int_0^T \quad (s)ds\quad ;\right.$$

which concludes the proof. □

Lemma B.2. For all and all ,

$$E_2(\ ;\ ) = \sum_{k=1}^{N} E\left[\ r\ \log p_{t_k}\left(X_{t_k}\right)\ -\ s\left(t_k, X_{t_k}\right)\right]^2 \int_{t_k}^{t_{k+1}} (t)dt\ ;$$

where $E_2(\ ;\ )$ is defined by (14).

Proof. By definition of $E_2(\ ;\ )$,

$$\begin{aligned} E_2(\ ;\ ) &= \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} (t) E\left[\ r\ \log p_{T-t_k}\left(X_{t_k}\right)\ -\ s\left(T-t_k, X_{t_k}\right)\right]^2 dt \\ &= \sum_{k=0}^{N-1} E\left[\ r\ \log p_{T-t_k}\left(X_{t_k}\right)\ -\ s\left(T-t_k, X_{t_k}\right)\right]^2 \int_{t_k}^{t_{k+1}} (t)dt \\ &= \sum_{k=0}^{N-1} E\left[\ r\ \log p_{t_k}\left(X_{t_k}\right)\ -\ s\left(t_k, X_{t_k}\right)\right]^2 \int_{t_k}^{t_{k+1}} (t)dt\ ; \end{aligned}$$

where the last equality comes from the fact that the forward and backward processes have same marginals since $X_T \sim p_T$. □

Lemma B.3. Assume that H1 holds. For all $T,\ > 0$, and all ,

$$E_3(\ ) \quad 2h\ (T) \max\left(\frac{h\ (T)}{4\ ^2}; 1\right) I\left(\ _{\text{data}} \| \ _1\right);$$

where $E_3(\ )$ is defined by (15).

Proof. By Lemma B.7, with $Y_t := r\ \log p_{T-t}(X_t)$,

$$dY_t = \frac{(t)}{2\ ^2} Y_t dt + \sqrt{(t)} Z_t dB_t\ :$$

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

$$d\|Y_t\|^2 = \left(-\frac{(t)}{2} \|Y_t\|^2 + (t)\|Z_t\|_{Fr}^2\right) dt + \sqrt{(t)}\, Y_t^\top Z_t dB_t\ :$$

Fix > 0. From Baldi (2017, Theorem 7.3, p.193), we have that $\left(\int_0^t g(s) Y_s^\top Z_s dB_s\right)_{t \in [0;T-]}$ is a square integrable martingale if

$$E\left[\int_0^{T-} g^2(s)\ \left\|Y_s^\top Z_s\right\|^2 ds\right] < 1\ :$$

From the Cauchy-Schwarz inequality, we get that

$$E\left[\left\|Y_s^\top Z_s\right\|_2^2\right] \quad E\left[\|Y_s\|_2^2 \|Z_s\|_{Fr}^2\right] \quad E\left[\|Y_s\|_2^4\right]^{1=2} E\left[\|Z_s\|_{Fr}^4\right]^{1=2}\ :$$

Applying Lemma B.8 and B.9, we get that both $\mathbb{E}[\|Y_s\|_2^4]$ and $\mathbb{E}[\|Z_s\|_2^4]$ are bounded by a quantity depending on $T^8 t$. As the term $T^8 t$ is uniformly bounded in $[0; T]$ and by Fubini's theorem, $\mathbb{E}[\int_0^T g^2(s)\|Y_s^\top Z_s\|^2 ds] = \int_0^T g^2(s)\mathbb{E}[\|Y_s^\top Z_s\|^2]ds < 1$. Therefore, $(\int_0^t g(s) Y_s^\top Z_s dB_s)_{t\in[0;T]}$ is a square integrable martingale. Therefore, we have

$$\mathbb{E}\|Y_t\|^2 - \mathbb{E}\|Y_{t_k}\|^2 = \mathbb{E}\int_{t_k}^{t} \frac{(s)}{2}\|Y_s\|^2 ds + \int_{t_k}^{t} (s)\|Z_s\|_{Fr}^2 ds;$$

and

$$\begin{aligned}
\mathbb{E}\left[\|Y_t - Y_{t_k}\|^2\right] &= \mathbb{E}\left[\left\|\int_{t_k}^{t} \frac{(s)}{2^{\ 2}} Y_s ds + \int_{t_k}^{t} \sqrt{\quad}\ (s) Z_s dB_s\right\|^2\right] \\
&\quad 2\mathbb{E}\left[\left\|\int_{t_k}^{t} \frac{(s)}{2^{\ 2}} Y_s ds\right\|^2\right] + 2\mathbb{E}\left[\left\|\int_{t_k}^{t} \sqrt{\quad}\ (s) Z_s dB_s\right\|^2\right] \\
&\quad 2\mathbb{E}\left[\left(\frac{1}{2}\int_{t_k}^{t} \sqrt{\quad}\ (s)\frac{\sqrt{\quad}\ (s)}{\quad} Y_s ds\right)^2\right] + 2\mathbb{E}\left[\left\|\int_{t_k}^{t} \sqrt{\quad}\ (s) Z_s dB_s\right\|^2\right] \\
&\quad \frac{1}{2^{\ 2}}\int_{t_k}^{t_{k+1}} (s) ds\, \mathbb{E}\left[\int_{t_k}^{t_{k+1}} \frac{(s)}{2}\|Y_s\|^2 ds\right] + 2\mathbb{E}\int_{t_k}^{t_{k+1}} (s)\|Z_s\|_{Fr}^2 ds \\
&\quad 2\max\left(\frac{\int_{t_k}^{t_{k+1}} (s) ds}{4^{\ 2}}; 1\right)\left(\mathbb{E}\|Y_{t_{k+1}}\|^2 - \mathbb{E}\|Y_{t_k}\|^2\right). \qquad (16)
\end{aligned}$$

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

$$\begin{aligned}
&\sum_{k=0}^{N\ 1}\int_{t_k}^{t_{k+1}} (t)\mathbb{E}\left[\left\| r \log p_{T\ t}(X_t) - r \log p_{T\ t_k}(X_{t_k})\right\|^2\right] dt \\
&= \sum_{k=0}^{N\ 1}\int_{t_k}^{t_{k+1}} (t)\mathbb{E}\left[\|Y_t - Y_{t_k}\|^2\right] dt \\
&\quad 2\sum_{k=0}^{N\ 1}\int_{t_k}^{t_{k+1}} (t)\max\left(\frac{\int_{t_k}^{t_{k+1}} (s) ds}{4^{\ 2}}; 1\right)\left(\mathbb{E}\|Y_{t_{k+1}}\|^2 - \mathbb{E}\|Y_{t_k}\|^2\right) dt \\
&\quad 2\sum_{k=0}^{N\ 1}\max\left(\frac{\int_{t_k}^{t_{k+1}} (s) ds}{4^{\ 2}}; 1\right)\left(\mathbb{E}\|Y_{t_{k+1}}\|^2 - \mathbb{E}\|Y_{t_k}\|^2\right)\int_{t_k}^{t_{k+1}} (t) dt \\
&\quad 2\sum_{k=0}^{N\ 1}\max\left\{\frac{\int_{t_k}^{t_{k+1}} (s) ds^{\ 2}}{4^{\ 2}}; \int_{t_k}^{t_{k+1}} (s) ds\right\}\left(\mathbb{E}\|Y_{t_{k+1}}\|^2 - \mathbb{E}\|Y_{t_k}\|^2\right) \\
&\quad 2\max_{0\ k\ N\ 1}\left\{\max\left\{\frac{\int_{t_k}^{t_{k+1}} (s) ds^{\ 2}}{4^{\ 2}}; \int_{t_k}^{t_{k+1}} (s) ds\right\}\right\}\mathbb{E}\left[\left\| r \log p_{T\ t_{N\ 1}}(X_{t_{N\ 1}})\right\|^2\right]:
\end{aligned}$$

By H1, $t \mapsto \gamma(t)$ is increasing, so that $t \mapsto \sigma(t)$ is decreasing. Therefore, using that since $X_0 \sim p_T$, X_T and X have the same distribution, yields,

$$\begin{aligned}
&\sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \gamma(t) \mathbb{E}\left[\left\| \nabla \log p_{T-t}(X_t) - \nabla \log p_{T-t_k}(X_{t_k}) \right\|^2 \right] dt \\
&\qquad \leq 2 \max_{0 \leq k \leq N-1} \max\left(\frac{(t_{k+1} - t_k)^2 \gamma(t_k)^2}{4\sigma^2(t_k)} ; (t_{k+1} - t_k)\gamma(t_k) \right) \\
&\qquad\qquad \times \mathbb{E}\left[\left\| \nabla \log p_{T-t_{N-1}}(X_{t_{N-1}}) - \nabla \log p_{T-t_{N-1}}(X_{t_{N-1}}) \right\|^2 \right] \\
&\qquad \leq 2 \max_{0 \leq k \leq N-1} \max\left(\frac{h^2 \gamma^2(t_k)}{4\sigma^2(t_k)} ; h\gamma(t_k) \right) I(p_T \| Q_T \otimes \gamma_{d,1}) \\
&\qquad \leq 2h\gamma(0) \max\left(\frac{h\gamma(0)}{4\sigma^2} ; 1 \right) I(p_T \| Q_T \otimes \gamma_{d,1}) \\
&\qquad \leq 2h\gamma(T) \max\left(\frac{h\gamma(T)}{4\sigma^2} ; 1 \right) I(p_T \| Q_T \otimes \gamma_{d,1}) .
\end{aligned}$$

Finally, following the steps of the proof of Conforti et al. (2023, Lemma 2), we can consider the limit when ε goes to zero, under Assumption H2, concluding the proof.

□

B.2 Technical results

Lemma B.4. Assume that H1 and H2 hold. Let $(X_t)_{t \geq 0}$ be a weak solution to the forward process (1). Then, the stationary distribution of $(X_t)_{t \geq 0}$ is Gaussian with mean 0 and variance $\sigma^2 I_d$.

Proof. Consider the process

$$X_t = \exp\left(\frac{1}{2\sigma^2} \int_0^t \gamma(s) ds \right) X_t .$$

Itô's formula yields

$$X_t = \exp\left(\frac{1}{2\sigma^2} \int_0^t \gamma(s) ds \right) X_0 + \int_0^t \sqrt{\gamma(s)} \exp\left(\int_0^s \gamma(u)/(2\sigma^2) du \right) dB_s . \tag{17}$$

First, we have that

$$\lim_{t \to \infty} \exp\left(\frac{1}{2\sigma^2} \int_0^t \gamma(s) ds \right) X_0 = 0 .$$

Secondly, we have that the second term in the r.h.s. of (17), by property of the Wiener integral, is Gaussian with mean 0 and variance $\sigma_t^2 I_d$, where

$$\sigma_t^2 = \exp\left(\frac{1}{\sigma^2} \int_0^t \gamma(s) ds \right) \int_0^t \gamma(s) e^{-\int_0^s \gamma(u)/\sigma^2 du} ds = \sigma^2 \left(1 - \exp\left(-\frac{1}{\sigma^2} \int_0^t \gamma(s) ds \right) \right) .$$

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

Lemma B.5. 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 $i \in \{1,2\}$,

$$dX_t^{(i)} = b_i\left(t, (X_s^{(i)})_{s\in[0,T]}\right)dt + \sqrt{(T-t)}\,dB_t \tag{18}$$

admits a unique strong solution with $X_0^{(i)} \sim \mu_0^{(i)}$. 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

$$\mathbb{E}\left[\exp\left(\frac{1}{2}\int_0^T \frac{1}{(T-s)}\left\|b_1\left(s, X_u^{(1)}\right)_{u\in[0,s]} - b_2\left(s, X_u^{(1)}\right)_{u\in[0,s]}\right\|^2 ds\right)\right] < +\infty. \tag{19}$$

Then,

$$\begin{aligned}\mathrm{KL}\left(\mu_0^{(1)}P_T^{(1)} \,\middle\|\, \mu_0^{(2)}P_T^{(2)}\right) &\leq \mathrm{KL}\left(\mu_0^{(1)} \,\middle\|\, \mu_0^{(2)}\right) \\ &+ \frac{1}{2}\int_0^T \frac{1}{(T-t)}\mathbb{E}\left[\left\|b_1\left(t, X_u^{(1)}\right)_{u\in[0,s]} - b_2\left(t, X_u^{(1)}\right)_{u\in[0,s]}\right\|^2\right] dt. \end{aligned} \tag{20}$$

Proof. Consider the probability space $(\Omega, (\mathcal{F}_t)_{0\leq t\leq T}, \mathbb{P})$ and 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 $X_0^{(i)} \sim \mu_0^{(i)}$. De ne $u(t,\omega)$ as

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

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

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

From (19), the Novikov's condition is satis ed (Karatzas and Shreve, 2012, Chapter 3.5.D), thus the process $(M_t)_{0\leq t\leq T}$ is a martingale. Applying Girsanov theorem, $dB_t = dB_t + u(t,(X_s^{(1)})_{s\in[0,t]})dt$ is a Brownian motion under the measure $\mathbb{Q}$. Therefore,

$$dX_t^{(1)} = b_1\left(t, X_u^{(1)}\right)_{u\in[0,t]}dt + \sqrt{(T-t)}\,dB_t = b_2\left(t, X_u^{(1)}\right)_{u\in[0,t]}dt + \sqrt{(T-t)}\,dB_t.$$

Using the uniqueness in law of (18), the law of $X^{(1)}$ under $\mathbb{P}$ is the same as the one of $X^{(2)}$ under $\mathbb{Q}$, with $X^{(2)}$ solution of (18) with $i = 2$ and $X_0^{(2)} = \mu_0^{(1)}$. Denote by $\mu^{(2)}$ the law of $X^{(2)}$. Therefore,

$$\mu^{(1)}(A) = \mathbb{P}(X^{(1)} \in A) = \mathbb{Q}(X^{(2)} \in A) = \int 1_A(X^{(2)}(\omega))\mathbb{Q}(d\omega);$$

which implies that

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

Hence, we obtain that

$$\begin{aligned}
\mathrm{KL}\left(\mathbb{P}^{(1)} \,\|\, \mathbb{P}^{(2)}\right) &= \mathrm{KL}\left(\mathbb{P}^{(1)}_0 \,\|\, \mathbb{P}^{(2)}_0\right) + \mathbb{E}\left[\log \frac{\mathrm{d}\mathbb{P}^{(1)}}{\mathrm{d}\mathbb{P}^{(2)}}\right] \\
&= \mathrm{KL}\left(\mathbb{P}^{(1)}_0 \,\|\, \mathbb{P}^{(2)}_0\right) + \mathbb{E}\left[\int_0^t u(s;\omega)^\top \mathrm{d}B_s + \frac{1}{2}\int_0^t \|u(s;\omega)\|^2 \mathrm{d}s\right] \\
&= \mathrm{KL}\left(\mathbb{P}^{(1)}_0 \,\|\, \mathbb{P}^{(2)}_0\right) + \frac{1}{2}\int_0^T \frac{1}{(T-t)}\mathbb{E}\left[\left\|b_1\left(t;(X^{(1)}_s)_{s\in[0;t]}\right) - b_2\left(t;(X^{(1)}_s)_{s\in[0;t]}\right)\right\|^2\right] \mathrm{d}t\,;
\end{aligned}$$

which concludes the proof. □

Lemma B.6. Let p be a probability density function on $\mathbb{R}^d$. For all $\sigma^2 > 0$,

$$\mathrm{KL}\left(p \,\|\, \mathcal{N}(0, \sigma^2 I_d)\right) = \int p(x) \log\left(\frac{p(x)}{\mathcal{N}(0,\sigma^2 I_d)(x)}\right) \mathrm{d}x \le \frac{\sigma^2}{2}\int \left\|\nabla \log\left(\frac{p(x)}{\mathcal{N}(0,\sigma^2 I_d)(x)}\right)\right\|^2 p(x)\,\mathrm{d}x.$$

Proof. Define $f_{\sigma^2} : x \mapsto p(x)/\mathcal{N}(0,\sigma^2 I_d)(x)$. Since $-\nabla^2 \log \mathcal{N}(0,\sigma^2 I_d)(x) = \sigma^{-2} I_d$, the Bakry-Emery criterion is satisfied with constant σ^{-2}, see Bakry et al. (2014); Villani (2021); Talagrand (1996). By the classical logarithmic Sobolev inequality,

$$\int f_{\sigma^2}(x) \log f_{\sigma^2}(x)\, \mathcal{N}(0,\sigma^2 I_d)(x)\mathrm{d}x \le \frac{\sigma^2}{2}\int \frac{\|\nabla f_{\sigma^2}(x)\|^2}{f_{\sigma^2}(x)}\, \mathcal{N}(0,\sigma^2 I_d)(x)\mathrm{d}x\,;$$

which concludes the proof. □

Lemma B.7. Define $Y_t := -\log p_{T-t}(X_t)$ and $Z_t := -\nabla \log p_{T-t}(X_t)$, where $\{X_t\}_{t\ge 0}$ is a weak solution to (10). Then,

$$\mathrm{d}Y_t = \frac{g^2(t)}{2}\left(\Delta \log p_{T-t}(X_t)\right)\mathrm{d}t - \frac{2}{\sigma^2}\frac{g^2(t)}{2\sigma^2}\left(\nabla \log p_{T-t}(X_t)\right)^\top \mathrm{d}X_t + g(t)\|Z_t\|\mathrm{d}B_t\,. \tag{21}$$

Proof. The Fokker-Planck equation associated with the forward process (10) is

$$\partial_t p_t(x) = \tfrac{1}{2}g(t)\,\mathrm{div}\left(\nabla p_t(x)\right) + \frac{g^2(t)}{2}\Delta p_t(x)\,; \tag{22}$$

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

$$\begin{aligned}
\partial \log p_t(x) = \mathrm{d}\left(\tfrac{1}{2}g(t)\right)\frac{g^2(t)}{2\sigma^2} + \left\langle \nabla \log p_t(x), x\right\rangle \left(\tfrac{1}{2}g(t)\right)\frac{g^2(t)}{2\sigma^2} + \frac{\|x\|^2}{2}\left(\frac{g^2(t)}{2\sigma^2}\right)'(t) \\
+ \frac{g^2(t)}{2}\frac{\Delta p_t(x)}{p_t(x)}\,.
\end{aligned} \tag{23}$$

Using that $\nabla \log \mathcal{N}(0,\sigma^2)(x) = -x/\sigma^2$, we have

$$\begin{aligned}
\mathrm{div}\left(\nabla p_t(x)\right) &= \mathrm{d}\,\Delta p_t(x) + p_t(x)\,\|\nabla \log p_t(x)\|^2 \\
&= \mathcal{N}(0,\sigma^2)(x)\left[\mathrm{d}\,\Delta p_t(x) + p_t(x)\left\|\nabla \log p_t(x) - \frac{x}{\sigma^2}\right\|^2\right] \\
&= \mathcal{N}(0,\sigma^2)(x)\left[\mathrm{d}\,\Delta p_t(x) + \|\nabla p_t(x)\|^2 - \frac{\|x\|^2}{\sigma^2}p_t(x)\right]\,:
\end{aligned}$$

Then, since $\nabla \varphi_2(x) = (\varphi_2(x)/\sigma^2)\, x$ and $\|x\|^2/\sigma^2\, d$, we get

$$\begin{aligned}\nabla p_t(x) &= \nabla \tilde{p}_t(x)\, \varphi_2(x) + 2\nabla \tilde{p}_t(x)^\top \nabla \varphi_2(x) + \varphi_2(x)\, \tilde{p}_t(x) \\ &= \varphi_2(x)\left(\frac{\nabla \tilde{p}_t(x)}{2} \frac{x\|x\|^2}{2}\, d \frac{2}{2}\nabla \tilde{p}_t(x)^\top x + \tilde{p}_t(x)\right).\end{aligned}$$

Combining these results with (22), we obtain

$$\begin{aligned}\partial_t \nabla \tilde{p}_t(x) = d\, \nabla \tilde{p}_t(x) \quad (t) \quad \frac{g^2(t)}{2^2} + \nabla \tilde{p}_t(x)^\top x \quad (t) \quad \frac{g^2(t)}{2} \\ + \nabla \tilde{p}_t(x) \frac{x\|x\|^2}{2} \quad \frac{g^2(t)}{2^2} \quad (t) + \frac{g^2(t)}{2} \tilde{p}_t(x).\end{aligned}$$

Hence, diving by $\tilde{p}_t$ yields (23).

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

$$\partial\ _t(x) = d \quad (t) \quad \frac{g^2(t)}{2^2} \quad \nabla\ _t(x)^\top x \quad (t) \quad \frac{g^2(t)}{2} \tag{24}$$

$$\frac{x\|x\|^2}{2} \quad \frac{g^2(t)}{2^2} \quad (t) \quad \frac{g^2(t)}{2} \quad _t(x) + \|\nabla\ _t(x)\|^2 : \tag{25}$$

Following the lines of Conforti et al. (2023, Proposition 1), we get that, since and g are continuous and non-increasing, the map p_t, solution to (22), belongs to $C^{1,2}((0,T] \times \mathbb{R}^d)$. By (11), as $Y_t = \nabla\ _t(X_t)$, we can apply Itô's formula and obtain, writing $(t) = (t)\, g(t)^2/^2$,

$$\begin{aligned}dY_t = \partial \nabla\ _t X_t + \nabla^2\ _t X_t \quad (t) X_t + g^2(t) \nabla\ _t X_t + \frac{g^2(t)}{2} \nabla\ _t X_t \, dt \\ + g(t) \nabla^2\ _t X_t \, dB_t \\ = \nabla \partial\ _t X_t + \frac{g^2(t)}{2} \quad _t X_t + \nabla\ _t X_t \quad ^2 + (t) \nabla^2\ _t X_t X_t \, dt \\ + g(t) \nabla^2\ _t X_t \, dB_t ;\end{aligned}$$

using that $2\nabla^2\ _t(x)\nabla\ _t(x) = \nabla\|\nabla\ _t(x)\|^2$. Using (24), we get

$$\begin{aligned}dY_t = \quad (t) \nabla\ _t X_t + \frac{2}{2} \quad (t) \quad \frac{g^2(t)}{2^2} X_t + (t) \nabla^2\ _t X_t X_t \, dt \\ + g(t) \nabla^2\ _t X_t \, dB_t ;\end{aligned}$$

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

$$\begin{aligned}dY_t &= \frac{g^2(t)}{2} \quad (t) \quad \nabla\ _t X_t + \frac{2}{2} \quad (t) \quad \frac{g^2(t)}{2^2} X_t \, dt + g(t) \nabla^2\ _t X_t \, dB_t \\ &= \frac{g^2(t)}{2} \quad (t) \quad Y_t + \frac{2}{2} \quad (t) \quad \frac{g^2(t)}{2^2} X_t \, dt + g(t) Z_t \, dB_t ;\end{aligned}$$

which concludes the proof. □

Lemma B.8. Let $Y_t := \nabla \log p_{T-t}(X_t)$, with X satisfying (11). There exists a constant $C > 0$ such that

$$\mathbb{E}\left[\|Y_t\|^4\right] \leq C\left(\bar{\alpha}_T^4 \sigma_t^4 \mathbb{E}\left[\|N\|^4\right] + \bar{\alpha}_T^8 \mathbb{E}\left[\|X_0\|^4\right]\right), \tag{26}$$

with $N \sim \mathcal{N}(0, I_d)$ and σ_t^2 as in (12).

Proof. The transition density $q_t(y; x)$ associated with the semi-group of the process (10) is given by

$$q_t(y; x) = (2\sigma_t^2)^{-d/2} \exp\left(-\frac{\|x - y \exp\left(\int_0^t \beta(s)ds\right)\|^2}{2\sigma_t^2}\right).$$

Therefore, we have

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

This, together with the definition of p_t, yields

$$\nabla \log p_{T-t}(X_{T-t}) = \bar{\alpha}_T^2 \sigma_t^2 \mathbb{E}\left[X_0 e^{\int_0^{T-t} \beta(u)du} \middle| X_{T-t}, X_{T-t}\right] + \sigma_t^2 X_{T-t}.$$

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

$$\begin{aligned}
\|\nabla \log p_{T-t}(X_{T-t})\|^4 &\leq C\left(\bar{\alpha}_T^8 \sigma_t^8 \left\|\mathbb{E}\left[X_0 e^{\int_0^{T-t} \beta(s)ds} \middle| X_{T-t}, X_{T-t}\right]\right\|^4 + \sigma_t^8 \|X_{T-t}\|^4\right) \\
&\leq C\left(\bar{\alpha}_T^8 \sigma_t^8 \mathbb{E}\left[\left\|X_0 e^{\int_0^{T-t} \beta(s)ds}\right\|^4 \middle| X_{T-t}, X_{T-t}\right] + \sigma_t^8 \|X_{T-t}\|^4\right).
\end{aligned}$$

Note that X_t has the same law as $\exp\left(\int_0^t \beta(s)ds\right) X_0 + \sigma_t N$, with $N \sim \mathcal{N}(0, I_d)$. This means that we have that

$$\mathbb{E}\left[\|\nabla \log p_{T-t}(X_{T-t})\|^4\right] \leq C\left(\bar{\alpha}_T^4 \sigma_t^4 \mathbb{E}\left[\|N\|^4\right] + \mathbb{E}\left[\|X_0\|^4\right]\right).$$

Finally,

$$\begin{aligned}
\mathbb{E}\left[\|Y_t\|^4\right] = \mathbb{E}\left[\|\nabla \log p_{T-t}(X_t)\|^2\right] &= \mathbb{E}\left[\|\nabla \log p_{T-t}(X_{T-t})\|^4\right] \\
&\leq \bar{\alpha}_T^4 \sigma_t^4 \mathbb{E}\left[\|N\|^4\right] \\
&\leq C\left(\bar{\alpha}_T^4 \sigma_t^4 \mathbb{E}\left[\|N\|^4\right] + \bar{\alpha}_T^8 \mathbb{E}\left[\|X_0\|^4\right]\right),
\end{aligned}$$

which concludes the proof. □

Lemma B.9. Let $Z_t := \nabla^2 \log p_{T-t}(X_t)$, where $\{X_t\}_{t \geq 0}$ is a weak solution to (11). There exists a constant $C > 0$ such that

$$\mathbb{E}\left[\|Z_t\|^4\right] \leq C\left(\alpha_T^{-8} \sigma_t^8 + \alpha_T^{-8}\, \mathbb{E}\left[\|Z\|_2^8\right] + d^4\right); \tag{27}$$

with $Z \sim N(0, I_d)$ and σ_t^2 as in (12).

Proof. Let $q_t(y; x)$ be the transition density associated to the semi-group of the process (10). Write

$$\begin{aligned}
&\nabla^2 \log p_{T-t}(x) \\
&= \nabla\left(\frac{1}{p_{T-t}(x)} \int p_0(y) \frac{y e^{-\int_0^{T-t} (s) ds} - x}{\alpha_{T-t}^2} q_{T-t}(y; x) dy\right) \\
&= \frac{\nabla p_{T-t}(x)}{p_{T-t}^2(x)} \left(\int p_0(y) \frac{y e^{-\int_0^{T-t} (s) ds} - x}{\alpha_{T-t}^2} q_{T-t}(y; x) dy\right)^{\top} \\
&\quad + \frac{1}{p_{T-t}(x)} \nabla \int p_0(y) \frac{y e^{-\int_0^{T-t} (s) ds} - x}{\alpha_{T-t}^2} q_{T-t}(y; x) dy \\
&= \frac{1}{\alpha_{T-t}^2 p_{T-t}(x)} \int \frac{\nabla p_{T-t}(x)}{p_{T-t}(x)} \left(y e^{-\int_0^{T-t} (s) ds} - x\right)^{\top} q_{T-t}(y; x) p_0(y) dy \\
&\quad - I_d + \frac{1}{\alpha_{T-t}^2} \int y e^{-\int_0^{T-t} (s) ds} - x \quad y e^{-\int_0^{T-t} (s) ds} - x \quad ^{\top} q_{T-t}(y; x) p_0(y) dy \; .
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\nabla^2 \log p_{T-t}\left(X_{T-t}\right) \\
&= \frac{1}{\alpha_{T-t}^2} \left(\mathbb{E}\left[\frac{\nabla p_{T-t}\left(X_{T-t}\right)}{p_{T-t}\left(X_{T-t}\right)} \left(X_0 e^{-\int_0^{T-t} (s) ds} - X_{T-t}\right)^{\top}\right] - \left(X_{T-t} - \alpha_{T-t} X_0\right)^{\top} + I_d\right) \\
&\quad + \alpha_{T-t}^4 \, \mathbb{E}\left[\left(X_0 e^{-\int_0^{T-t} (s) ds} - X_{T-t}\right) \left(X_0 e^{-\int_0^{T-t} (s) ds} - X_{T-t}\right)^{\top}\right] + \alpha^2 I_d \\
&= \alpha_{T-t}^4 \, \mathbb{E}\left[X_0 e^{-\int_0^{T-t} (s) ds} - X_{T-t} \mid X_{T-t}\right] \mathbb{E}\left[X_0 e^{-\int_0^{T-t} (s) ds} - X_{T-t} \mid X_{T-t}\right]^{\top} \\
&\quad + \alpha^2 - \alpha_{T-t}^2 I_d \\
&\quad + \alpha_{T-t}^4 \, \mathbb{E}\left[\left(X_0 e^{-\int_0^{T-t} (s) ds} - X_{T-t}\right) \left(X_0 e^{-\int_0^{T-t} (s) ds} - X_{T-t}\right)^{\top} \mid X_{T-t}\right] \; .
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}\left[\left\| r^2 \log p_{T-t}(\overleftarrow{X}_{T-t}) \right\|_{\mathrm{Fr}}^4\right] \\
&\quad \le \frac{C}{(T-t)^{16}} \mathbb{E}\Big[\mathbb{E}\big[\overleftarrow{X}_0 e^{\int_0^{T-t} (s) ds} \overleftarrow{X}_{T-t} \,\big|\, \overleftarrow{X}_{T-t}\big] - \mathbb{E}\big[\overleftarrow{X}_0 e^{\int_0^{T-t} (s) ds} \overleftarrow{X}_{T-t} \,\big|\, \overleftarrow{X}_{T-t}\big]\Big\|_{\mathrm{Fr}}^4\Big] \\
&\qquad + C (T-t)^8 + {}^2 \\
&\qquad + \frac{C}{(T-t)^{16}} \mathbb{E}\Big[\Big\| \mathbb{E}\big[\overleftarrow{X}_0 e^{\int_0^{T-t} (s) ds} \overleftarrow{X}_{T-t} \,\big|\, \overleftarrow{X}_{T-t}\big] \mathbb{E}\big[\overleftarrow{X}_0 e^{\int_0^{T-t} (s) ds} \overleftarrow{X}_{T-t} \,\big|\, \overleftarrow{X}_{T-t}\big] \Big\|_{\mathrm{Fr}}^4\Big] .
\end{aligned}$$

As in the previous proof, we note that $\overleftarrow{X}_t$ has the same law as $e^{\int_0^t (s) ds} X_0 + {}_t Z$, with $Z \sim \mathcal{N}(0; I_d)$ independent of X_0. Therefore, using Jensen's inequality,

$$\begin{aligned}
&\mathbb{E}\Big[\mathbb{E}\big[\overleftarrow{X}_0 e^{\int_0^{T-t} (s) ds} \overleftarrow{X}_{T-t} \,\big|\, \overleftarrow{X}_{T-t}\big] \mathbb{E}\big[\overleftarrow{X}_0 e^{\int_0^{T-t} (s) ds} \overleftarrow{X}_{T-t} \,\big|\, \overleftarrow{X}_{T-t}\big]^{\top}\Big] \\
&\quad \mathbb{E}\Big[\mathbb{E}\big[\overleftarrow{X}_0 e^{\int_0^{T-t} (s) ds} \overleftarrow{X}_{T-t} \,\big|\, \overleftarrow{X}_{T-t}\big]^4 \mathbb{E}\big[\overleftarrow{X}_0 e^{\int_0^{T-t} (s) ds} \overleftarrow{X}_{T-t} \,\big|\, \overleftarrow{X}_{T-t}\big]^4\Big]^{\frac{1}{2}} \\
&\quad \mathbb{E}\Big[\mathbb{E}\big[\overleftarrow{X}_0 e^{\int_0^{T-t} (s) ds} \overleftarrow{X}_{T-t} \,\big|\, \overleftarrow{X}_{T-t}\big]^8\Big]^{\frac{1}{2}} \\
&\quad {}_t^8 \mathbb{E}\left[\|Z\|_2^8\right]^{\frac{1}{2}}
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\Big[\Big\| \mathbb{E}\big[\overleftarrow{X}_0 e^{\int_0^{T-t} (s) ds} \overleftarrow{X}_{T-t} \,\big|\, \overleftarrow{X}_{T-t}\big] \mathbb{E}\big[\overleftarrow{X}_0 e^{\int_0^{T-t} (s) ds} \overleftarrow{X}_{T-t} \,\big|\, \overleftarrow{X}_{T-t}\big] \Big\|_{\mathrm{Fr}}^4\Big] \\
&\quad \mathbb{E}\Big[\Big\| \mathbb{E}\big[\overleftarrow{X}_0 e^{\int_0^{T-t} (s) ds} \overleftarrow{X}_{T-t} \,\big|\, \overleftarrow{X}_{T-t}\big] \Big\|_2^4 \Big\| \mathbb{E}\big[\overleftarrow{X}_0 e^{\int_0^{T-t} (s) ds} \overleftarrow{X}_{T-t} \,\big|\, \overleftarrow{X}_{T-t}\big] \Big\|_2^4\Big] \\
&\quad = \mathbb{E}\Big[\Big\| \mathbb{E}\big[\overleftarrow{X}_0 e^{\int_0^{T-t} (s) ds} \overleftarrow{X}_{T-t} \,\big|\, \overleftarrow{X}_{T-t}\big] \Big\|_2^8\Big] \\
&\quad {}_t^8 \mathbb{E}\left[\|Z\|_2^8\right] .
\end{aligned}$$

Hence, we can conclude that

$$\mathbb{E}\left[\|Z_t\|_{\mathrm{Fr}}^4\right] = \mathbb{E}\left[\left\| r^2 \log p_{T-t}(\overleftarrow{X}_{T-t}) \right\|_{\mathrm{Fr}}^4\right] \quad C (T-t)^8 + {}^8 \mathbb{E}\left[\|Z\|_2^8\right] + d^4 .$$

□

C Proofs of Section 4

C.1 Gaussian case: proof of Lemma 4.1

In the case where $_{data}$ is the Gaussian probability density with mean $_0$ and variance $_0$, we have

$$r \log p_t(x) = \quad m_t^2 \ _0 + \ _t^2 I_d \quad ^1 (x \quad m_t \ _0) + \quad ^2 x ;$$

with $m_t = \exp \quad \int_0^t (s)ds=(2 \ ^2)$ and $_t = \ ^2(1 \quad m_t^2)$. Let $_t = m_t^2 \ _0 + \ _t^2 I_d$ be the covariance of the forward process X_t and $b_t = \ _t \ ^1 m_t \ _0$ so that

$$r \log p_t(x) = A_t x + b_t \quad \text{with} \quad A_t = \quad _t \ ^1 \quad \ ^2 I_d \ : \tag{28}$$

Note that, if we denote by $_0^1 \quad _0^d$ the eigenvalues of $_0$, which are positive as $_0$ is positive de nite, we have that the eigenvalues of A_t are

$$_t^i := \quad \frac{1}{m_t^2 \ _0^i + \ _t^2} + \frac{1}{2} :$$

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

$$(r \log p_t(x) \quad r \log p_t(y))^> (x \quad y) \quad _t^d kx \quad yk^2 ;$$
$$kr \log p_t(x) \quad r \log p_t(y)k \quad \max \quad _t^1 ; \ _t^d \quad kx \quad yk ;$$

which entails that we can de ne

$$L_t := \max \quad _t^1 ; \ _t^d \quad ; \qquad C_t := \quad _t^d ;$$

and apply Proposition C.2.

The condition $_t^d \quad 0$, or equivalently $^2 \quad _{max}(\ _0)$, yields a contraction in 2 Wasserstein distance in the backward process as well in the forward process from Proposition C.2. This shows that, in speci c 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 ows.

As a consequence, note that

$$W_2(\ _{data} ; \ _1 Q_T)^2 \quad W_{\ 2}(p_T ; \ _1)^2 \exp \quad \frac{1}{\ ^2} \int_0^T \quad (t)(1 + 2 C_t \ ^2)dt \quad :$$

Using Talagrand's T_2 inequality for the Gaussian measure $W_2(\ ; \ _1)^2 \quad 2 \ ^2 KL(\ k \ _1)$ and Lemma B.1 we get

$$W_2(\ _{data} ; \ _1 Q_T)^2 \quad 2 \ ^2 KL(\ _{data} k \ _1) \exp \quad \frac{2}{\ ^2} \int_0^T \quad (t)(1 + 2 C_t \ ^2)dt \quad :$$

Proposition C.1. Assume that $_{data}$ is a Gaussian distribution $N(\ _0; \ _0)$ such that $_{max}(\ _0) \quad ^2$ where $_{max}(\ _0)$ denotes the largest eigenvalue of $_0$. Then,

$$KL(\ _{data} k \ _1 Q_T) \quad KL(\ _{data} k' \ _2) \exp \quad \frac{2}{\ ^2} \int_0^T \quad (s)ds \quad :$$

Proof. In this Gaussian case, the backward process is linear (see (28)) and the associated in nitesimal generator writes, for $g \in C^2$,

$$L_t g(x) = r\, g(x)^\top \left(\frac{(t)}{2^{\ 2}} + \quad (t)(A_t x + b_t) \right) + \frac{1}{2} \ (t) \quad g(x);$$

where $A_t = A_{T\ \ t}$ and $b_t = b_{T\ \ t}$.

Our objective is to monitor the evolution of the Kullback-Leibler divergence, $KL(p_T Q_t \| '_{\ 2} Q_t)$, 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 $_t = ' _{\ 2} Q_t$ two densities that satisfy the Fokker-Planck equation, involving the dual operator L_t of the in nitesimal generator L

$$\begin{aligned} @q_t &= L_t q_t; \qquad q_0(x) = p_T(x) \\ @\ _t &= L_t\ _t; \qquad _0(x) = ' _{\ 2}(x): \end{aligned}$$

Let $f_t = q_t / _t$. By de nition of $KL(q_t \| _t) = \int \ln(f_t(x))\, q_t(x)dx$ we have

$$\begin{aligned} @KL(q_t \| _t) &= \int \ln(f_t(x))\, @q_t(x)dx + \int @\ln(f_t(x))\, q_t(x)dx \\ &= \int \ln(f_t(x))\, @q_t(x)dx \quad \int f_t(x) @\ _t(x)dx: \end{aligned}$$

By employing the Fokker-Planck equation and the adjoint relation, which states that $\int f(x) L_t(g)(x)dx = \int L_t f(x) g(x)dx$ we obtain

$$@KL(q_t \| _t) = \int L \ln(f_t)(x) q_t(x)dx \quad \int L f_t(x)\ _t(x)dx:$$

The in nitesimal generator L satis es the change of variables formula (see Bakry et al., 2014) so that

$$L_t(\ln(f)) = \frac{1}{f} L_t f \quad \frac{1}{2f^{\ 2}} \ _t(f;f);$$

where $_t$ is the carré du champ operator associated with L_t de ned by $_t(f;f)(x) = (t)|r f(x)|^2$. We then obtain

$$\begin{aligned} @KL(q_t \| _t) &= \int L f_t(x) \frac{q_t(x)}{f_t(x)} dx \quad \int \frac{(t)}{2} \frac{|r f_t(x)|^2}{f_t^2(x)} q_t(x)dx \quad \int L f_t(x)\ _t(x)dx \\ &= \quad \frac{(t)}{2} \int \frac{|r f_t(x)|^2}{f_t(x)}\ _t(x)dx: \end{aligned} \tag{29}$$

To obtain a control of the Kullback-Leibler divergence we need a logarithmic Sobolev inequality for the distribution of density $_t = ' _{\ 2} Q_t$. In this Gaussian case, if $X_0 \quad N(0; \ ^2)$ then for all $t \in [0;T]$ the law of X_t is a centered Gaussian with covariance matrix $_t$ given by

$$_t = \ ^2 \exp\left(\int_0^t \frac{(s)}{2} + 2\ _s A_s ds \right) + \int_0^t (s) \exp\left(\int_s^t \frac{(u)}{2} + 2\ (u) A_u du \right) ds;$$

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 $\lambda_{\max}(\Sigma_t) \leq \sigma^2$. We recall the logarithmic Sobolev inequality for a normal distribution (see Chafai, 2004, Corollary 9)

$$\mathrm{KL}(q_t \| \gamma_t) \leq \frac{1}{2} \int \frac{1}{f_t(x)} \| \nabla f_t(x) \|^2 \gamma_t(x) dx \leq \frac{\lambda_{\max}(\Sigma_t)}{2} \int \frac{|\nabla f_t(x)|^2}{f_t(x)} \gamma_t(x) dx .$$

Plugging this into (29) we get

$$\partial_t \mathrm{KL}(q_t \| \gamma_t) \leq -\frac{\gamma(t)}{2} \mathrm{KL}(q_t \| \gamma_t) .$$

Therefore, recalling that $q_0 = p_T$ and $\gamma_0 = \mathcal{N}(0, \sigma^2 I_d)$

$$\mathrm{KL}(q_T \| \mathcal{N}(0, \sigma^2 Q_T)) \leq \mathrm{KL}(p_T \| \mathcal{N}(0, \sigma^2)) \exp\left(- \int_0^T \frac{\gamma(s)}{2} ds \right) .$$

We conclude using Lemma B.1. □

C.2 Proof of Theorem 4.2

EI scheme. Using the fact that

$$\int_{t_k}^{t} e^{\int_s^t (\nu)/(2\sigma^2) d\nu} \gamma(s) ds = 2\sigma^2 \left(1 - e^{\int_{t_k}^t (\nu)/(2\sigma^2) d\nu} \right) ,$$

the Exponential Integrator scheme that we consider consists in the following discretization, recursively given with respect to the index k,

$$X_t = e^{\int_{t_k}^t (s)/(2\sigma^2) ds} X_{t_k} + 2\sigma^2 \left(1 - e^{\int_{t_k}^t (s)/(2\sigma^2) ds} \right) \nabla \log p_{T-t_k}(X_{t_k}) + \sqrt{1 - e^{\int_{t_k}^t (s)/\sigma^2 ds}} Z_k , \quad (30)$$

where Z_k are i.i.d. Gaussian random vectors $\mathcal{N}(0, I_d)$. In particular, we have that

$$X_t = e^{\int_{t_k}^t (s)/(2\sigma^2) ds} X_{t_k} + 2\sigma^2 \left(1 - e^{\int_{t_k}^t (s)/(2\sigma^2) ds} \right) s_\theta(T - t_k, X_{t_k}) + \sqrt{1 - e^{\int_{t_k}^t (s)/\sigma^2 ds}} Z_k , \quad (31)$$

and $X_0 \sim \mathcal{N}(0, \sigma^2 I_d)$. Note that

$$W_2(\hat{\gamma}_{\mathrm{data}}, \hat{b}_N^{(\cdot,\cdot)}) \leq W_2(\hat{\gamma}_{\mathrm{data}}, \gamma_1 Q_T) + W_2(\gamma_1 Q_T, \gamma_1 Q_T^{N,\cdot}) , \quad (32)$$

where

$$W_2(\hat{\gamma}_{\mathrm{data}}, \gamma_1 Q_T) = W_2(p_T Q_T, \gamma_1 Q_T) ,$$

which corresponds to the discrepancy between the same process (3) with two different initializations. The first term of (32) is upper bounded by Proposition C.2.

Proposition C.2. Assume that $W_2(\ _{data}\ ;\ _1\ )^2 < +1$. The marginal distribution at the end of the forward phase satis es

$$W_2(p_T\ ;\ _1\ )^2 \quad W_{\ 2}(\ _{data}\ ;\ _1\ )^2 \exp\left(\int_0^T \frac{(t)}{2}dt\right) : \tag{33}$$

Assume that H4(ii) holds. Then,

$$W_2(\ _{data}\ ;\ _1\ Q_T)^2 \quad W_{\ 2}(p_T\ ;\ _1\ )^2 \exp\left(\int_0^T \frac{(t)}{2}\ 1 \quad 2L_t\ ^2\ dt\right)$$
$$W_{\ 2}(\ _{data}\ ;\ _1\ )^2 \exp\left(\int_0^T \frac{(t)}{2}\ 2 \quad 2L_t\ ^2\ dt\right) : \tag{34}$$

Moreover, under Assumption H4(i) , we have

$$W_2(\ _{data}\ ;\ _1\ Q_T)^2 \quad W_{\ 2}(p_T\ ;\ _1\ )^2 \exp\left(\int_0^T \frac{(t)}{2}\ 1 + 2C_t\ ^2\ dt\right)$$
$$W_{\ 2}(\ _{data}\ ;\ _1\ )^2 \exp\left(\int_0^T \frac{(t)}{2}\ 2 + 2C_t\ ^2\ dt\right) : \tag{35}$$

Proof of Proposition C.2. Let $x \ 2\ R^d$ (resp. $y \ 2\ R^d$) and denote by $\overleftarrow{X}^x$ (resp. $\overleftarrow{X}^y$) the solution of (1), with initial condition $\overleftarrow{X}_0^x = x$ (resp. $\overleftarrow{X}_0^x = y$). Applying the chain rule, we get

$$\overleftarrow{X}_t^x \quad \overleftarrow{X}_t^y\ ^2 = kx \quad yk^2 + 2\int_0^t \frac{(s)}{2\ ^2}\ \overleftarrow{X}_s^x \quad \overleftarrow{X}_s^y\ ^2 ds :$$

Therefore, applying Grönwall's lemma, we obtain

$$E\left[\sup_{t\ 2\ [0;T\ ]}\ \overleftarrow{X}_t^x \quad \overleftarrow{X}_t^y\ ^2\right] \quad \exp\left(\int_0^T \frac{(t)}{2}dt\right) \quad kx \quad yk^2 :$$

From this, we can show contraction (33) in 2 Wasserstein distance by taking the in mum over all couplings.

Now, let $x \ 2\ R^d$ (resp. $y \ 2\ R^d$) and denote by X^x (resp. X^y) the solution of (3), with initial condition $X_0^x = x$ (resp. $X_0^x = y$). Applying the chain rule and using Cauchy-Schwarz inequality, we get

$$X_t^x \quad X_t^y\ ^2 = kx \quad yk^2 + 2\int_0^t \frac{(s)}{2\ ^2}\ X_s^x \quad X_s^y\ ^2 ds$$
$$+ 2\int_0^t (s)\ r\ \log p_{T\ s}\ X_s^x \quad r\ \log p_{T\ s}\ X_s^y\ ^{>}\ X_s^x \quad X_s^y\ ds$$
$$k\ x \quad yk^2 \int_0^t \frac{(s)}{2}\ 1 \quad 2L_s\ ^2\ X_s^x \quad X_s^y\ ^2 ds :$$

Therefore, applying Grönwall's lemma, we obtain

$$\mathbb{E}\Big[\sup_{t\in[0;T]} \big\|X_t^x - X_t^y\big\|^2\Big] \le \exp\Big(\int_0^T \frac{\gamma(t)}{2}\big(1 - 2L_t\big)^2\,dt\Big)\,\|x-y\|^2 .$$

From this, we can show contraction (34) in 2 Wasserstein distance by taking the in mum over all couplings.

To establish (35) note that, under Assumption H4(i), we have

$$\begin{aligned}\big\|X_t^x - X_t^y\big\|^2 &= \|x-y\|^2 + 2\int_0^t \frac{\gamma(s)}{2^2}\big\|X_s^x - X_s^y\big\|^2 ds \\ &\quad + 2\int_0^t \gamma(s)\big(\nabla \log p_{T-s}(X_s^x) - \nabla \log p_{T-s}(X_s^y)\big)^\top \big(X_s^x - X_s^y\big)\,ds \\ &\le \|x-y\|^2 \int_0^t \frac{\gamma(s)}{2}\big(1 + 2C_s\big)^2\big\|X_s^x - X_s^y\big\|^2 ds .\end{aligned}$$

Therefore, applying Grönwall's lemma, we obtain

$$\mathbb{E}\Big[\sup_{t\in[0;T]} \big\|X_t^x - X_t^y\big\|^2\Big] \le \exp\Big(\int_0^T \frac{\gamma(t)}{2}\big(1 + 2C_t\big)^2\,dt\Big)\,\|x-y\|^2 .$$

From this, we can show contraction (35) in the 2 Wasserstein distance by taking the in mum over all couplings. □

Note that a similar assumption as Assumption H4(i) is used in De Bortoli et al. (2021, Proposition 10,11,12), in particular to bound the conditional moments of X_0 given X_t for $t > 0$. However, in this paper the authors also require additional assumptions, in particular that the score of $_{\text{data}}$ has a linear growth.

Second term. The second term of (36) can be handled as follows

$$W_2\big(\,_1 Q_T;\,{}_1 Q_T^{N;}\big) \le \big\|X_T^{\frac{1}{T}} - X_T\big\|_{L_2} .$$

To upper bound $\|X_T^{\frac{1}{T}} - X_T\|_{L_2}$, we aim at controlling $\|X_{t_{k+1}}^{\frac{1}{t_{k+1}}} - X_{t_{k+1}}\|_{L_2}$ by $\|X_{t_k}^{\frac{1}{t_k}} - X_{t_k}\|_{L_2}$ to resort subsequently to a telescopic sum.

Proposition C.3. Assume that H4, H5 and H6 hold. Consider the regular discretization $\{t_k; 0 \le k \le N\}$ of $[0;T]$ of constant step size h such that for all t_k with $0 \le k \le N-1$,

$$h < \frac{2C_t}{\gamma(t_k)\,\max_{t_k \le s \le t_{k+1}} L_s\, L_t}\,\frac{\mathrm{e}^{\bar{m}_{t_{k+1}}}}{\mathrm{e}^{\bar{m}_{t_k}}} ;$$

where $\mathrm{e}^{\bar{m}_t} := \exp\big(\int_0^t \gamma(s)ds/(2^2)\big)$, $m_t := \exp\big(\int_0^t \gamma(s)ds/(2^2)\big)$. Then,

$$\begin{aligned}\big\|X_T^{\frac{1}{T}} - X_T\big\|_{L_2} &\le \varepsilon_T(T) + M h_T(T)(1+2B) \\ &\quad + \sum_{k=0}^{N-1}\Big(\int_{t_k}^{t_{k+1}} L_t\,\frac{\mathrm{e}^{\bar{m}_t}}{\mathrm{e}^{\bar{m}_{t_k}}}(t)dt\Big)\,\frac{\sqrt{2h\gamma(T)}}{} + m_T\int_{t_k}^{t_{k+1}}\Big(\frac{1}{2^2} + 2L_t\Big)(t)dt\Big)\,B ;\end{aligned}$$

where M is defined in H6 and $B := (\mathbb{E}[\|X_0\|^2] + \sigma^2 d)^{1/2}$.

Proof. Using (31) and the triangular inequality, we have

$$
\begin{aligned}
&\left\| X^1_{t_{k+1}} - X_{t_{k+1}} \right\|_{L_2} \\
&= \left\| \frac{\alpha_{t_{k+1}}}{\alpha_{t_k}} X^1_{t_k} - \frac{\alpha_{t_{k+1}}}{\alpha_{t_k}} X_{t_k} + \int_{t_k}^{t_{k+1}} \frac{\alpha_t}{\alpha_{t_k}} \left(\beta_T - t \right) \nabla \log p_{T-t}(X^1_t) - s_\theta(T - t_k, X_{t_k}) \, dt \right\|_{L_2} \\
&\quad \left\| \frac{\alpha_{t_{k+1}}}{\alpha_{t_k}} X^1_{t_k} - \frac{\alpha_{t_{k+1}}}{\alpha_{t_k}} X_{t_k} + \int_{t_k}^{t_{k+1}} \frac{\alpha_t}{\alpha_{t_k}} \left(\beta_T - t \right) \nabla \log p_{T-t}(X^1_t) - \nabla \log p_{T-t}(X_{t_k}) \, dt \right\|_{L_2} \\
&\quad + \int_{t_k}^{t_{k+1}} \frac{\alpha_t}{\alpha_{t_k}} \left(\beta_T - t \right) \left\| \nabla \log p_{T-t}(X^1_t) - \nabla \log p_{T-t}(X^1_{t_k}) \right\| dt \\
&\quad + \int_{t_k}^{t_{k+1}} \frac{\alpha_t}{\alpha_{t_k}} \left(\beta_T - t \right) \left\| \nabla \log p_{T-t}(X_{t_k}) - s_\theta(T - t_k, X_{t_k}) \right\| dt \, .
\end{aligned} \tag{36}
$$

Using the strong concavity and Lipschitz properties of the modified score function, we have that the first term of r.h.s. of (36) can be bounded as follows

$$
\begin{aligned}
&\left\| \frac{\alpha_{t_{k+1}}}{\alpha_{t_k}} X^1_{t_k} - \frac{\alpha_{t_{k+1}}}{\alpha_{t_k}} X_{t_k} + \int_{t_k}^{t_{k+1}} \frac{\alpha_t}{\alpha_{t_k}} \left(\beta_T - t \right) \nabla \log p_{T-t}(X^1_t) - \nabla \log p_{T-t}(X_{t_k}) \, dt \right\|^2 \\
&= \frac{\alpha^2_{t_{k+1}}}{\alpha^2_{t_k}} \left\| X^1_{t_k} - X_{t_k} \right\|^2 + \left\| \int_{t_k}^{t_{k+1}} \frac{\alpha_t}{\alpha_{t_k}} \left(\beta_T - t \right) \nabla \log p_{T-t}(X^1_t) - \nabla \log p_{T-t}(X_{t_k}) \, dt \right\|^2 \\
&\quad + \frac{\alpha_{t_{k+1}}}{\alpha_{t_k}} 2 \int_{t_k}^{t_{k+1}} \frac{\alpha_t}{\alpha_{t_k}} \left(\beta_T - t \right) \left\langle X^1_{t_k} - X_{t_k}, \nabla \log p_{T-t}(X^1_t) - \nabla \log p_{T-t}(X_{t_k}) \right\rangle dt \\
&\quad \left\| X^1_{t_k} - X_{t_k} \right\|^2 \frac{\alpha^2_{t_{k+1}}}{\alpha^2_{t_k}} + \int_{t_k}^{t_{k+1}} L_t \frac{\alpha_t}{\alpha_{t_k}} (\beta_T - t) \, dt \right)^2 - 2 \frac{\alpha_{t_{k+1}}}{\alpha_{t_k}} \int_{t_k}^{t_{k+1}} C_t \frac{\alpha_t}{\alpha_{t_k}} (\beta_T - t) \, dt \, .
\end{aligned}
$$

Using the Lipschitz property of the modified score and Proposition C.6, the second term of the r.h.s. of (36) can be controlled as follows

$$
\begin{aligned}
&\int_{t_k}^{t_{k+1}} \frac{\alpha_t}{\alpha_{t_k}} \left(\beta_T - t \right) \left\| \nabla \log p_{T-t}(X^1_t) - \nabla \log p_{T-t}(X^1_{t_k}) \right\| dt \\
&\quad \int_{t_k}^{t_{k+1}} L_{T-t} \frac{\alpha_t}{\alpha_{t_k}} (\beta_T - t) \, dt \sup_{t_k \le t \le t_{k+1}} \left\| X^1_t - X^1_{t_k} \right\|_{L_2} \\
&\quad \int_{t_k}^{t_{k+1}} L_t \frac{\alpha_t}{\alpha_{t_k}} (\beta_T - t) \, dt \left(\frac{1}{\sqrt{h}} \sqrt{\gamma(T)} + \exp\left(\int_0^{t_k} \frac{\gamma(s)}{2} \left(1 + C_s^2 \right) ds \right) B \right) .
\end{aligned}
$$

Using Assumption H5, we can control the third term of the r.h.s. of (36) as follows

$$\begin{aligned}
&\int_{t_k}^{t_{k+1}} \frac{\partial_t}{\partial_{t_k}}(t)\, r \log p_{T-t}\big(X_{t_k}\big) - s - T - t_k; X_{t_k}\, dt \Big\|_{L^2} \\
&\int_{t_k}^{t_{k+1}} \frac{\partial_t}{\partial_{t_k}}(t)\, r \log p_{T-t_k}\big(X_{t_k}\big) - s - T - t_k; X_{t_k}\, dt \Big\|_{L^2} \\
&+ \int_{t_k}^{t_{k+1}} \frac{\partial_t}{\partial_{t_k}}(t)\, r \log p_{T-t}\big(X_{t_k}\big) - r \log p_{T-t_k}\big(X_{t_k}\big)\, dt \Big\|_{L^2} \\
&\int_{t_k}^{t_{k+1}} \frac{\partial_t}{\partial_{t_k}}(t)dt + \int_{t_k}^{t_{k+1}} \frac{\partial_t}{\partial_{t_k}}(t)\, r \log p_{T-t}\big(X_{t_k}\big) - r \log p_{T-t_k}\big(X_{t_k}\big)\, dt \\
&\int_{t_k}^{t_{k+1}} \frac{\partial_t}{\partial_{t_k}}(t)dt + hM\Big(1 + \big\|X_{t_k}\big\|_{L^2}\Big) \int_{t_k}^{t_{k+1}} \frac{\partial_t}{\partial_{t_k}}(t)dt.
\end{aligned}$$

Note that X_t has the same law as $m_t X_0 + \sqrt{(1 - m_t^2)}G$, with G a standard Gaussian random variable independent of X_0. We have that $X_0^{\frac{1}{2}} \sim N(0; \sigma^2 I_d)$. Define $(X_t)_{t\in[0;T]}$ satisfying (3) but initialized at

$$X_0 = m_T X_0 + \sqrt{(1 - m_T^2)} X_0^{\frac{1}{2}};$$

with $X_0 \sim \text{data}$. Employing Proposition C.5 and (42), we obtain

$$\big\|X_{t_k}\big\|_{L^2} \le \big\|X_{t_k} - X_{t_k}^{\frac{1}{2}}\big\|_{L^2} + \big\|X_{t_k}^{\frac{1}{2}} - X_{t_k}\big\|_{L^2} + \big\|X_{t_k}\big\|_{L^2} \le \big\|X_{t_k} - X_{t_k}^{\frac{1}{2}}\big\|_{L^2} + 2B.$$

Therefore, combining the previous bounds, together with (36), we obtain

$$\begin{aligned}
&\big\|X_{t_{k+1}}^{\frac{1}{2}} - X_{t_{k+1}}\big\|_{L^2} \\
&\le \Big\|X_{t_k}^{\frac{1}{2}} - X_{t_k}\Big\| \frac{\partial_{t_{k+1}}^2}{\partial_{t_k}^2} + \int_{t_k}^{t_{k+1}} L_t \frac{\partial_t}{\partial_{t_k}}(t)dt\Big)^2 - 2\frac{\partial_{t_{k+1}}}{\partial_{t_k}} \int_{t_k}^{t_{k+1}} C_t \frac{\partial_t}{\partial_{t_k}}(t)dt\Big)^{1=2} \\
&+ \int_{t_k}^{t_{k+1}} L_t \frac{\partial_t}{\partial_{t_k}}(t)dt \cdot \frac{1}{\sqrt{2h\,(T)}} + m_T \int_{t_k}^{t_{k+1}} \frac{1}{2\,^2} + 2L_t\Big)(t)dt \cdot B \\
&+ \int_{t_k}^{t_{k+1}} \frac{\partial_t}{\partial_{t_k}}(t)dt + hM\Big(1 + \big\|X_{t_k}\big\|_{L^2}\Big) \int_{t_k}^{t_{k+1}} \frac{\partial_t}{\partial_{t_k}}(t)dt.
\end{aligned}$$

By the assumption on h and Proposition C.4,

$$0 < 1 + \frac{\partial_{t_k}^2}{\partial_{t_{k+1}}^2} \int_{t_k}^{t_{k+1}} L_t \frac{\partial_t}{\partial_{t_k}}(t)dt\Big)^2 - 2\frac{\partial_{t_k}}{\partial_{t_{k+1}}} \int_{t_k}^{t_{k+1}} C_t \frac{\partial_t}{\partial_{t_k}}(t)dt < 1;$$

and, using that $\sqrt{1-x} \leq 1 - x/2$ for $x \in [0,1]$, we conclude that

$$
\begin{aligned}
&\left\| X^1_{t_{k+1}} - X_{t_{k+1}} \right\|_{L_2} \\
&\quad \leq \left\| X^1_{t_k} - X_{t_k} - \frac{\mathrm{d}t_{k+1}}{\mathrm{d}t_k} \left(1 + \frac{1}{2} \frac{\mathrm{d}t_k}{\mathrm{d}t_{k+1}} \right) \int_{t_k}^{t_{k+1}} L_t \frac{\mathrm{d}t}{\mathrm{d}t_k}(t)\,\mathrm{d}t - \frac{\mathrm{d}t_k}{\mathrm{d}t_{k+1}} \int_{t_k}^{t_{k+1}} C_t \frac{\mathrm{d}t}{\mathrm{d}t_k}(t)\,\mathrm{d}t + \right. \\
&\qquad \left. + \frac{\mathrm{d}t_k}{\mathrm{d}t_{k+1}} Mh \int_{t_k}^{t_{k+1}} \frac{\mathrm{d}t}{\mathrm{d}t_k}(t)\,\mathrm{d}t \right\|_{L_2} \\
&\quad + \left\| \int_{t_k}^{t_{k+1}} L_t \frac{\mathrm{d}t}{\mathrm{d}t_k}(t)\,\mathrm{d}t - \frac{1}{\sqrt{2h}} \sqrt{\Sigma(T) + m_T} \int_{t_k}^{t_{k+1}} \frac{1}{2} + 2L_t \,\mathrm{d}t \right\|_B \\
&\quad + \left\| \int_{t_k}^{t_{k+1}} \frac{\mathrm{d}t}{\mathrm{d}t_k}(t)\,\mathrm{d}t + hM(1+2B) \int_{t_k}^{t_{k+1}} \frac{\mathrm{d}t}{\mathrm{d}t_k}(t)\,\mathrm{d}t \right\|_{L_2} .
\end{aligned}
$$

Define

$$
\begin{aligned}
\alpha_k := \frac{\mathrm{d}t_{k+1}}{\mathrm{d}t_k} &\left\| 1 + \frac{1}{2} \frac{\mathrm{d}t_k}{\mathrm{d}t_{k+1}} \int_{t_k}^{t_{k+1}} L_t \frac{\mathrm{d}t}{\mathrm{d}t_k}(t)\,\mathrm{d}t - \frac{\mathrm{d}t_k}{\mathrm{d}t_{k+1}} \int_{t_k}^{t_{k+1}} C_t \frac{\mathrm{d}t}{\mathrm{d}t_k}(t)\,\mathrm{d}t \right. \\
&\left. + \frac{\mathrm{d}t_k}{\mathrm{d}t_{k+1}} Mh \int_{t_k}^{t_{k+1}} \frac{\mathrm{d}t}{\mathrm{d}t_k}(t)\,\mathrm{d}t \right\|_{L_2} \\
&\left\| 1 + \frac{1}{2} \frac{\mathrm{d}t_k}{\mathrm{d}t_{k+1}} \int_{t_k}^{t_{k+1}} L_t \frac{\mathrm{d}t}{\mathrm{d}t_k}(t)\,\mathrm{d}t - \frac{\mathrm{d}t_k}{\mathrm{d}t_{k+1}} \int_{t_k}^{t_{k+1}} C_t \frac{\mathrm{d}t}{\mathrm{d}t_k}(t)\,\mathrm{d}t \right. \\
&\left. + \frac{\mathrm{d}t_k}{\mathrm{d}t_{k+1}} Mh \int_{t_k}^{t_{k+1}} \frac{\mathrm{d}t}{\mathrm{d}t_k}(t)\,\mathrm{d}t \right\|_{L_2} .
\end{aligned}
$$

By Proposition C.4, $\alpha_k \leq 1$ for any $0 \leq k \leq N-1$, which yields

$$
\begin{aligned}
\left\| X^1_T - X_T \right\|_{L_2} &\leq \prod_{k=0}^{N-1} \alpha_k \left\| X^1_0 - X_0 \right\|_{L_2} + \varepsilon h \Sigma(T) + Mh^2 \Sigma(T)(1+2B) \sum_{k=0}^{N-1} \prod_{\ell=k}^{N-1} \alpha_\ell \\
&\quad + \sum_{k=0}^{N-1} \left\| \int_{t_k}^{t_{k+1}} L_t \frac{\mathrm{d}t}{\mathrm{d}t_k}(t)\,\mathrm{d}t - \frac{1}{\sqrt{2h}} \sqrt{\Sigma(T) + m_T} \int_{t_k}^{t_{k+1}} \frac{1}{2} + 2L_t \,\mathrm{d}t \right\|_B \prod_{\ell=k}^{N-1} \alpha_\ell \\
&\quad \varepsilon_T \Sigma(T) + MhT \Sigma(T)(1+2B) \\
&\quad + \sum_{k=0}^{N-1} \left\| \int_{t_k}^{t_{k+1}} L_t \frac{\mathrm{d}t}{\mathrm{d}t_k}(t)\,\mathrm{d}t - \frac{1}{\sqrt{2h}} \sqrt{\Sigma(T) + m_T} \int_{t_k}^{t_{k+1}} \frac{1}{2} + 2L_t \,\mathrm{d}t \right\|_B .
\end{aligned}
$$

□

Final bound. Finally, combining the results of Proposition C.2 and Proposition C.3, we conclude that

$$W_2\left(\mu_{\text{data}}, b_N^{(\gamma,\eta)}\right) \le W_2(\mu_{\text{data}}, \nu_1)\exp\left(\int_0^T \frac{\alpha(t)}{2}\left(1+C_t^2\right)dt\right) + \sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} L_t\frac{\rho_t}{\rho_{t_k}}\alpha(t)dt\left(\frac{1}{\sqrt{2h}}\sqrt{\alpha(T)+m_T}\int_{t_k}^{t_{k+1}}\left(\frac{1}{2\sigma^2}+2L_t\right)\alpha(t)dt\, B\right) + \varepsilon_T\alpha(T) + Mh\,T\,\alpha(T)(1+2B).$$

C.3 Technical results for Wasserstein upper bound

Proposition C.4. Assume that H4 and H6 hold. Consider the regular discretization $\{t_k; 0 \le k \le N\}$ of $[0,T]$ of constant step size h. Assume that $h > 0$ is such that for all t_k with $0 \le k \le N-1$,

$$h < \frac{2C_t}{\alpha(t_k)\max_{t_k \le s \le t_{k+1}} L_s L_t}\frac{\rho_{t_{k+1}}}{\rho_{t_k}}; \tag{37}$$

where $\rho_t := \exp(\int_0^t \beta(s)ds/(2\sigma^2))$, $m_t := \exp(\int_0^t \beta(s)ds/(2\sigma^2))$. Then, for all $0 \le k \le N-1$,

$$0 < 1 + \frac{\rho_{t_k}^2}{\rho_{t_{k+1}}^2}\int_{t_k}^{t_{k+1}} L_t\frac{\rho_t}{\rho_{t_k}}\alpha(t)dt\Big)^2 - 2\frac{\rho_{t_k}}{\rho_{t_{k+1}}}\int_{t_k}^{t_{k+1}} C_t\frac{\rho_t}{\rho_{t_k}}\alpha(t)dt < 1.$$

In addition, if

$$h < \frac{2C_t}{M + \alpha(t_k)\max_{t_k \le s \le t_{k+1}} L_s L_t}\frac{\rho_{t_{k+1}}}{\rho_{t_k}}; \tag{38}$$

then, for all $0 \le k \le N-1$,

$$0 < 1 + \frac{1}{2}\frac{\rho_{t_k}^2}{\rho_{t_{k+1}}^2}\int_{t_k}^{t_{k+1}} L_t\frac{\rho_t}{\rho_{t_k}}\alpha(t)dt\Big)^2 - \frac{\rho_{t_k}}{\rho_{t_{k+1}}}\int_{t_k}^{t_{k+1}} C_t\frac{\rho_t}{\rho_{t_k}}\alpha(t)dt + \frac{\rho_{t_k}^2}{\rho_{t_{k+1}}^2}Mh\int_{t_k}^{t_{k+1}}\frac{\rho_t}{\rho_{t_k}}\alpha(t)dt < 1.$$

Proof. Denote λ_1 and λ_2 the following quantities

$$\lambda_1 = 1 + \frac{\rho_{t_k}^2}{\rho_{t_{k+1}}^2}\int_{t_k}^{t_{k+1}} L_t\frac{\rho_t}{\rho_{t_k}}\alpha(t)dt\Big)^2 - 2\frac{\rho_{t_k}}{\rho_{t_{k+1}}}\int_{t_k}^{t_{k+1}} C_t\frac{\rho_t}{\rho_{t_k}}\alpha(t)dt; \tag{39}$$

$$\lambda_2 = 1 + \frac{\rho_{t_k}^2}{\rho_{t_{k+1}}^2}\int_{t_k}^{t_{k+1}} L_t\frac{\rho_t}{\rho_{t_k}}\alpha(t)dt\Big)^2 - 2\frac{\rho_{t_k}}{\rho_{t_{k+1}}}\int_{t_k}^{t_{k+1}} C_t\frac{\rho_t}{\rho_{t_k}}\alpha(t)dt + \frac{\rho_{t_k}^2}{\rho_{t_{k+1}}^2}Mh\int_{t_k}^{t_{k+1}}\frac{\rho_t}{\rho_{t_k}}\alpha(t)dt. \tag{40}$$

First, we prove that γ_1 is positive. Completing the square, we obtain

$$\begin{aligned}\gamma_1 &= 1 - \frac{\sigma_{t_k}}{\sigma_{t_{k+1}}} \int_{t_k}^{t_{k+1}} L_t \frac{\sigma_t}{\sigma_{t_k}} \alpha(t)dt \Big)^2 + 2\frac{\sigma_{t_k}}{\sigma_{t_{k+1}}} \int_{t_k}^{t_{k+1}} L_t \frac{\sigma_t}{\sigma_{t_k}} \alpha(t)dt \\ &\quad 2\frac{\sigma_{t_k}}{\sigma_{t_{k+1}}} \int_{t_k}^{t_{k+1}} C_t \frac{\sigma_t}{\sigma_{t_k}} \alpha(t)dt \\ &= 1 - \frac{\sigma_{t_k}}{\sigma_{t_{k+1}}} \int_{t_k}^{t_{k+1}} L_t \frac{\sigma_t}{\sigma_{t_k}} \alpha(t)dt \Big)^2 + 2\frac{\sigma_{t_k}}{\sigma_{t_{k+1}}} \int_{t_k}^{t_{k+1}} L_t \quad C_t \frac{\sigma_t}{\sigma_{t_k}} \alpha(t)dt.\end{aligned}$$

The first term if the r.h.s. of the previous equality is a square, therefore always positive. The second term is always positive as well, as $L_t \geq C_t$ for any t, as the Lipschitz constant and the log-concavity coefficient of the score function respectively. Moreover, the previous is always strictly positive as

$$\frac{\sigma_{t_k}}{\sigma_{t_{k+1}}} \int_{t_k}^{t_{k+1}} L_t \frac{\sigma_t}{\sigma_{t_k}} \alpha(t)dt > 0.$$

Secondly, proving that the previous quantity is smaller than 1 is equivalent to show that

$$\frac{\sigma_{t_k}^2}{\sigma_{t_{k+1}}^2} \int_{t_k}^{t_{k+1}} L_t \frac{\sigma_t}{\sigma_{t_k}} \alpha(t)dt \Big)^2 - 2\frac{\sigma_{t_k}}{\sigma_{t_{k+1}}} \int_{t_k}^{t_{k+1}} C_t \frac{\sigma_t}{\sigma_{t_k}} \alpha(t)dt < 0.$$

As $\alpha(t)$ is a decreasing function, we obtain the following bound

$$\begin{aligned}&\frac{\sigma_{t_k}^2}{\sigma_{t_{k+1}}^2} \int_{t_k}^{t_{k+1}} L_t \frac{\sigma_t}{\sigma_{t_k}} \alpha(t)dt \Big)^2 - 2\frac{\sigma_{t_k}}{\sigma_{t_{k+1}}} \int_{t_k}^{t_{k+1}} C_t \frac{\sigma_t}{\sigma_{t_k}} \alpha(t)dt \\ &\quad \frac{\sigma_{t_k}}{\sigma_{t_{k+1}}} \Big(\max_{s \in [t_k, t_{k+1}]} L_s \Big) (t_k)h \Big) \frac{\sigma_{t_k}}{\sigma_{t_{k+1}}} \int_{t_k}^{t_{k+1}} L_t \frac{\sigma_t}{\sigma_{t_k}} \alpha(t)dt - 2\frac{\sigma_{t_k}}{\sigma_{t_{k+1}}} \int_{t_k}^{t_{k+1}} C_t \frac{\sigma_t}{\sigma_{t_k}} \alpha(t)dt \\ &= \frac{\sigma_{t_k}}{\sigma_{t_{k+1}}} \int_{t_k}^{t_{k+1}} \frac{\sigma_{t_k}}{\sigma_{t_{k+1}}} \Big(\max_{s \in [t_k, t_{k+1}]} L_s \Big) (t_k)h - L_t - 2C_t \Big) \frac{\sigma_t}{\sigma_{t_k}} \alpha(t)dt.\end{aligned}$$

This means that, if we have

$$\frac{\sigma_{t_k}}{\sigma_{t_{k+1}}} \Big(\max_{s \in [t_k, t_{k+1}]} L_s \Big) (t_k)hL_t - 2C_t < 0$$

for $t_k \leq t \leq t_{k+1}$, we have $\gamma_1 < 1$. Isolating h in the previous inequality, we obtain that it is equivalent to the condition (37).

Now we focus on γ_2. This quantity is clearly positive as the $\gamma_2 \geq \gamma_1$. Moreover, following the same lines as to prove that $\gamma_1 < 1$, we have

$$\gamma_2 \leq 1 - \frac{\sigma_{t_k}}{\sigma_{t_{k+1}}} \int_{t_k}^{t_{k+1}} \frac{\sigma_{t_k}}{\sigma_{t_{k+1}}} \Big(\max_{s \in [t_k, t_{k+1}]} L_s \Big) (t_k)hL_t + \frac{\sigma_{t_k}}{\sigma_{t_{k+1}}} Mh - 2C_t \Big) \frac{\sigma_t}{\sigma_{t_k}} \alpha(t)dt.$$

This means that, if we have

$$\frac{\mathrm{e}^{t_k}}{\mathrm{e}^{t_{k+1}}} \max_{t_k \, s \, t_{k+1}} L_s \quad (t_k) h L_t + \frac{\mathrm{e}^{t_k}}{\mathrm{e}^{t_{k+1}}} M h \quad 2C_t < 0$$

for $t_k \quad t \quad t_{k+1}$, we have $_2 < 1$. Isolating h in the previous inequality, we obtain that it is equivalent to the condition (38).

□

Proposition C.5. Assume that H2 holds. For all $t \quad 0$,

$$\sup_{0 \, t \, T} \quad X_t \quad _{L_2} \quad \sup_{0 \, t \, T} \; m_t^2 E\left[kX_0k^2 \right] + (1 \quad m_t^2) \; ^2 d \; ^{1=2} \quad E\left[kX_0k^2 \right] + \; ^2 d \; ^{1=2};$$

where $m_t = \exp(\int_0^t (s)ds=2 \; ^2)$.

Proof. Recall the following equality in law

$$! \; X_t = m_t X_0 + \; ^q \overline{(1 \quad m_t^2)} G :$$

with $X_0 \quad _{\text{data}}$ and $G \quad N \; (0; I_d)$.

Therefore, for any $t \; 2 \; [0; T]$

$$E \quad X_T \quad t \; ^2 = E \; ! \; X_t \; ^2 \quad m_t^2 E\left[kX_0k^2 \right] + \; ^2 \; 1 \quad m_t^2 \; E\left[kGk^2 \right]$$
$$m_t^2 E\left[kX_0k^2 \right] + \; ^2 \; 1 \quad m_t^2 \; d :$$

□

Proposition C.6. Assume that H2 holds. For all $t_k \quad t \quad t_{k+1}$,

$$\sup_{t_k \; t \; t_{k+1}} \quad X_t^1 \quad X_{t_k}^1 \quad _{L_2} \quad \frac{1}{-} p \overline{2h \; (T)} + m_T \int_{t_k}^{t_{k+1}} \frac{1}{2 \; ^2} + 2L_t \quad (t)dt \quad B ; \tag{41}$$

$$\sup_{0 \, t \, T} \quad X_t^1 \quad X_t \quad _{L_2} \quad \left[kX_0k^2 \right] + \; ^2 d \; ^{1=2} \exp \; ! \; \int_0^T \frac{(s)}{2 \; ^2} ds \; ; \tag{42}$$

where $m_t = \exp(\int_0^t (s)ds=2 \; ^2)$ and $B = (E[kX_0k^2] + \; ^2 d)^{1=2}$.

Proof. Note that $! \; X_t$ has the same distribution as $m_t X_0 + \; ^p \overline{(1 \quad m_t^2)} G$ where $G \quad N \; (0; I_d)$ is independent of X_0. We have that $X_0^1 = G \quad N \; (0; \; ^2 I_d)$. De ne $(X_t)_{t \; 2 \; [0;T]}$ satisfying (3) but initialized at

$$X_0 = m_T Y + \; ^q \overline{(1 \quad m_T^2)} G ; \tag{43}$$

with $Y \quad _{\text{data}}$ independent of G (G being shared by X_0 and X_0^1).

On the one hand, following the same proof as in Proposition C.2, we have that

$$\left\| X_t^1 - X_t \right\|_{L_2} \le \left\| X_0^1 - X_0 \right\|_{L_2} \exp\left(\int_0^t \frac{(s)}{2^{\ 2}} \left(1 + 2C_s^{\ 2} \right) ds \right) \left[\left\| Y \right\|_{k}^2 + {}^2 d \right]^{1=2} m_T ;$$

where we have used (43) as well as the fact that

$$\left\| X_0 - G \right\|_{k L_2} = \left[\left\| Y \right\|_{k}^2 + \left\| G \right\|_{k}^2 \right]^{1=2} = B :$$

Therefore,

$$\sup_{0 \ t \ T} \left\| X_t^1 - X_t \right\|_{L_2} \le \left[\left\| Y \right\|_{k}^2 + {}^2 d \right]^{1=2} \exp\left(\int_0^T \frac{(s)}{2^{\ 2}} ds \right) ;$$

corresponding to (42).

On the other hand, we have that

$$\left\| X_t^1 - X_{t_k}^1 \right\|_{L_2} \le \left\| X_t - X_{t_k} \right\|_{L_2} + \left\| X_t^1 - X_t - X_{t_k}^1 + X_{t_k} \right\|_{L_2} :$$

The process $(X_t^1 - X_t)_{t \ 0}$ is determined by the following ODE:

$$d\left(X_t^1 - X_t \right) = \frac{(t)}{2^{\ 2}} \left(X_t^1 - X_t \right) + 2 \ (t) \ r \log p_{T \ t}\left(X_t^1 - r \log p_{T \ t} X_t \right) dt :$$

Then,

$$\begin{aligned} &\left\| X_t^1 - X_t - X_{t_k}^1 + X_{t_k} \right\|_{L_2} \\ &= \int_{t_k}^{t} \frac{(s)}{2^{\ 2}} \left\| X_s^1 - X_s + 2 \ (s) \ r \log p_{T \ s}\left(X_s^1 - r \log p_{T \ s} X_s \right) \right\|_{L_2} ds \\ &\quad \sup_{t_k \ t \ t_{k+1}} \left\| X_t^1 - X_t \right\|_{L_2} \int_{t_k}^{t_{k+1}} \left(\frac{1}{2^{\ 2}} + 2L_t \right) (t) dt \\ &\quad Bm_T \int_{t_k}^{t_{k+1}} \left(\frac{1}{2^{\ 2}} + 2L_t \right) (t) dt : \end{aligned}$$

Write $(X_t)_{t2[0;T]}$ the time reversal of $(X_t)_{t2[0;T]}$, which clearly satisfies (1). Using the following equality in law

$$X_{T \ t_k} = \frac{m_{T \ t_k}}{m_{T \ t}} X_{T \ t} + \left(1 - \frac{m_{T \ t_k}}{m_{T \ t}} \right)^{2} {}^{1=2} G ;$$

with $G \sim N(0; I_d)$, we get

$$X_t - X_{t_k}\|_{L_2} = \|X_{T-t_k} - X_{T-t}\|_{L_2} = \left(1 - \frac{m_{T-t_k}}{m_{T-t}}\right)^{2} \|X_{T-t}\|^2 + \sigma^2 d^{1/2}$$
$$\left(1 - \frac{m_{T-t_k}}{m_{T-t}}\right)^{2} \sqrt{2B};$$

where we have applied Proposition C.5 in the last inequality. Since

$$1 - \frac{m_{T-t_k}}{m_{T-t}}^2 = 1 - \exp\left(-\frac{1}{2}\int_{T-t}^{T-t_k} (s)ds\right)$$
$$= \frac{1}{2}\int_{T-t}^{T-t_k} \exp\left(-\frac{1}{2}\int_{T-t}^{T-u} (s)ds\right) (u)du$$
$$\frac{1}{2} h\ (T);$$

which concludes the proof of (41). □

D Discussion on the hypotheses

Proposition D.1. Assume that $\log_{\text{data}}$ is C-strongly concave and that $C > 1/\sigma^2$. Then, the modified score function $\log p_t(x)$ is, for any $t \in (0, T]$, C_t-strongly concave, with

$$m_t = \exp\left(-\frac{1}{2\sigma^2}\int_0^t (s)ds\right);$$
$$C_t = \frac{1}{m_t^2/C + \sigma^2(1 - m_t^2)} \quad \frac{1}{2}:$$

Moreover, we have that $C_t \quad C \quad 1/\sigma^2$ for any $t \quad 0$.

Proof. This result is also proved in Saremi et al. (2023). We provide an alternative proof here for completeness. For all $1 \quad t \quad T$, X_t has the same law has $m_t X_0 + \sqrt{1 - m_t^2} Z$ where $X_0 \sim_{\text{data}}$ and $Z \sim N(0; I_d)$ are independent. Therefore, writing $p_0 = {}_{\text{data}}$,

$$p_t(y) = \int_{\mathbb{R}^d} (2\sigma^2 1 - m_t^2)^{d/2} \exp\left(\frac{k y - x_0 m_t k^2}{2\sigma^2(1 - m_t^2)}\right) p_0(x_0)dx_0: \tag{44}$$

This implies that

$$\log p_t(y) = \frac{d}{2}\log 2\sigma^2 1 - m_t^2 + \log \int_{\mathbb{R}^d} \exp\left(\frac{ky - x_0 m_t k^2}{2\sigma^2(1 - m_t^2)}\right) p_0(x_0)dx_0$$
$$= \frac{d}{2}\log 2\sigma^2 1 - m_t^2 + \log \int_{\mathbb{R}^d} \exp\left(\frac{ky - uk^2}{2\sigma^2(1 - m_t^2)}\right) p_0\left(\frac{u}{m_t}\right) du$$
$$+ \frac{d}{2\sigma^2}\int_0^t (s)ds:$$

Since $\log p_0$ is C -strongly concave, the function $x \mapsto p_0(u/m_t)$ is C/m_t^2-strongly log-concave. Moreover, we have that the function $y \mapsto \exp(-\|y\|^2/(2\sigma^2(1-m_t^2)))$ is $(\sigma^2(1-m_t^2))^{-1}$-strongly log-concave. Applying Saumard and Wellner (Proposition 7.1 2014), since p_t is a convolution of the previous two functions up to terms independent in space, we have that $\log p_t$ is $m_t^2/C + \sigma^2(1-m_t^2)^{-1}$-strongly concave. Note that if $C \geq 1/\sigma^2$,

$$\frac{C}{m_t^2 + \sigma^2 C(1-m_t^2)} \geq \frac{1}{2\sigma^2}.$$

This entails that $\log p_t$ is C_t-strongly concave, with

$$C_t = \frac{1}{m_t^2/C + \sigma^2(1-m_t^2)} \geq \frac{1}{2\sigma^2}.$$

Finally, finding the maximum $t \mapsto C_t$, is equivalent to find the maximum of the following function on $[0,1]$:

$$\varphi : z \mapsto \frac{C}{z + \sigma^2 C(1-z)} \geq \frac{1}{2\sigma^2}.$$

We have that $\varphi(0) = C \geq 1/\sigma^2$, $\varphi(1) = 0$ and for all $z \in [0,1]$,

$$\varphi'(z) = \frac{\sigma^2 - 1/C}{(z/C + \sigma^2(1-z))^2},$$

which is negative since $C \geq 1 \geq 1/\sigma^2$. Therefore, we get $0 \leq C_t \leq C \leq 1/\sigma^2$. □

Proposition D.2. If $\log p_{\text{data}}$ is L -smooth, then for all $0 \leq t \leq T$, $\nabla \log p_t$ is L_t-Lipschitz in the space variable with

$$L_t = \min\left(\frac{1}{\sigma^2(1-m_t^2)}, \frac{L}{m_t^2}\right) + \frac{1}{2\sigma^2}.$$

Moreover, if $L > 1/\sigma^2$, we can choose L_t as follows:

$$L_t = \min\left(\frac{1}{\sigma^2(1-m_t^2)}, \frac{L}{m_t^2}\right) \geq \frac{1}{2\sigma^2}.$$

Moreover, in this case, we have that $L_t \leq L$ for any $t \geq 0$.

Proof. In the proof of Proposition D.1, we proved that, if $\log p_{\text{data}}$ is C -strongly concave, $\log p_t$ is $m_t^2/C + \sigma^2(1-m_t^2)^{-1}$-strongly concave i.e.,

$$-\nabla^2(\log p_t)(x) \leq \frac{1}{m_t^2/C + \sigma^2(1-m_t^2)} I_d.$$

For $p_0 := p_{\text{data}}$, we have that p_t is given by (44). This means that p_t is the density of the sum of two independent random variables $X_1 + X_0$ of density respectively q_0 and q_1, such that

$$q_0(x) := \frac{1}{m_t^d} p_0\left(\frac{u}{m_t^d}\right) = e^{-\varphi_0(x)};$$

$$q_1(x) := \frac{1}{(2\pi\sigma^2(1-m_t^2))^{d/2}} \exp\left(-\frac{\|y\|^2}{2\sigma^2(1-m_t^2)}\right) = e^{-\varphi_1(x)};$$

for two functions $_0$ and $_1$. Therefore, as in the proof of Saumard and Wellner (Proposition 7.1 2014), we get

$$\begin{aligned} r^2(\ \log p_t)(x) &= \mathrm{Var}(r\ _0(X_0)jX_0 + X_1 = x) + E[r^2\ _0(X_0)jX_0 + X_1 = x] \\ &= \mathrm{Var}(r\ _1(X_1)jX_0 + X_1 = x) + E[r^2\ _1(X_1)jX_0 + X_1 = x]: \end{aligned}$$

Since $r\ \log p_0$ is L -Lipschitz and from the de nition of q_1,

$$r^2\ _0 \ 4\ \frac{L}{m_t^2} I_d; \quad r^2\ _1 \ 4\ \frac{1}{^2(1\ \ m_t^2)} I_d:$$

Hence,

$$r^2(\ \log p_t)(x) \ 4\ \min\ \frac{1}{^2(1\ \ m_t^2)}; \frac{L_0}{m_t^2}\ \ I_d:$$

Therefore, since the di erence between $r\ \log p_t$ and $r\ \log p_t$ is a linear function, we can choose L_t as follows:

$$L_t = \min\ \frac{1}{^2(1\ \ m_t^2)}; \frac{L}{m_t^2}\ \ + \frac{1}{^2}:$$

Clearly we have that $0\ \ m_t^2\ \ 1$, therefore $1=m_t^2\ \ 1$ and $1=\ 1\ \ m_t^2\ \ \ 1$. This means that, if $L\ \ 1=\ ^2$,

$$\min\ \frac{1}{^2(1\ \ m_t^2)}; \frac{L}{m_t^2}\ \ \ \frac{1}{^2}:$$

Thus, we can choose L_t to be

$$L_t = \min\ \frac{1}{^2(1\ \ m_t^2)}; \frac{L}{m_t^2}\ \ \ \frac{1}{^2}:$$

Finally, since $m_0 = 1$, we have that $L_0 = L\ \ 1=\ ^2$. This function increases up to the point where $L\ =m_t^2 = (\ ^2(1\ \ m_t^2)\ ^1$, achieved for $m_t^2 = (\ ^2L\ )=(\ ^2L\ + 1)$. At this point, we have that $L_t = L$. After this point the Lipschitz constant decreases to 0, as $m_t\ !\ 0$ for $t\ !\ 1$. This means that for any t, L_t is bounded by L . □

Proposition D.3. Assume that $\log\ _{\text{data}}$ is L -smooth and C -strongly concave. Consider the regular discretization $f\,t_k; 0\ \ k\ \ N\,g$ of $[0; T]$ of constant step size h. By choosing $h > 0$ such that for all t_k with $0\ \ k\ \ N\ \ 1$,

$$h\ \ \min\ \frac{\log(2)2\ ^2}{(T)}; \frac{^2C\ \ \ 1}{^2C\ (\ ^2L\ + 1)\ L\ \ (T)}; \frac{^2C\ \ \ 1}{(\ ^2L\ \ \ 1)\ L\ \ (T)}\ \ ; \tag{45}$$

then, for all $0\ \ k\ \ N\ \ 1$,

$$0 < 1 + \frac{\text{@}_{t_k}^2}{\text{@}_{t_{k+1}}^2} \int_{t_k}^{t_{k+1}} L_t \frac{\text{@}_t}{\text{@}_{t_k}}\ (t)dt\ \ ^2 \ 2\frac{\text{@}_{t_k}}{\text{@}_{t_{k+1}}} \int_{t_k}^{t_{k+1}} C_t \frac{\text{@}_t}{\text{@}_{t_k}}\ (t)dt < 1:$$

In addition, if

$$h \quad \min \quad \frac{\log(2)2^{\ 2}}{(T)}; \frac{{}^2C \quad 1}{{}^2M + \ (T)L \ (\ {}^2L \quad 1)}; \frac{{}^2C \quad 1}{{}^2C} \frac{m_T^2 \quad 1 \quad m_T^2}{{}^2M \ (1 \quad m_T^2) + \ (T)L \ m_T^2}; \frac{{}^2C \quad 1 \ L}{{}^2C \ (\ {}^2L \ + 1)(M + \ (T)L^2)} \quad ; \quad (46)$$

then, for all $0 \quad k \quad N \quad 1$,

$$0 < 1 + \frac{1}{2} \frac{\text{@}_{t_k}^2}{\text{@}_{t_{k+1}}^2} \int_{t_k}^{t_{k+1}} L_t \frac{\text{@}_t}{\text{@}_{t_k}} (t)dt \quad {}^2 \quad \frac{\text{@}_{t_k}}{\text{@}_{t_{k+1}}} \int_{t_k}^{t_{k+1}} C_t \frac{\text{@}_t}{\text{@}_{t_k}} (t)dt + \frac{\text{@}_{t_k}^2}{\text{@}_{t_{k+1}}^2} Mh \int_{t_k}^{t_{k+1}} \frac{\text{@}_t}{\text{@}_{t_k}} (t)dt < 1:$$

Proof. De ne $_1$ and $_2$ as in (39)-(40). From Proposition C.4, we have that $_i \ 2 \ (0; 1)$, for $i = 1; 2$, if we have (37)-(38).

First, we prove that (45) implies (37). From Proposition D.2, we have that L_t is bounded by L everywhere. Moreover, since $\text{@}_{t_{k+1}} = \text{@}_{t_k} = \exp \int_{t_k}^{t_{k+1}} (s)=2 \ {}^2ds$, we can nd h small enough such that $2\text{@}_{t_k} = \text{@}_{t_{k+1}} \quad 1$. This is equivalent to $\int_{t_k}^{t_{k+1}} (s)=2 \ {}^2ds \quad \log(2)$ and it is implied by

$$h \quad \frac{\log(2)2^{\ 2}}{(T)}:$$

Now, we study the function $t \ 7! \ C_t=L_t$. From the proof of the Proposition D.1, we have that

$$C_t = \frac{1}{m_t^2=C \ + \ {}^2(1 + m_t^2)} \quad \frac{1}{2};$$

which is a decreasing function. Moreover, from the proof of the Proposition D.2, we have that

$$L_t = \min \quad \frac{1}{{}^2(1 \quad m_t^2)}; \frac{L}{m_t^2} \quad \frac{1}{2};$$

which is an increasing function from 0 up to t, such that $m_t^2 = \frac{{}^2L}{{}^2L \ +1}$ and decreasing for $t \quad t$. On the one hand, this means that for $t \ 2 \ [0; t \ ]$, the function $t \ 7! \ C_t=L_t$ is decreasing, therefore reaching its minimum $\ {}^2C \quad 1 = \ {}^2L \quad 1$ in 0, which is a positive quantity. On the other hand, for $t \quad t$, we have that

$$\frac{C_t}{L_t} = \frac{1}{m_t^2=C \ + \ {}^2(1 + m_t^2)} \frac{1}{2} \frac{{}^2C \quad 1 \ m_t^2}{C} \frac{{}^2 \ 1 \quad m_t^2}{m_t^2} \frac{1}{2} \frac{{}^2C \quad 1}{C} \ 1 \quad m_t^2 \frac{1}{2} \frac{{}^2C \quad 1}{C} \ 1 \quad m_t^2 = \frac{{}^2C \quad 1}{{}^2C \ (\ {}^2L \ + 1)}:$$

Therefore, combining the previous inequalities, we have that condition (45) implies (37).

Secondly, we prove (46) implies (38). Take h to satisfy

$$h \quad \frac{\log(2)2^{\ 2}}{(T)} :$$

We now need to study the function $t \ 7! \ \frac{C_t}{M + (T)L \ L_t}$. On the one hand, this function is decreasing for $t \ 2 \ [0; t \ ]$, therefore reaching its minimum $\frac{^2C \quad 1}{^2M + (T)L \ (^2L \quad 1)}$ in 0, which is a positive quantity. On the other hand, for $t \quad t$, we have that

$$\frac{C_t}{M + (T)L \ L_t} = \frac{1}{m_t^2=C + {}^2(1 + m_t^2)} \ \frac{1}{2} \ \frac{^2C \quad 1 \ m_t^2}{C} \ \frac{^2 \ 1 \quad m_t^2}{^2M \ (1 \quad m_t^2) + (T)L \ m_t^2}$$
$$\frac{1}{2} \ \frac{^2C \quad 1}{C} \ \frac{m_t^2 \ 1 \quad m_t^2}{^2M \ (1 \quad m_t^2) + (T)L \ m_t^2} :$$

Controlling from below the previous quantity, boils down to control from below the function $(y) = \frac{y(1 \quad y)}{^2M \ (1 \quad y) + (T)L \ y}$ for $y \ 2 \ [m_t^2 \ ; m_T^2]$. We see that $\quad$ in this interval can be bounded by $\min f \quad m_t^2 \quad ; \quad m_T^2 \ g$. Therefore, we get

$$\frac{C_t}{M + (T)L \ L_t} \quad \frac{^2C \quad 1}{^2C} \min \left(\frac{m_T^2 \ 1 \quad m_T^2}{^2M \ (1 \quad m_T^2) + (T)L \ m_T^2}; \frac{L}{(^2L_0 + 1)(M + (T)L^2)} \right) :$$

Therefore, combining the previous inequalities, we have that condition (46) implies (38).

□

E Details on numerical experiments

This section is divided into two parts. The first part is dedicated to providing detailed implementation choices for the numerical experiments presented in Section 5. The second part displays additional experiments and more details about the experiments of Section 5. All experiments were conducted on a local computer CPU equipped with an Apple M3 processor (8GB of unified memory). This setup is sufficient to replicate the experiments of this paper.

E.1 Implementation choices

E.1.1 Exact score and metrics in the Gaussian case

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

$$\overset{!}{dX}_t = \quad \frac{(t)}{2 \ ^2} X_t dt + \sqrt{(t)} dB_t ; \quad \overset{!}{X}_0 \quad _0;$$

is initialised with $_0$ the Gaussian probability density function with mean $_0$ and variance $_0$. Then, the score function of (1) is:

$$r \ \log p_t : x \ 7! \quad (m_t^2 \ _0 + \ ^2_t I_d) \ ^1(x \quad m_t \ _0) ;$$

where p_t is the probability density function of $\overset{!}{X}_t$, $m_t = \exp\left\{ \int_0^t (s)ds=(2 \ ^2) \right\}$ and $^2_t = \ ^2(1 \quad m_t^2)$.

Proof. Note that X_t has the same law as $m_t X_0 + \gamma_t Z$ where $Z \sim N(0, I_d)$ is independent of X_0. Therefore $X_t \sim N(m_t \mu_0, \Sigma_t)$ with $\Sigma_t = m_t^2 \Sigma_0 + \gamma_t^2 I_d$ which concludes the proof.

□

Lemma E.2. Let μ_1, μ_2 in $\mathbb{R}^d$ and Σ_1 and Σ_2 be two definite positive matrices in $\mathbb{R}^{d \times d}$. Then,

$$KL(\mathcal{N}(\mu_1, \Sigma_1) \| \mathcal{N}(\mu_2, \Sigma_2)) = \frac{1}{2} \log \frac{|\Sigma_2|}{|\Sigma_1|} - d + \mathrm{Tr}\left(\Sigma_2^{-1}\Sigma_1\right) + (\mu_2 - \mu_1)^\top \Sigma_2^{-1}(\mu_2 - \mu_1). \quad (47)$$

Lemma E.3. Let μ_1, μ_2 in $\mathbb{R}^d$ and Σ_1 and Σ_2 be two definite positive matrices in $\mathbb{R}^{d \times d}$. Then,

$$W_2^2(\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)) = \|\mu_2 - \mu_1\|^2 + \mathrm{Tr}\left(\Sigma_1 + \Sigma_2 - 2\left(\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2}\right)^{1/2}\right). \quad (48)$$

Lemma E.4. The relative Fisher information between $X_0 \sim N(\mu_0, \Sigma_0)$ and $X_1 \sim N(0, \sigma^2 I_d)$ is given by:

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

Proof. The relative Fisher information between X_0 and X_1 is given by

$$I(\rho_{\mu_0, \Sigma_0} | \gamma_{\sigma^2}) = \int \left\| \nabla \log \frac{\rho_{\mu_0, \Sigma_0}(x)}{\gamma_{\sigma^2}(x)} \right\|^2 \rho_{\mu_0, \Sigma_0}(x) dx.$$

Write

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

so that,

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

First,

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

Then,

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

Using that $\mathbb{E}\left[X_0 X_0^\top\right] = {}_0 + {}_0 {}_0^\top$ yields

$$\begin{aligned}
\mathbb{E}\left[\frac{2}{2} X_0^\top {}_0^{-1}(X_0 \quad {}_0)\right] &= \frac{2}{2}\,\mathrm{Tr}\left({}_0^{-1} \quad {}_0 + {}_0 {}_0^\top \quad {}_0^\top {}_0^{-1} {}_0\right) \\
&= \frac{2}{2}\left(d + \mathrm{Tr}\left({}_0^{-1} {}_0 {}_0^\top \quad {}_0^\top {}_0^{-1} {}_0\right)\right) \\
&= \frac{2d}{2}.
\end{aligned}$$

Finally,

$$\begin{aligned}
\mathbb{E}\left[(X_0 \quad {}_0)^\top {}_0^{-2}(X_0 \quad {}_0)\right] &= \mathbb{E}\left[\mathrm{Tr}\left((X_0 \quad {}_0)^\top {}_0^{-2}(X_0 \quad {}_0)\right)\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\left({}_0^{-2}(X_0 \quad {}_0)(X_0 \quad {}_0)^\top\right)\right] \\
&= \mathrm{Tr}\left({}_0^{-2}\,\mathbb{E}\left[(X_0 \quad {}_0)(X_0 \quad {}_0)^\top\right]\right) \\
&= \mathrm{Tr}\left({}_0^{-2} {}_0\right) \\
&= \mathrm{Tr}\left({}_0^{-1}\right),
\end{aligned}$$

which concludes the proof. □

Proposition E.5. Under the same assumptions as in Lemma E.1, the Euclidean norm of the score function admits the following upper bound for $t_1 \quad t_2$:

$$\sup_{t_1 \quad t \quad t_2} \left\| \nabla \log p_{t_1}(x) \quad \nabla \log p_t(x) \right\| \quad (t_2 \quad t_1) \max\left\{ \| {}_0 \|_2;\ {}_1 \right\}(1 + \|x\|),$$

with

$${}_1 := \frac{m_{t_1}^2 \frac{(t_2)}{2} \quad \min^{\quad 2}}{{}^2 + m_{t_1}^2 (\quad \min^{\quad 2})\left({}^2 + m_{t_2}^2 (\quad \min^{\quad 2})\right)};$$

and

$${}_2 := \frac{m_{t_1} \frac{(t_2)}{2^{\,2}} \quad m_{t_1} m_{t_2} \quad \min^{\quad 2} \quad {}^2}{{}^2 + m_{t_1}^2 (\quad \min^{\quad 2})\left({}^2 + m_{t_2}^2 (\quad \min^{\quad 2})\right)};$$

where $\min$ is the smallest eigenvalue of ${}_0$.

Proof. Let $t_1 \quad t_2$,

$$\begin{aligned}
\left\| \nabla \log p_{t_1}(x) \quad \nabla \log p_{t_2}(x) \right\| = \quad & \left(m_{t_1}^2 \quad {}_0 + {}_{t_1}^2 I_d\right)^{-1}(x \quad m_{t_1} \quad {}_0) + \left(m_{t_2}^2 \quad {}_0 + {}_{t_2}^2 I_d\right)^{-1}(x \quad m_{t_2} \quad {}_0) \\
& m_{t_1} \left(m_{t_1}^2 \quad {}_0 + {}_{t_1}^2 I_d\right)^{-1} \quad m_{t_2} \left(m_{t_2}^2 \quad {}_0 + {}_{t_2}^2 I_d\right)^{-1} \quad {}_0 \\
+ \quad & \left(m_{t_1}^2 \quad {}_0 + {}_{t_1}^2 I_d\right)^{-1} \quad \left(m_{t_2}^2 \quad {}_0 + {}_{t_2}^2 I_d\right)^{-1} x \ :
\end{aligned}$$

Writing $M_t = m_t^2 \ _0 + \ _t^2 I_d \ ^1$ we have, for $t_1 \ t_2$,

$$kM_{t_1} \ M_{t_2}k \ \frac{1}{m_{t_1}^2 \ _{\min} + \ _{t_1}^2} \ \frac{1}{m_{t_2}^2 \ _{\min} + \ _{t_2}^2} \ \frac{m_{t_2}^2 \ m_{t_1}^2 \ _{\min} \ ^2}{^2 + m_{t_1}^2 (\ _{\min} \ ^2) \ ^2 + m_{t_2}^2 (\ _{\min} \ ^2)}$$
$$(t_2 \ t_1) \ \underbrace{\frac{m_{t_1}^2 \ \frac{(t_2)}{^2} \ _{\min} \ ^2}{^2 + m_{t_1}^2 (\ _{\min} \ ^2) \ ^2 + m_{t_2}^2 (\ _{\min} \ ^2)}}_{z_1} :$$

Moreover, for $t_1 \ t_2$,

$$km_{t_1} M_{t_1} \ m_{t_2} M_{t_2}k \ \frac{m_{t_1}}{m_{t_1}^2 \ _{\min} + \ _{t_1}^2} \ \frac{m_{t_2}}{m_{t_2}^2 \ _{\min} + \ _{t_2}^2}$$
$$\frac{m_{t_1} m_{t_2}^2 \ m_{t_2} m_{t_1}^2 \ _{\min} \ ^2 + \ ^2 (m_{t_1} \ m_{t_2})}{^2 + m_{t_1}^2 (\ _{\min} \ ^2) \ ^2 + m_{t_2}^2 (\ _{\min} \ ^2)}$$
$$\frac{jm_{t_2} \ m_{t_1} j \ m_{t_1} m_{t_2} (\ _{\min} \ _2) \ ^2}{^2 + m_{t_1}^2 (\ _{\min} \ ^2) \ ^2 + m_{t_2}^2 (\ _{\min} \ ^2)}$$
$$(t_2 \ t_1) \ \underbrace{\frac{m_{t_1} \ \frac{(t_2)}{2 \ ^2} \ m_{t_1} m_{t_2} \ _{\min} \ ^2 \ ^2}{^2 + m_{t_1}^2 (\ _{\min} \ ^2) \ ^2 + m_{t_2}^2 (\ _{\min} \ ^2)}}_{z_2} :$$

Finally,

$$kr \ \log p_{t_1}(x) \ r \ \log p_{t_2}(x)k \ (t_2 \ t_1) k \ _0 k \ _2 + (t_2 \ t_1) \ _1 kxk$$
$$(t_2 \ t_1) \max fk \ _0 k \ _2; \ _1 g(1 + kxk) :$$

□

E.1.2 Stochastic di erential equation exact simulation

In certain cases, exact simulation of stochastic di erential 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

$$! X_t = m_t X_0 + \ _t Z ;$$

with $Z \ N \ (0; I_d)$ independent of X_0, $X_0 \ _0$, $m_t = \exp f \ _0^{R_t} \ (s)ds=(2 \ ^2)g$ and $_t^2 = \ ^2(1 \ \exp f \ _0^{R_t} \ (s)= \ ^2 ds g)$. Therefore, sampling from the forward process only necessitates access to samples from $_0$ and $N(0; I_d)$.

E.1.3 Noise schedules

Linear and parametric noise schedules. In Section 5, we introduced parametric noise schedules of the form

$$\bar{\alpha}_a(t) \propto (e^{at} - 1)/(e^{aT} - 1),$$

with $a \in \mathbb{R}$ ranging from -10 to 10 (see Figure 6). For all a, with a time horizon of $T = 1$, the initial and final values have been set to match exactly the schedule prescribed by Song et al. (2021) (i.e. $\bar{\alpha}_a(0) = 0.1$ and $\bar{\alpha}_a(1) = 20$) when $a = 0$ (linear schedule).

Figure 6: Evolution of noise schedules $\bar{\alpha}_a$ w.r.t. time, for different values of parameter a between -10 to 10. The linear case $a = 0$ (Song and Ermon, 2019; Song et al., 2021) is dashed.

As shown in Section E.1.2 m_t and σ_t are the two quantities of interest in the calibration of the noising procedure of the forward proces. Their values for different choices of a are displayed in Figure 7.

Cosine noise schedule. We consider the cosine schedule introduced in Nichol and Dhariwal (2021) for which the forward process is defined for $t \in \{1, \dots, T\}$ as

$$X_t := \sqrt{\bar{\alpha}_t} X_0 + \sqrt{1 - \bar{\alpha}_t} Z,$$

with $X_0 \sim \pi_{\text{data}}$, $Z \sim \mathcal{N}(0, I_d)$ and with

$$\bar{\alpha}_t = \frac{f(t)}{f(0)}, \quad f(t) = \cos\left(\frac{t/T + s}{1 + s} \frac{\pi}{2}\right)^2.$$

To use this noise schedule in the SDE setting we notice that the forward process writes, for $t \in [0, T]$,

$$X_t \overset{!}{=} m_t X_0 + \sigma_t Z,$$

m_t $\quad$ $_t$

Figure 7: Evolution of m_t and $_t$ over time, for di erent choices of a in the noise schedule $_a$ used in see Section 5. The stationary distribution of the forward process ^{2} is set to 1. The range fora spans from -10 to 10, with the dashed line representing the linear schedule as proposed originally in the VPSDE models (Song et al., 2021).

with $m_t = \exp\{\int_0^t (s)ds=(2\ ^2)\}$, $_t^2 = \ ^2(1 \quad m_t^2)$ and $Z \quad N \ (0; I_d)$. Therefore, we can simply identify $_{cos}$ by solving

$$\int_0^t \frac{_{cos}(s)}{2\ ^2} ds = \log(\ _t) \ ;$$

which yields the following noising function:

$$_{cos}(t) := \ ^2 \frac{}{T\ (s+1)} \tan\left(\frac{(s+\ t=T)}{2\ (s+1)}\right) \ : \tag{49}$$

Finally, to ensure fair comparison with the linear schedule and the parametric schedules de ned in Section 5, we set in all our experiments$s = 0:021122$so that $_{cos}(0) \quad _a(0) = 0:1$ for any a.

E.1.4 Discretization details of the di usion 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:

$$dX_t = \quad \frac{(t)}{2\ ^2} X_t + \ (t)r \ \log p_{T\ t}(X_t) \ dt + \sqrt{(t)}dB_t; \quad X_0 \quad _1 :$$

Consider time intervals $0 \quad t_k \quad t \quad t_{k+1} \quad T$, with $t_k = \sum_{`=1}^{k} \ `$ and $T = \sum_{k=1}^{N} \ _k$.

In our theoretical analysis, we have considered the Exponential Integrator discretization, de ned recursively for $t \ 2 \ [t_k; t_{k+1}]$ by

$$dX_t^{EI} = \ (t) \quad \frac{1}{2\ ^2} X_t^{EI} + r \ \log p_{T\ t_k}(T \quad t_k; X_{t_k}^{EI}) \ dt + \sqrt{(t)}dB_t; \quad X_0^{EI} \quad _1 :$$

Figure 8: Evolution of noising functions under the cosine schedule (orange, $_{cos}$) compared to the linear schedule ($_0$, blue) over time with $^2 = 1$ and $s = 0:021122$ Additionally, since $_0$ increases unboundedly near T, we clip its value to 200 for better visualization.

m_t $_t$

Figure 9: Evolution of m_t and $_t$ for both the cosine schedule (orange) and the linear schedule (blue) w.r.t. time, with $s = 0:021122$ and $^2 = 1$. We clip the value of $_{cos}$ by 200 for better visualization.

In the numerical experiments, we have given priority to the Euler-Maruyama discretization, which is widely used, and de ned recursively for $t \in [t_k; t_{k+1}]$ by

$$dX_t^{EM} = \frac{(t_k)}{2\ ^2} X_{t_k}^{EM} + (t_k)r \log p_{T\ t_k}(X_{t_k}^{EM})\,dt + \sqrt{(t_k)}dB_t; \quad X_0^{EM} \quad _1 : \qquad (50)$$

To ensure transparency, the graphs presented in Figure 2 of Section 5.1 are reproduced in Figure 10 using an Exponential Integrator discretization scheme. As expected for ne discretization steps (here 500 steps were used) the two schemes produce nearly identical results.

(a) Isotropic setting (b) Heteroscedastic setting (c) Correlated setting

Figure 10: Comparison of the empirical KL divergence (mean± std over 30 runs) (top) and W_2 distance (mean± std over 10 runs) (bottom) between $_{data}$ and $b_N^{(\ ;\ )}$ (orange) and the related upper bounds (blue) from Theorem 3.1 and Theorem 4.2 across parameter for noise schedule $_a$, d = 50 with Exponential Integrator discretization scheme. We also show the metrics for the linear VPSDE model (dashed line) and our model (dotted line) with exact score evaluation.

E.1.5 Implementation of the score approximation in the Gaussian setting

Although the score function is explicit when $_{data}$ is Gaussian (see Lemma E.1), we implement SGMs as done in applications, i.e., we train a deep neural network to witness the e ect of the noising function on the approximation error. We train a neural network architecture $s\ (t;x)\ 2\ [0;T]\ \ R^d\ 7!\ R^d$ using the actual score function as a target:

$$L_{\text{explicit}}\ (\ ) = E\ \ s\ \ ;\overset{!}{X}\ \ r\ \ \log p\ \ \overset{!}{X}\ \ ^2$$

$$= E\ \ s\ \ ;\overset{!}{X}\ \ (m^2\ _0 + \ ^2 I_d)\ \ ^1\overset{!}{(X}\ \ + m\ \ _0)\ \ ^2\ \ ;$$

where $t\ !\ \ m_t$ and $t\ !\ \ _t$ are de ned in Lemma E.1 and $\ \ U\ (0;T)$ is independent of $\overset{!}{X}$. The neural network architecture chosen for this task is described in Figure 11. The width of each dense layer mid_features is set to 256 throughout the experiments.

Figure 11: 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.

E.2 Details on the experiments and additional results

E.2.1 Illustration of the KL bound in the Gaussian setting

Target distributions. We investigate the relevancy of the upper bound from Theorem 3.1 for different noise schedules in the Gaussian setting. We use as a training sample 10^4 samples with distribution $\mathcal{N}(1_d; \Sigma)$ for d the dimension of the target distribution with different choices of covariance structure.

1. (Isotropic) $\Sigma^{\text{iso}} = 0.5 I_d$.
2. (Heteroscedastic) $\Sigma^{\text{heterosc}} \in \mathbb{R}^{d \times d}$ is a diagonal matrix such that $\Sigma^{\text{heterosc}}_{jj} = 1$ for $1 \leq j \leq d$, and $\Sigma^{\text{heterosc}}_{jj} = 0.01$ 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 $\Sigma^{\text{corr}}_{jj'} = 1/\sqrt{jj'}$ for $1 \leq j \neq j' \leq d$.

The resulting data distributions are respectively denoted by $q_{\text{data}}^{(\text{iso})}$, $q_{\text{data}}^{(\text{heterosc})}$ and $q_{\text{data}}^{(\text{corr})}$.

Upper bound evaluation. We leverage the Gaussian nature of the target distribution to compute explicitly all the terms in the bound. On the one hand, the relative entropy in E_1^{KL}, $KL(q_{\text{data}} \| \gamma_1)$

is computed using the analytical formula for KL-divergence between two random Gaussian variable (Lemma E.2). On the other, the relative Fisher information in E_3^{KL}, $I(\ _{data} \mid\ _1)$, is computed using Lemma (E.4). Moreover, as the noise schedule function $_a$ and its primitive are analytically known, every occurrences of either of them are explicitly computed. Finally, it remains to estimate the expectation in $E_2^{KL}(\ ;\ )$. This is done via Monte Carlo estimation on 500 samples from the forward process (see Section E.1.2) for every step forward:

$$\frac{1}{500}\sum_{k=0}^{N-1}\sum_{i=1}^{500} r\log p_{T-t_k}\left(X_{T-t_k}^{(i)}\right) s-T-t_k;X_{T-t_k}^{(i)}\Big)^2\int_{T-t_{k+1}}^{T-t_k} {}_a(t)dt:$$

SGM data generation in dimension 50. In Figures 2 (top) of the main paper, we represent the following quantities in the same graph, in dimension $d = 50$, for different values of a.

- In blue the upper bound from Theorem 3.1. The dark blue color is used to refer to the upper bound with the contraction argument in equation (7) from Proposition C.1 while the ligther blue bound is the same bound without the contraction argument.
- In orange (dotted line) the KL divergence between the target distribution $_{data}$ and the empirical mean and covariance of the data generated using the true score function from Lemma E.1.
- In orange (plain line) we represent $KL(\ _{data} \| b_N^{(\ _a;\ )})$ for $a \in \{-10, -9, -8, \dots, 10\}$. That is, the KL divergence between the target distribution $_{data}$ and the empirical mean and covariance of the data generated using the neural network architecture described in Figure 11 to approximate the score function.
- In orange (dashed line) we represent $KL(\ _{data} \| b_N^{(\ _0;\ )})$. That is, the KL divergence between the target data $_{data}$ and the empirical mean and covariance of the data generated by the linear schedule VPSDE presented in Song et al. (2021) with the neural network architecture described in Figure 11.

We generate 10 000 samples. The batch size is set to 64 and neural networks are optimized with Adam. All the KL divergences written above are computed using Lemma E.2. Due to the stochastic nature of our experiments, they are repeated 30 times so that the corresponding mean value and standard deviations of these results are respectively depicted using plain and ll-in-between plots.

To disentangle the effect of each error term it is possible to plot the mixing time error $E_1^{KL}(\ )$, the approximation error $E_2^{KL}(\ ;\ )$ and the discretization error $E_3^{KL}(\ )$ on a same graph for different values of a. However, for the schedule choice presented in Figure 1 as (T) is set to be 20 for every a values it is pointless to display $E_3^{KL}(\ )$ as it would not vary for different choices of schedule from $_a$. The three error terms for Theorem 3.1, corresponding to the example in Figure 2 (top) are provided below in Figure 12.

Optimal schedule versus classical choices. We investigate the gain from using the parametric schedule with $a^?$ minimising the upper bound from Theorem 3.1 for $d \in \{5, 10, 25, 50\}$ compared to using the linear and cosine schedules (see Appendix E.1.3) in the isotropic and correlated settings

(a) Isotropic setting (b) Heteroscedastic setting (c) Correlated setting

Figure 12: Error terms contribution from Theorem 3.1 displayed from the same examples as in Figure 2 (top).

(as mentioned in Section 5.1, up to rescaling the heteroscedastic setting boils down to an isotropic setting).

To determine the optimal value $a^\star$, upper bounds were initially calculated across various dimensions for a range of a values from $\{-10, -9, \dots, 10\}$. This initial calculation aimed to identify a preliminary minimum value. Subsequently, the search was refined around these preliminary values using finer step-sizes of 0.25 to more precisely locate $a^\star$.

Results are given in tabular form in Table 1 and in Figure 13. The parametric schedule optimized to minimize the upper bound $_{a^\star}$ consistently surpasses the linear schedule, delivering significant improvements. This enhanced performance is shown by lower average Kullback-Leibler divergence between $_{\text{data}}$ and the generated sample distribution, as well as a reduction in the standard deviation of these divergences, which contributes to more stable generation. These results are competitive with or even exceed those obtained with state-of-the-art schedules such as the cosine schedule, particularly in higher dimensions $d = 25$ and $d = 50$. However, one should note that this comparison may not be entirely fair, as the cosine schedule increases unboundedly near T, whereas we capped the parametric schedule at $(T) = 20$ to align with the linear schedule described in Song et al. (2021).

E.2.2 Illustration of the Wasserstein bound in the Gaussian setting

Target distributions. The target distributions are Gaussian and are the same as for the the Kullback-Leibler bound: $^{(\text{iso})}_{\text{data}}$, $^{(\text{heterosc})}_{\text{data}}$ and $^{(\text{corr})}_{\text{data}}$.

Upper bound evaluation. We leverage the Gaussian nature of the target distribution to compute explicitly all the terms in the bound from Theroem 4.2. For the mixing time $E_1^{W_2}$, the strong log-concavity constant C_t is derived using Lemma 4.1 and $W_2(\ _{\text{data}}\ ;\ _1\ )$ is derived using Lemma E.3. For $E_2^{W_2}$, the analytical expressions for L_t is given in Lemma 4.1 and an upper bound to M is derived in Proposition E.5. All non analytically solvable integrals estimated numerically using the trapezoidal rule, implemented with the built-in PyTorch function `torch.trapezoid`. To estimate $"$, we use Monte-Carlo simulations with 500 samples (in the same manner as for the Kullback-Leibler

	Dimension	5		10		25		50	
Isotropic	Upper bound min $a^?$	1.75		1.00		1.50		2.00	
	Generation value in $a^?$	0.001607	0.000462	0.005343	0.001155	0.026724	0.004046	0.095981	0.005485
	VPSDE (linear sched.)	0.001935	0.000405	0.005594	0.001377	0.031748	0.006158	0.105592	0.019529
	Cosine schedule	0.001390	0.000296	0.005097	0.001064	0.026900	0.001859	0.099917	0.004375
	% gain (vs VPSDE)	+16.93 %		+4.48 %		+15.80 %		+9.10 %	
	% gain (vs Cosine)	-15.61 %		-4.83 %		+0.66 %		+3.94 %	
Correlated	Upper bound min $a^?$	2.25		1.75		1.75		2.25	
	Generation value in $a^?$	0.001861	0.000880	0.005871	0.001165	0.033156	0.003785	0.109649	0.008056
	VPSDE (linear sched.)	0.002568	0.002708	0.006210	0.001816	0.038434	0.010313	0.134716	0.016541
	Cosine schedule	0.001197	0.000332	0.005515	0.000775	0.040430	0.003475	0.110515	0.004646
	% gain (vs VPSDE)	+27.53 %		+5.46 %		+13.74 %		+18.63 %	
	% gain (vs Cosine)	-55.47 %		-6.46 %		+17.98 %		+0.78 %	
Parameters	Learning rate	1e-4		1e-4		1e-3		1e-3	
	Epochs	20		30		75		150	

Table 1: Comparison of the KL divergence between the target value and the generated value at $a^?$ (the minimum value of the upper bound from Theorem 3.1) with the KL divergence between the generated value by VPSDE with linear schedule and the target distribution. We display average KL divergences plus or minus standard deviations over 10 runs. The target distributions are chosen to be Gaussian with di erent covariance structures: isotropic ($_{\text{data}}^{(\text{iso})}$), heteroscedastic ($_{\text{data}}^{(\text{heterosc})}$) and correlated ($_{\text{data}}^{(\text{corr})}$).

(a) Isotropic setting $_{\text{data}}^{(\text{iso})}$ (b) Correlated setting $_{\text{data}}^{(\text{corr})}$

Figure 13: Comparison of the empirical KL divergence (mean value std over 10 runs) between $_{\text{data}}$ and the generative distribution ^ for di erent values of the dimension. The generative distributions considered are $b_N^{(a^?;\ )}$) obtained by the time-inhomogeneous SGM for $a^?$ (blue plain), $b_N^{(0;\ )}$ obtained by a standard linear VPSDE model (yellow dashed) and $b_N^{(\cos;\ )}$ obtained by using a cosine schedule (orange dotted).

bound):

$$\sup_{k \in \{0,\dots,N-1\}} \sqrt{\frac{1}{500}\sum_{i=1}^{500} r \log p_T\left(t_k, X_T^{(i)}(t_k)\right) - s - T\left(t_k; X_T^{(i)}(t_k)\right)^2}:$$

SGM data generation dimension 50. In Figures 14 (and Figures 2 (bottom) of the main paper) we represent on the same graph, in dimension $d = 50$, for different values of a:

- in blue the upper bound from Theorem 4.2.
- in orange (dotted line) the W_2 distance between the target distribution $_{data}$ and the empirical mean and covariance of the data generated using the true score function from Lemma E.1.
- in orange (plain line) we represent $W_2(\ _{data}\ ; b_N^{(\ a\ ;\ )})$ for $a \in \{\ 10;\ \ 9;\ \ 8; :::; 10\}$. That is, the W_2 distance between the target distribution $_{data}$ and the empirical mean and covariance of the data generated using the neural network architecture described in Figure 11 to approximate the score function
- in orange (dashed line) we represent $W_2(\ _{data}\ ; b_N^{(\ 0;\ )})$. That is, the W_2 distance between the target data $_{data}$ and the empirical mean and covariance of the data generated by the linear schedule VPSDE presented in Song et al. (2021) with the neural network architecture described in Figure 11.

First results. We generate 10 000 samples. The batch size is set to 64 and neural networks are optimized with Adam. All the W_2 distances written above are computed using Lemma E.3. Due to the stochastic nature of our experiments, they are repeated ten times so that the corresponding mean value and standard deviations of these results are respectively depicted using plain and ll-in-between plots.

(a) Isotropic setting (b) Heteroscedastic setting (c) Correlated setting

Figure 14: Comparison of the empirical 2-Wasserstein distance (mean value std over 10 runs) between $_{data}$ and $b_N^{(\ a\ ;\ )}$ (in orange) and the upper bound from Theorem 4.2 (in blue) w.r.t. the parameter a used in the de nition of the noise schedule $_a$, for $d = 50$. We also represent the 2-Wasserstein distances obtained with the linear VPSDE model (dashed line) and the one obtained with the parametric model (dotted line) when the score is not approximated but exactly evaluated. The data distribution $_{data}$ is chosen Gaussian, corresponding to (a) $^{(iso)}_{data}$, (b) $^{(heterosc)}_{data}$ and (c) $^{(corr)}_{data}$.

Performances obtained from raw distributions for $^{(iso)}_{data}$, $^{(heterosc)}_{data}$ and $^{(corr)}_{data}$ are displayed in Figure 14. In the isotropic case (Figure 14 (a)) the curve for the upper bound (blue line) points a global minimum near the minimal values obtain by $W_2(\ _{data}\ ; b_N^{(\ a\ ;\ )})$ (plain orange line), which underlines

that the upper bound is indeed informative in such a case. However, the upper bounds obtained for the heteroscedastic and correlated settings (Figure 14 (b,c)) are not in line with the generation results.

These observed discrepancies can be linked to the conditioning of the covariance matrices. In both heteroscedastic and correlated cases, the largest eigenvalue of the covariance matrices is not smaller than the variance stationary distribution of the forward process (set to $^2 = 1$ in those experiments) violating the requirements of Lemma 4.1 ($_{max}^{(heterosc)} = 1$ and $_{max}^{(corr)}$ 15). This induces the default of strong log-concavity of the renormalized densities p_t. In this way, the Gaussian scenario highlights the critical in uence of the covariance matrix conditioning on SGMs. Additionally, a smaller $_{min}()$ would increase L_t and M, which in turn would increase the bound from Theorem 4.2.

Data preprocessing. As frequently done in practice, we expect better conditioning by running SGMs on a standardized distribution. In this way, note that if X_0 $_{data}$ we consider the centered standardized distribution $X_{stand} = D(X_0$) with $D = diag(_1; : : : ; _d) 2 R^{d\ d}$ a diagonal matrix with diagonal entries $_j$ corresponding to the standard deviation of the j-th component of X_0 and with $= [E[X_{0;1}]; :::; E[X_{0;d}]]^>$. A last transformation shrinks the data into a rescaled version of X_{stand} de ned as $X_{scale} = D(X_0$) with $:= 1=(2\ _{max}^{(stand)})^{1=2}$, where $_{max}^{(stand)}$ is the largest eigenvalue of the covariance matrix of X_{stand}. We then train SGMs to approximate the distribution of X_{scale}. By doing so we ensure the applicability of Lemma 4.1 (with $^2 = 1$), as the largest eigenvalue of the covariance matrix of X_{scale} is no larger than 0.5.

Adapted upper bound. We can nally adapt the upper bound from Theorem 4.2 to a rescaled setting by noting that

$$W_2(\ _{data}; \sim) \quad \frac{1}{} \max_{1\ j\ d}\ _j\ W_2\ \ _{scale}; b_{N;\ scale}^{(\ a;\ )}\ ; \tag{51}$$

where

- $\hat{}$ $_{scale}$ is the distribution of scaled sample X_{scale}
- $\hat{}$ $b_{N;\ scale}^{(\ a;\ )}$ corresponds to the distribution of SGM trained on X_{scale}
- $\hat{}$ $\sim$ is the distribution of the descaled generated samples, i.e., the distribution of $D^{\ 1}X = \ +$ with X $b_{N;\ scale}^{(\ a;\ )}$.

Therefore, we can evaluate the upper bound of Theorem 4.2 for scaled samples (r.h.s. of (51)), and transfer it up to a constant to descaled generated samples (l.h.s. of (51)).

Results with scaled data preprocessing. The results are detailed in Figure 15 for the heteroscedastic case (e) $_{scale}^{(heterosc)}$ and the correlated case (f) $_{scale}^{(corr)}$, and are discussed extensively in Section 5.1 of the main paper. Note that the minima of the evaluated bounds now align closely with the empirical metrics. However, the upper bound pro le for the correlated case has been shifted up. This increase was anticipated due to the e ect of rescaling by the largest eigenvalue of $_{(stand)}^{corr}$, approximately 15, which reduces the magnitude of the values in $_{scale}^{(corr)}$. This tends to increase the values of L_t and M through the e ect on $_{min}(\ _{(stand)}^{(corr)})$ as explained above. Despite this e ect, these experiments con rm the overall utility of the bound for selecting the appropriate noise schedule. The e ect of

data rescaling on the Lispschitz continuity and log concavity of the true score function $r \log p_t$ are illustrated in Figure 16 on the Heteroscedastic setting.

(d) Heteroscedastic rescaled setting (e) Correlated rescaled setting

Figure 15: Comparison of the empirical 2-Wasserstein distance on rescaled datasets for (d) $\Sigma_{\text{scale}}^{(\text{heterosc})}$, (e) $\Sigma_{\text{scale}}^{(\text{corr})}$.

Optimal schedule versus classical choices. We investigate the gain from using SGM with the schedule $a^\star$ minimising the upper bound from Theorem 4.2 for $d \in \{5, 10, 25, 50\}$ compared to the linear and cosine schedules (see Appendix E.1.3). To determine the optimal value $a^\star$, upper bounds were initially calculated across various dimensions for a range of a values from $\{-10, -9, \dots, 10\}$. This initial calculation aimed to identify a preliminary minimum value. Subsequently, the search was refined around these preliminary values using finer step-sizes of 0.25 to more precisely locate $a^\star$.

Results for the isotropic, heteroscedastic, and correlated cases are presented in both tabular form in Table 2 and visually in Figure 3 within the main paper. These findings are discussed in Section 5.1 of the main paper.

E.2.3 Numerical experiments on more complex synthetic data

In the context of complex data distributions, the Kullback-Leibler bound (Theorem 3.1) appears to be of limited practical applicability. Specifically, $E_2^{KL}(\cdot, \cdot)$ implies that for each noise schedule tested, a distinct score approximation $s_\cdot(t, x)$ must be trained. This requirement renders the bound computationally intensive and therefore not realistically usable. Additionally, $E_3^{KL}(\cdot)$ is independent of the schedule choice over $(0, T)$, as it depends solely on its final value $\gamma(T)$ which is set constant in our empirical setting (for all a, $\gamma_a(T) = 20$). As a consequence, the last remaining error term to analyse the bound through the lens of noise schedules is the mixing time $E_1^{KL}(\cdot)$. However, relying exclusively on $E_1^{KL}(\cdot)$ would suggest selecting a schedule $t \mapsto \gamma(t)$ that maximises $\int_0^T \gamma(t)dt$. As demonstrated in Section 5.1, this approach clearly fails to yield the schedule choices near the optimal solution.

Therefore, a more reliable choice would be to use the W_2 bound of Theorem 4.2 for which most of the terms can be computed explicitly with reasonable computational cost in the Gaussian setting.

	Dimension	5	10	25	50
Isotropic	Upper bound min $a^?$	4.5	4.25	3.75	4.25
	Generation value in $a^?$	0.039241 0.012572	0.059274 0.009438	0.130829 0.014245	0.233812 0.010584
	VPSDE (linear sched.)	0.036995 0.004663	0.063939 0.010876	0.141601 0.020447	0.256384 0.032709
	Cosine schedule	0.030996 0.003254	0.060649 0.007117	0.131234 0.004794	0.251959 0.005588
	% gain (vs VPSDE)	-6.07 %	+7.30 %	+7.61 %	+8.79 %
	% gain (vs Cosine)	-26.60 %	+2.26 %	+0.31 %	+7.20 %
Heterosc. (with rescaling)	Upper bound min $a^?$	4.00	3.25	2.00	2.75
	Generation value in $a^?$	0.096592 0.003062	0.143224 0.004899	0.242493 0.004769	0.372292 0.004694
	VPSDE (linear sched.)	0.098889 0.003604	0.147478 0.009638	0.249144 0.011394	0.385612 0.009333
	Cosine schedule	0.096437 0.002380	0.143701 0.002460	0.250520 0.004448	0.374868 0.003243
	% gain (vs VPSDE)	+2.32 %	+2.89 %	+2.67 %	+3.46 %
	% gain (vs Cosine)	-0.16 %	+0.33 %	+3.20 %	+0.69 %
Correlated (with rescaling)	Upper bound min $a^?$	8.00	8.75	10.50	11.00
	Generation value in $a^?$	0.066548 0.013873	0.107291 0.028454	0.261075 0.029533	0.676151 0.123277
	VPSDE (linear sched.)	0.072068 0.019861	0.138240 0.031119	0.302986 0.045539	0.897584 0.079860
	Cosine schedule	0.048276 0.008605	0.112898 0.011284	0.391753 0.030112	0.765524 0.022376
	% gain (vs VPSDE)	+7.65 %	+22.36 %	+13.81 %	+24.68 %
	% gain (vs Cosine)	-37.77 %	+4.96 %	+33.31 %	+11.67 %
Parameters	Learning rate	1e-4	1e-4	1e-3 (1e-4 for Corr.)	1e-3 (1e-4 for Corr.)
	Epochs	20	30	75	150

Table 2: Comparison of the W_2 distance between the target value and the generated value at $a^?$ (the minimum value of the upper bound from Theorem 4.2) with the W_2 distance between the generated value by VPSDE and the target distribution. We display averages plus or minus standard deviations over 10 runs. The target distributions are chosen to be Gaussian with different covariance structures: isotropic, heteroscedastic (with rescaling applied), and correlated (with rescaling applied).

Figure 16: Comparison of the ratio strong concavity / Lipschitz continuity for the true score function $r \log p_t$ in the Heteroscedastic setting before rescaling (Figure 14 (b)) and after rescaling (Figure 15 (d)) throughout diffusion time $t \in (0; T]$.

In particular, we leverage the Gaussian framework to estimate the constant terms and apply the rescaling defined in Appendix E.2.2 to ensure that C_t is non negative for $t \in (0; T]$. More precisely,

- $\hat{L}_t$ and $\hat{C}_t$ are given in Lemma 4.1 and are computed using the empirical covariance matrix associated with π_{scale} (and using when applicable the refinements in Propositions D.1 and D.2),
- $\hat{M}$ is derived with Proposition E.5 with appropriate empirical estimators,
- $W_2(\pi_{\text{data}}, \pi_1)$ is computed using closed-form formulas for Gaussian distributions, involving empirical estimators of the mean and covariance of π_{scale},
- the term ε is deliberately omitted to avoid the prohibitively high computational costs associated with training distinct models for different noise schedules.

The experiments are run using the same neural network architecture as in the Gaussian illustrations of Appendices E.2.1 and E.2.2 (i.e., a dense neural network with 3 hidden layers of width 256). The network was trained over 200 epochs for $a \in \{-10, -9, \dots, 19, 20\}$. Contrary to the Gaussian case, conditional score matching $L_{\text{score}}(\theta)$ (5) is used, as being closer to what is done in practice (explicit scores are now out of reach). To assess the quality of the data generation three metrics are used:

(a) an estimator of the KL-divergence based on k-nearest neighbors (Wang et al., 2009) with $k = 100$ neighbors,

(b) the sliced 2-Wasserstein distance (Flamary et al., 2021) with 2000 projections,

(c) the negative log likelihood computed on 1000 samples defined as $\frac{1}{1000}\sum_{i=1}^{1000} \log \pi_{\text{data}}(x_i)$ with $(x_i)_{1 \le i \le 1000}$ samples from the generated distribution and π_{data} the probability density function to be estimated.

Funnel distribution. The first distribution considered is the Funnel distribution (Thin et al., 2021) in dimension 50, defined as

$$\pi_{\text{data}}(x) = \mathcal{N}\left(a^2(x_1)\right) \prod_{j=2}^{d} \mathcal{N}\left(\exp(2bx_1)(x_j)\right),$$

with $a = 1$ and $b = 0.5$. To ensure the applicability of Theorem 4.2 and Lemma 4.1 the samples are standardized and rescaled according to the method described in Appendix E.2.2. The results, illustrated in Figures 4 and 17, show that the upper bounds effectively mirror the generation outcomes across the three metrics considered. Moreover, the generation results for the parametric schedule a^2 (the one that minimizes the upper bound) outperforms in all three metrics both the linear and cosine schedules (see Table 3).

Gaussian mixture models. The second distribution considered is a Gaussian mixture model with 25 modes in dimension 50, defined as

$$\pi_{\text{data}}(x) = \frac{1}{25} \sum_{(j,k) \in \{-2, \dots, 2\}^2} \mathcal{N}_{\mu_{jk}, \Sigma_d}(x)$$

with $\mathcal{N}_{\mu_{jk}, \Sigma_d}$ denoting the probability density function of the Gaussian distribution with covariance matrix $\Sigma_d = \text{diag}(0.01, 0.01, 0.1, \dots, 0.1)$ and mean vector $\mu_{jk} = [j, k, 0, 0, \dots, 0]^\top$. The results shown in Figure 18 and Table 3 confirm the relevance of the upper bound even for non-Gaussian datasets.

(b) KL divergence with k-nearest neighbors estimate (c) Negative log-likelihood

Figure 17: Upper bound and empirical distances between the data distribution and the generated samples for different metrics on a Funnel dataset in dimension 50.

(a) Sliced-W_2 distance (b) k-nearest neighbors estimator (c) Negative log-likelihood

Figure 18: Upper bound and empirical distances between the data distribution and the generated samples for different metrics on a mixture of 25 Gaussian variables dataset in dimension 50.

E.3 Numerical experiments on real-world datasets

To evaluate the impact of the noise schedule on the performance of score-based generative models we evaluate the parametric family $_a$ introduced in Equation (9) using CIFAR 10 dataset. We suggest to analyse the FID (Fréchet Inception Distance) score on 50 000 samples generated for different noise schedules (different values of a in $_a$, see Figure 1) on CIFAR 10.

We use pretrained models from Karras et al. (2022) with the recommended hyperparameters designed to replicate the experiments in Song et al. (2021) corresponding to our linear schedule ($a = 0$) as shown in Figure 1. In particular, we let $T = 1$, $(0) = 0:1$, $(T) = 20$, 1000 discretization steps and sample over the diffusion $[\ ; 1]$ with $= 10^{-3}$.

The training process in Karras et al. (2022) is slightly different, though equivalent, to the original implementation. In particular, the networks are not trained to directly estimate $r \log p_t(X_t)$. Instead, a denoiser function $D(X; \ )$ is trained to isolate the noise from the signal for some noise level (see Equations (2) and (3) in Karras et al. (2022)). With appropriate rescaling this denoiser

	Metric	Sliced-Wasserstein		k-nn (Kullback-Leibler)		NLL	
Funnel distribution	Generation value in a?	0.218498	0.049882	4.242455	0.450224	82.25179	3.12809
	VPSDE (linear sched.)	0.240664	0.036578	6.048403	0.726221	87.02893	3.40642
	Cosine schedule	0.221851	0.054309	4.927209	0.510968	83.73294	3.53262
	% gain (vs VPSDE)	+9.21 %		+29.88 %		+5.49 %	
	% gain (vs Cosine)	+1.51 %		+13.91 %		+1.77 %	
Gaussian mixture models	Generation value in a?	0.043388	0.005222	2.433759	0.180652	35.033176	1.97863
	VPSDE (linear sched.)	0.057763	0.004450	3.063054	0.126697	40.49867	3.13705
	Cosine schedule	0.046816	0.008402	2.541213	0.158563	34.76353	2.20980
	% gain (vs VPSDE)	+24.91 %		+20.55 %		+13.49 %	
	% gain (vs Cosine)	+7.32 %		+4.23 %		-0.77 %	
Parameters	Learning rate	1e-3		1e-3		1e-3	
	Epochs	200		200		200	

Table 3: Comparison of the sliced W_2 distance, KL divergence coupled with k-nearest neighbors estimate and negative log-likelihood between the target distribution and the SGM-generated one. For the latter, the SGM is either trained with linear, cosine and $a^?$ schedules. We display averages plus or minus standard deviations over 10 runs. The target distributions are chosen are Funnel and Gaussian mixture models.

can be used in the VP setting by letting

$$s^{\star}(X_t, t) = \frac{\sigma_t^2}{m_t} \quad D\left(\frac{X_t}{m_t}; \frac{\sigma_t}{m_t}\right) \quad \frac{X_t}{m_t} \quad ;$$

where $s^{\star}$ is the score approximation as defined in our paper, $m_t = \exp\left(\int_0^t (s)ds/(2\ ^2)\right)$ and $\sigma_t^2 = \ ^2(1 \quad m_t^2)$. This formulation bridges the denoising approach with score-based methods in the VP framework.

Figure 19 displays the FID score for samples generated using the Euler-Maruyama discretization of the backward process for different choices of $_a$ with $a \in \{\ 10,\ \ 9, \ldots, 10\}$ and cosine schedule $_{\cos}$. Although the assumptions of our results cannot be verified in such a setting, it is interesting to note that the empirical performance follows the same dynamics as in the toy numerical experiments. This indicates that the analysis and optimization of noise schedules is an interesting problem to be explored further for complex cases.

F Conditional training in the Gaussian setting

Section 5.1 of this paper is dedicated to the illustration of the theoretical upper bounds and their relevance in the Gaussian setting (i.e., when $_{\text{data}}$ is Gaussian). This choice has been motivated by the fact that, under this setting, all constants in the upper bounds from Theorem 3.1 and Theorem 4.2 are either analytically available or could be precisely estimated (see Appendices E.2.1 and E.2.2).

In particular, both upper bounds display error terms proportional to $L_{\text{explicit}}(\ )$ (4), which has motivated the use of explicit score matching during the training. To do so, we used a deep neural architecture (see Figure 11) trained to minimize $L_{\text{explicit}}(\ )$ (4) using as a target the true score function. This is possible because in the Gaussian setting, the true score function is analytically known (Lemma E.1). However, in most applications the score function is not available, because

Figure 19: FID scores on 50 000 generated samples using pre-trained models from Karras et al. (2022) with different noise schedules from the parametric family (9) and cosine schedule.

the data distribution is not known and has to be learned. This is the reason why, in practice we rely on conditional score matching (i.e., the minimization of $L_{\text{score}}(\ )$ (5)). This approach is particularly relevant given the relationship between the explicit and conditional score functions: $L_{\text{explicit}}(\ ) = L_{\text{score}}(\ ) \quad E\ k\nabla \log p_t(X\ ) - \nabla \log p_t(X\ |X_0)k^2$:

Consequently, all the theoretical upper bounds discussed in Sections 3 and 4 can be adjusted by a constant (with respect to) to account for discrepancies between the score function learned through L_{score} or L_{explicit} .

The rest of this section demonstrates the numerical effects of employing conditional score matching instead of explicit score matching, following the numerical set-up of Appendices E.2.1 and E.2.2. In Figure 20, the Kullback-Leibler upper bound from Theorem 3.1 is depicted in varying shades of blue, while the empirical $KL(\ _{\text{data}}\ jj\ b_N^{(\ a;\ )})$ across parameters $a \in f\ 10;\ \ 9;\ \ 8; :::; 10g$ is shown in varying shades of orange.

In Figure 20, three learning scenarios are presented: one using explicit score matching (which exactly matches the results of Figure 2 (top)), another with conditional score matching over 150 epochs, and a third with 300 epochs. Both the generation results and the upper bounds show diminished performance as the curves are shifted upwards. Nonetheless, the overall curve shapes are similar, and the optimal points remain closely aligned. Interestingly, both the upper bounds and the generation outcomes in the conditional scenarios demonstrate more pronounced peaks near the minimum values. This suggests that precise noise schedule selection may yield even better performance gain when SGMs are trained using conditional score matching.

Additionally, Figure 21 demonstrates that increasing the number of training iterations when using conditional score matching provides results more and more similar to that obtained with explicit score matching. This effect is noticeable in both the KL divergence and the W_2 distance.

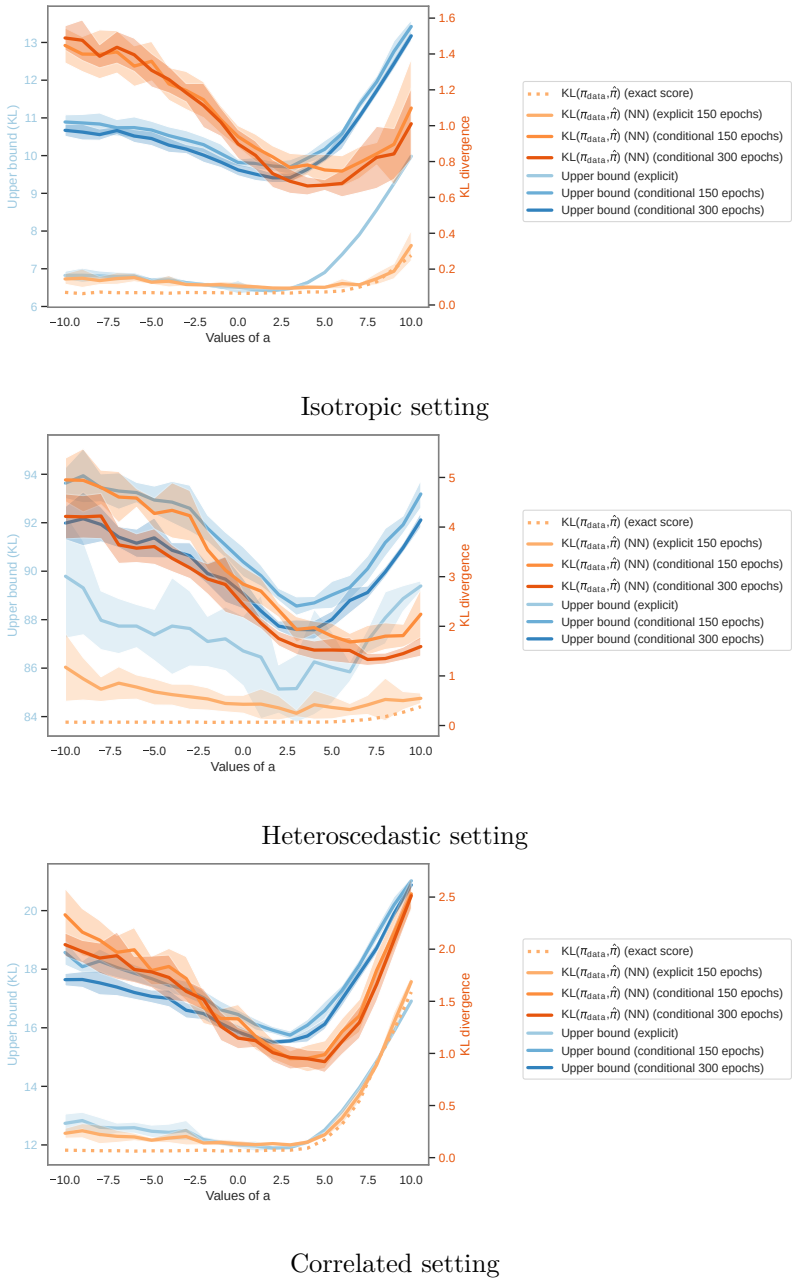

Figure 20: Comparison of the empirical KL divergence (mean value ± std over 10 runs) between π_{data} and $\overline{\pi}_N^{(\beta_a,\theta)}$ (in orange) and the upper bound of Theorem 3.1 (in blue) w.r.t. the parameter a used in the definition of the noise schedule β_a, for $d = 50$.

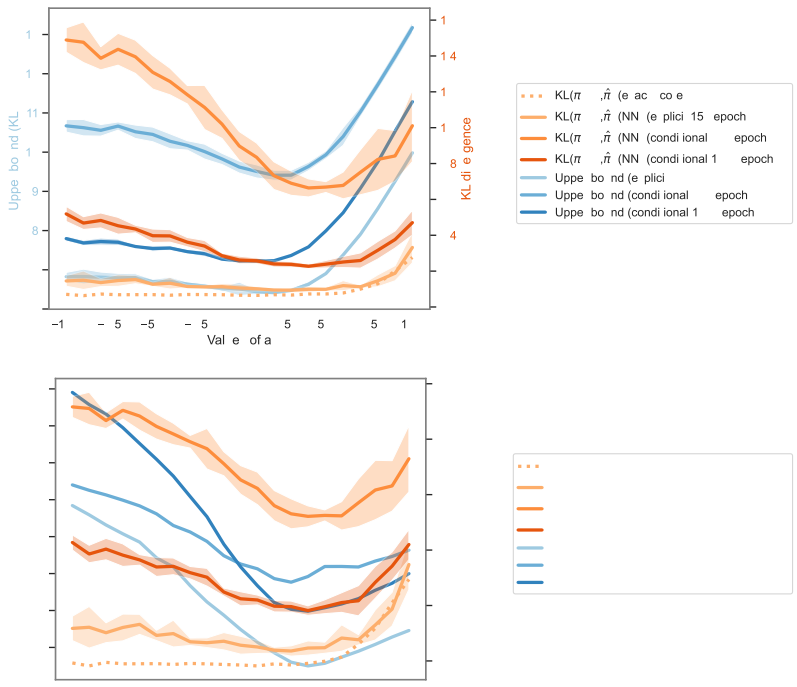

Figure 21: Comparison of the empirical KL divergence (top) and the W_2 distance (bottom) (mean value ± std over 10 runs) between $\pi_{\text{data}} = \pi_{\text{data}}^{(\text{iso})}$ and $\hat{\pi}_N^{(\beta_a,\theta)}$ (in orange) and the upper bound of Theorem 3.1 (top) and of Theorem 4.2 (bottom) (in blue) w.r.t. the parameter a used in the definition of the noise schedule β_a, for $d = 50$.