On Forgetting and Stability of Score-based Generative models

Anonymous Authors[1]

Abstract

Understanding the stability and long-time behavior of generative models is a fundamental problem in modern machine learning. This paper provides quantitative bounds on the sampling error of score-based generative models by leveraging stability and forgetting properties of the Markov chain associated with the reverse-time dynamics. Under weak assumptions, we provide the two structural properties to ensure the propagation of initialization and discretization errors of the backward process: a Lyapunov drift condition and a Doeblin-type minorization condition. A practical consequence is quantitative stability of the sampling procedure, as the reverse diffusion dynamics induces a contraction mechanism along the sampling trajectory. Our results clarify the role of stochastic dynamics in score-based models and provide a principled framework for analyzing propagation of errors in such approaches.

1. Introduction

Score-based generative models have recently emerged as a unifying framework for high-dimensional generative modeling, with remarkable empirical success across a wide range of practical problems. They are rooted in the framework of denoising diffusion probabilistic models (DDPMs) (see, e.g., Ho et al., 2020; Song et al., 2021; 2022), and of score-matching techniques introduced in Hyvärinen & Dayan (2005); Vincent (2011). These models can be interpreted as defining stochastic samplers whose associated Markov semigroups are designed to transform a simple reference distribution into a complex target distribution through a sequence of score-driven updates. This viewpoint naturally raises fundamental questions regarding the stability and convergence of the resulting sampling procedures.

Theoretical analysis of score-based generative models has been explored, in particular in Wasserstein distances, with non-asymptotic bounds highlighting initialization and discretization errors, along with score matching training errors. Following first guarantees in Wasserstein distances, such as (Chen et al., 2022), most recent works use strong assumptions of the target distributions, such as log-concavity (Bruno & Sabanis, 2025; Strasman et al., 2025a; Yu & Yu, 2025), bounded support (Beyler & Bach, 2025) or regularity conditions on the score function (Tang & Zhao, 2024; Gao et al., 2025a). Some recent works have moved beyond the traditional constraints of log-concavity in data distributions and regularity assumptions for the score function and proposed relaxed convexity assumptions (Gentiloni-Silveri & Ocello, 2025; Strasman et al., 2025b). In the most recent results, the data distribution is modeled as a perturbed strongly log-concave law and use one-sided Lipschitz conditions which ensure that the taget distribution has sub-Gaussian tails, and consequently, all its polynomial moments are finite, (Brigati & Pedrotti, 2025; Stéphanovitch, 2025).

Contributions. These works do not leverage the forgetting properties of Markov chains to characterize how initialization, score-approximation, and discretization errors propagate along the sampling trajectory. In this paper, we propose an analysis grounded in the Harris stability framework for Markov processes (Hairer & Mattingly, 2008), which provides quantitative, exponential decay of the dependence on the initial condition. We show that the backward Markov chain satisfies a Lyapunov drift condition and a localized minorization condition. Following (Hairer & Mattingly, 2008), these properties yield contraction of the associated Markov semigroup in a Wasserstein-type sense. A key in our work is a dissipativity condition on the score function outside a compact set. Notably, this holds *without early stopping*, and yields explicit constants that quantitatively transfer structural assumptions on the data score to the diffusion flow. Such dissipativity conditions are standard in the sampling and optimization literature and have been used extensively to establish stability and convergence of Langevin-type dynamics (Meyn & Tweedie, 2009; Löcherbach, 2015). Importantly, this requirement is weaker than the usual assumptions on the target distribution that underpins much of the existing theoretical analysis of diffusion-based samplers (Meyn & Tweedie, 1993; 2012).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Moreover, our assumptions do not impose global Lipschitz continuity of the score function, nor any form of one-sided Lipschitz regularity. They allow for polynomial growth of the Jacobian of the score, thereby accommodating target distributions with polynomially super-Gaussian tails. As a result, the proposed framework covers nonconvex and multimodal target distributions.

2. Background

2.1. Notations

We denote by $\mathbf{I}_d$ the $d \times d$ identity matrix, by $\dot{f}$ the derivative of a function f, and by $\mathcal{P}_p(\mathbb{R}^d)$ the set of probability distributions μ on $\mathbb{R}^d$ such that $\mathbb{E}_\mu [\|\mathbf{X}\|^p] < \infty$. For a vector $v \in \mathbb{R}^d$, we write $v^{\otimes 2} := vv^{\mathrm{T}}$. We write $X \overset{\mathcal{L}}{=} Y$ when the random variables X and Y have the same distribution. We denote by $\mathcal{C}^k\left(\mathbb{R}^{d_0}; \mathbb{R}^{d_1}\right)$ the space of k-times continuously differentiable functions from $\mathbb{R}^{d_0}$ to $\mathbb{R}^{d_1}$. When $d_1 = 1$, we simplify the notation to $\mathcal{C}^k\left(\mathbb{R}^{d_0}\right)$. When there is no ambiguity, $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^d$, and $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm for matrices.

Score-based generative models. Score-based generative models (SGMs) have emerged as a flexible framework for sampling from high-dimensional probability distributions using diffusion processes and a time-reversal argument to bridge the data distribution $\pi_{\mathrm{data}} \in \mathcal{P}(\mathbb{R}^d)$ to a Gaussian distribution $\pi_\infty \in \mathcal{P}(\mathbb{R}^d)$. The forward process is known as the noising process and is solution to the following stochastic differential equation (SDE) on a fixed time horizon $t \in [0, T]$: $\overrightarrow{\mathbf{X}}_0 \sim \pi_{\mathrm{data}}$ and

$$\mathrm{d}\overrightarrow{\mathbf{X}}_t = -\alpha\beta_t \overrightarrow{\mathbf{X}}_t \mathrm{d}t + \sqrt{2\beta_t}\mathrm{d}\,\mathrm{B}_t\,, \tag{1}$$

where $(\mathrm{B}_t)_{t\in[0,T]}$ is a d-dimensional Brownian motion, $\alpha \geq 0$ and $[0, T] \to \beta_t \in \mathbb{R}_+$ is a non-decreasing noise schedule. For all $0 \leq t \leq T$, p_t denotes the probability density function of the random vector $\overrightarrow{\mathbf{X}}_t$. Specific choices of (α, β) recover either the *Variance Exploding (VE)* formulation with $\alpha = 0$ with $\beta_t = \sigma_{0|t}\dot{\sigma}_{0|t}$ and $\sigma^2_{0|t} = 2\int_0^t \beta_s \mathrm{d}s$ (see, *e.g.*, Song & Ermon (2019)) or the Variance Preserving (VP) with $\alpha = 1$ (Sohl-Dickstein et al., 2015; Ho et al., 2020). More broadly, throughout this work we refer to the case $\alpha > 0$ as *Variance Preserving (VP)*: the forward diffusion (1) is an Ornstein–Uhlenbeck process with stationary distribution $\pi_\infty \sim \mathcal{N}\left(0, \alpha^{-1}\mathbf{I}_d\right)$ (Strasman et al., 2025a).

The time-reversal of (1) admits a diffusion representation (Haussmann & Pardoux, 1986)

$$(\overleftarrow{\mathbf{X}}_t)_{t\in[0,T]} \overset{\mathcal{L}}{=} (\overrightarrow{\mathbf{X}}_{T-t})_{t\in[0,T]}\,,$$

where $(\overleftarrow{\mathbf{X}}_t)_{t\in[0,T]}$ is called the *backward process* (or *reverse-time process*) and is defined as the solution to the SDE: $\overleftarrow{\mathbf{X}}_0 \sim p_T$[1] and

$$\mathrm{d}\overleftarrow{\mathbf{X}}_t = \left(\alpha\bar{\beta}_t\overleftarrow{\mathbf{X}}_t + 2\bar{\beta}_t\;\mathrm{S}_{T-t}\left(\overleftarrow{\mathbf{X}}_t\right)\right)\mathrm{d}t + \sqrt{2\bar{\beta}_t}\mathrm{d}\,\mathrm{B}_t\,, \tag{2}$$

with $\mathrm{S}_{T-t}(\mathbf{x}) := \nabla \log p_{T-t}(\mathbf{x})$ and $\bar{\beta}_t := \beta_{T-t}$. We denote by $(\mathrm{Q}_{t|s})_{0\leq s\leq t\leq T}$ the Markov semigroup associated with the backward diffusion (2), defined for any bounded measurable function f and $\mathbf{x} \in \mathbb{R}^d$ by

$$\mathrm{Q}_{t|s}f(\mathbf{x}) := \mathbb{E}\left[f(\overleftarrow{\mathbf{X}}_t)\middle|\overleftarrow{\mathbf{X}}_s = \mathbf{x}\right]. \tag{3}$$

Backward process approximations. Sampling from π_{data} amounts to simulating the marginal $\overleftarrow{\mathbf{X}}_T$ starting from $\overleftarrow{\mathbf{X}}_0 \sim p_T$ as $\pi_{\mathrm{data}} = p_T\mathrm{Q}_T$[2]. In practice, to turn this identity into a generative model, three approximations are required, which we now describe.

1. **Mixing time error**. The distribution p_T is given by a Gaussian convolution of π_{data} and is therefore typically intractable (see Lemma C.1). It is usually replaced by a simple reference distribution π_∞. In the VE case, we usually take $\pi_\infty = \mathcal{N}\left(0, \sigma^2_{T|0}\mathbf{I}_d\right)$, while in the VP case we take the stationary distribution of (1) $\pi_\infty = \mathcal{N}\left(0, \alpha^{-1}\mathbf{I}_d\right)$. This approximation is independent of π_{data} and induces an *initialization bias*.
2. **Score approximation**. The drift term of (2) involves the data-dependent score function $\mathrm{S}_t(\mathbf{x})$, which is, generally intractable. It is approximated using a neural network $\mathrm{s}_\theta : (0, T] \times \mathbb{R}^d \mapsto \mathbb{R}^d$ parameterized by $\theta \in \Theta$, and trained to minimize the conditional score matching loss (Vincent, 2011):
$$\mathcal{L}(\theta) = \mathbb{E}\left[\left\|\mathrm{s}_\theta\left(\overrightarrow{\mathbf{X}}_\tau, \tau\right) - \nabla\log p_{\tau|0}\left(\overrightarrow{\mathbf{X}}_\tau|\overrightarrow{\mathbf{X}}_0\right)\right\|_2^2\right],$$
where τ is uniformly distributed over $(0, T]$ and independent of $\overrightarrow{\mathbf{X}}_0$, and $\overrightarrow{\mathbf{X}}_\tau \sim p_{\tau|0}\left(\cdot|\overrightarrow{\mathbf{X}}_0\right)$.
3. **Discretization error**. The transition kernels $\mathrm{Q}_{t|s}$ of the backward diffusion do not admit closed-form expressions in general and must be approximated numerically. Since (2) is time-inhomogeneous through the scalar schedule $u \mapsto \bar{\beta}_u$, we discretize it using a *time-changed Euler–Maruyama scheme* with step sizes $\Delta_k := \int_{t_k}^{t_{k+1}} \bar{\beta}_u\,\mathrm{d}u$. This choice exactly integrates the noise schedule so that the scheme matches exactly the covariance of the Brownian increment over each step.

Combining the three approximations above yields a discrete-time Markov chain $(\bar{\mathbf{X}}^\theta_{t_k})_{0\leq k\leq N}$ over a finite discretization

[1] With abuse of notation we identify probability distribution with their density with respect to the Lebesgue measure.

[2] With abuse of notation we refer, for any $t > 0$ to Q_t as $\mathrm{Q}_{t|0}$.

grid $\mathcal{T}_N = (t_0, \cdots, t_N)$, for $N \in \mathbb{N}^*$, defined by $\bar{\mathbf{X}}^\theta_{t_0} \sim \pi_\infty$ and for $k = 0, \ldots, N-1$, by the recursion

$$\bar{\mathbf{X}}^\theta_{t_{k+1}} = \bar{\mathbf{X}}^\theta_{t_k} + \Delta_k \Big(\alpha \bar{\mathbf{X}}^\theta_{t_k} + 2\, \mathrm{s}_\theta \left(\bar{\mathbf{X}}^\theta_{t_k}, T - t_k \right) \Big) + \sqrt{2\Delta_k}\xi_k \,, \quad (4)$$

where $(\xi_k)_{k \geq 0}$ is an i.i.d. sequence of standard Gaussian random vectors in $\mathbb{R}^d$. For each $k \in \{0, \ldots, N-1\}$, we denote by $\mathrm{Q}^\theta_{t_{k+1}|t_k}$ the one-step Markov kernel of the discrete-time scheme (4), defined for any bounded measurable function f and $\mathbf{x} \in \mathbb{R}^d$, by

$$\mathrm{Q}^\theta_{t_{k+1}|t_k} f(\mathbf{x}) := \mathbb{E}\left[f(\bar{\mathbf{X}}^\theta_{t_{k+1}}) \middle| \bar{\mathbf{X}}^\theta_{t_k} = \mathbf{x} \right] .$$

Generation error analysis. Fix a discretization grid $\mathcal{T}_N = (t_0, \ldots, t_N)$ of $[0, T]$ and for integers $0 \leq k < \ell \leq N$, define the composed SGM kernel as $\mathbb{Q}^\theta_{k:\ell}$ by

$$\mathbb{Q}^\theta_{k:\ell} f(\mathbf{x}) := \int f(\mathbf{x}_\ell) \prod_{r=k}^{\ell-1} \mathrm{Q}^\theta_{t_{r+1}|t_r}(\mathbf{x}_r, \mathrm{d}\mathbf{x}_{r+1}) \,,$$

with $\mathbf{x}_k := \mathbf{x} \in \mathbb{R}^d$ and f a bounded measurable function. We denote the resulting generated distribution of the SGM at time t_k

$$\hat{\pi}^\theta_k := \pi_\infty \mathbb{Q}^\theta_{0:k} \quad (5)$$

and keep the dependency on $\mathcal{T}$ implicit. Our goal is then to compare the target distribution π_{data} with the distribution $\hat{\pi}^\theta_N$ produced by the SGM.

3. Forgetting of the backward process

Harris-type stability and weighted total variation distance. Our analysis relies on a classical stability framework for Markov processes known as Harris theory. At a high level, Harris-type results establish quantitative exponential forgetting of the initial condition by combining two ingredients: (i) a *Lyapunov drift condition* ensuring that the dynamics is pulled back toward a central region of the state space, and (ii) a *localized minorization (Doeblin-type) condition* ensuring a uniform mixing component when the process visits that region. Together, these two properties yield contraction of the Markov semigroup in a sense that we make precise below.

To quantify the forgetting property, we work with the weighted total variation distance $\rho_b(\cdot, \cdot)$ defined, for $\mu_1, \mu_2 \in \mathcal{P}(\mathbb{R}^d)$, by

$$\rho_b(\mu_1, \mu_2) := \int_{\mathbb{R}^d} \left(1 + b\, \mathrm{V}_2(\mathbf{x})\right) |\mu_1 - \mu_2|(\mathrm{d}\mathbf{x}) \quad (6)$$

where $\mathrm{V}_2(\mathbf{x}) := \|\mathbf{x}\|^2$, $b > 0$, and $|\mu_1 - \mu_2|$ denotes the total variation measure of the signed measure $\mu_1 - \mu_2$. This metric is standard in Harris-type theorems for unbounded state spaces (Hairer & Mattingly, 2008). Moreover, any contraction estimate proved in $\rho_b(\cdot, \cdot)$ yields quantitative guarantees in both total variation and Wasserstein distance. In particular, for any $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$, we have (see Lemma B.3)

$$\|\mu_1 - \mu_2\|_{\mathrm{TV}} \leq \frac{1}{2} \rho_b(\mu_1, \mu_2) \,,$$

and

$$\mathcal{W}_2^2(\mu_1, \mu_2) \leq \frac{2}{b} \rho_b(\mu_1, \mu_2) \,.$$

Assumptions on the data distribution. To establish forgetting of the backward Markov chain (2), we require mild assumptions on the data. The data distribution π_{data} admits a density $p_0 \in \mathcal{C}^2\left(\mathbb{R}^d\right)$ with respect to the Lebesgue measure. Moreover, p_0 satisfies the following regularity and Lyapunov-type conditions.

Assumption 3.1. There exist constants $\gamma_0 > \alpha/2$, $\kappa_0 \geq 0$, $\mathrm{C}_0 > 0$, and $p \geq 1$ such that the following assumptions hold:

(I) For all $\mathbf{x} \in \mathbb{R}^d$,

$$\left\langle \mathrm{S}_0\left(\mathbf{x}\right), \mathbf{x} \right\rangle \leq -\gamma_0 \left\|\mathbf{x}\right\|^2 + \kappa_0 \,.$$

(II) For all $\mathbf{x} \in \mathbb{R}^d$,

$$\left\|\nabla \mathrm{S}_0\left(\mathbf{x}\right)\right\|_{\mathrm{F}} \leq \mathrm{C}_0 \left(1 + \left\|\mathbf{x}\right\|^p\right) .$$

Assumption 3.1 imposed on the initial distribution play a crucial role in the proof of the main theorem. In particular, Assumption 3.1 (I) ensures that the target distribution π_{data} exhibits sub-Gaussian tail behavior (Lemma A.1), which in turn guarantees sufficient integrability and moment bounds. Note that the condition $\gamma_0 > \alpha/2$ is not restrictive in practice as it simply amounts to calibrating the stationary distribution of Equation (1). Moreover, Assumption 3.1 (II) provides a control on the growth of the score function $\mathrm{S}_0\left(x\right) = \nabla \log p_0(x)$ as well as on the quantity $\nabla \mathrm{S}_0\left(x\right) + \mathrm{S}_0\left(x\right)^{\otimes 2}$, which plays a central role in the control of the Froebenius norm of the score function for any $t \in [0, T]$ (Corollary D.5).

Comparison with recent literature. Assumption 3.1 (I) requires a coercive behavior of the score function $\mathrm{S}_0\left(\mathbf{x}\right)$ outside a compact set. This assumption is commonly referred to as a *dissipativity condition* and is standard in the sampling and optimization literature (see, e.g., Eberle, 2016; Raginsky et al., 2017; Zhang et al., 2017; Erdogdu & Hosseinzadeh, 2021; Erdogdu et al., 2022; Vacher et al., 2025). It relaxes global strong convexity of the potential function $x \mapsto -\log p_0(x)$ (Durmus & Moulines, 2017; Dalalyan &

Karagulyan, 2019; Strasman et al., 2025a) therefore cover highly nonconvex and multimodal distributions, which fall outside the scope of works relying on global convexity of the potential. In particular, this includes mixtures of Gaussian distributions, which are shown to satisfy this condition in Section A of Gentiloni-Silveri & Ocello (2025) and Section 3.2 of Vacher et al. (2025), covering a broad and practically relevant class of target distributions.

Moreover, our assumptions do not require global Lipschitz continuity of the score, nor any one-sided Lipschitz condition; instead, Assumption 3.1 (II) allows for polynomial growth of the Jacobian $\nabla \mathrm{S}_0(\mathbf{x})$. Consequently, our framework accommodates targets with *polynomially super-Gaussian tails* (e.g., $p(x) \propto \exp(-\|x\|^p)$ with $p > 2$), for which $\|\nabla \mathrm{S}_0(\mathbf{x})\|$ typically diverges as $\|\mathbf{x}\| \to \infty$. These targets that encode strong confinement in some regions of the space are excluded from analyses that impose (one-sided) Lipschitz-type regularity of the score or, equivalently, uniform bounds on $\nabla^2(-\log p)$, as in Chen et al. (2023a;b); Strasman et al. (2025a); Gentiloni-Silveri & Ocello (2025); Gao et al. (2025b). Targets with super-Gaussian tails can be interpreted as enforcing a strong form of regularization: although their support is the whole space, the rapid decay of the density concentrates almost all mass inside a bounded region, effectively mimicking compact support while preserving smoothness.

Lyapunov drift for the backward chain. We prove the first property of Harris-type stability results, the *Lyapunov drift* inequality for the backward Markov semigroup $(\mathrm{Q}_{t|s})_{0\le s\le t\le T}$. Such an estimate prevents trajectories from drifting to infinity by ensuring a quadratic pull toward the center of the state space. A key technical step is to show that the dissipativity of the data score at initial time propagates along the forward diffusion. Under Assumption 3.1 (I), there exist positive continuous functions $t \mapsto \gamma_t$ and $t \mapsto \kappa_t$ such that, for all $t \in [0, T]$ and $\mathbf{x} \in \mathbb{R}^d$,

$$\langle \mathrm{S}_t(\mathbf{x}), \mathbf{x}\rangle \le -\gamma_t \|\mathbf{x}\|^2 + \kappa_t \,. \tag{7}$$

We refer to Proposition D.1 for the precise statement and explicit constants. Combining (7) with the infinitesimal generator of (2) and Dynkin's formula (followed by Grönwall's inequality) yields the following Lyapunov drift inequality.

Proposition 3.2. *Suppose that Assumption 3.1 holds and let $\mathrm{V}_\ell(\mathbf{x}) := \|\mathbf{x}\|^\ell$ for $\ell \ge 2$. Then, there exist continuous functions $\tilde{\gamma}_{\cdot,\ell}, \tilde{\kappa}_{\cdot,\ell} : [0, T] \to \mathbb{R}_+$ such that, for all $0 \le s < t \le T$ and all $\mathbf{x} \in \mathbb{R}^d$,*

$$\mathrm{Q}_{t|s}\,\mathrm{V}_\ell(\mathbf{x}) \le \lambda^{(\ell)}_{t|s}\,\mathrm{V}_\ell(\mathbf{x}) + \mathrm{K}^{(\ell)}_{t|s} \,,$$

where $\lambda^{(\ell)}_{t|s} := \exp\left(-\int_s^t \tilde{\gamma}_{v,\ell}\,\mathrm{d}v\right)$ and $\mathrm{K}^{(\ell)}_{t|s} := \int_s^t \exp\left(-\int_u^t \tilde{\gamma}_{v,\ell}\,\mathrm{d}v\right)\tilde{\kappa}_{u,\ell}\,\mathrm{d}u$.

In particular, for $\ell = 2$,

$$\tilde{\gamma}_{t,2} = 2\bar{\beta}_t\big(2\gamma_{T-t} - \alpha\big) \text{ and } \tilde{\kappa}_{t,2} = 2\,\bar{\beta}_t\big(2\kappa_{T-t} + d\big) \,.$$

Proof. The proof is deferred to Appendix D.2. □

In the sequel, Harris-type contraction is stated with the quadratic Lyapunov function $\mathrm{V}_2(\mathbf{x}) = \|x\|^2$; we nevertheless prove the more general ℓ-moment drift bound since it is useful for moment and integrability estimates.

Minorization for the backward chain. In addition to the Lyapunov drift condition, a second key property in Harris-type arguments is a minorization property, which ensures sufficient mixing of the backward dynamics. In unbounded state-spaces such as $\mathbb{R}^d$, a global Doeblin condition is generally too strong, so one instead localizes it to an appropriate "small" or "petite" set (Meyn & Tweedie, 2009; Löcherbach, 2015). We establish a localized Doeblin condition in Proposition 3.3. This guarantees that, on an appropriate subset, $\mathrm{Q}_{t|s}$ dominates a state-independent measure and hence can forget its initial condition.

Proposition 3.3. *Let $0 \le s < t \le T$ and suppose Assumption 3.1 (I) holds. Fix $r > 0$ and define the set*

$$\mathcal{C}_r := \mathrm{B}_r(0) = \{\mathbf{x} \in \mathbb{R}^d : \mathrm{V}_2(\mathbf{x}) \le r^2\} \,.$$

Then, there exist a probability measure $\upsilon_{s|t}$ on $\mathbb{R}^d$ and a constant $\varepsilon^{(r)}_{t|s} = \tilde{\varepsilon}_{t|s}\exp(-r^2/\sigma^2_{T-s|T-t}) \in (0, 1)$ such that, for all $\mathbf{x} \in \mathcal{C}_r$ and all $\mathrm{A} \in \mathcal{B}(\mathbb{R}^d)$,

$$\mathrm{Q}_{t|s}(\mathbf{x}, \mathrm{A}) \ge \varepsilon^{(r)}_{t|s}\,\upsilon_{t|s}(\mathrm{A}) \,. \tag{8}$$

Proof. The proof is deferred to Appendix E, where we also provide explicit constants. □

Forgetting via Harris contraction. The drift and minorization conditions above imply a quantitative forgetting property of the backward dynamics in the weighted total variation metric (6). Although Harris-type theorems are typically stated for time-homogeneous Markov kernels, we use them here in a *local-in-time* form: for each interval $[s, t]$, the transition kernel $\mathrm{Q}_{t|s}$ satisfies a one-step contraction.

Proposition 3.4. *Fix $0 \le s < t \le T$ and suppose that Assumption 3.1 holds. Set*

$$r^2 > r_c^2 = \frac{2\mathrm{K}^{(2)}_{t|s}}{1 - \lambda^{(2)}_{t|s}} \,,$$

where $\mathrm{K}^{(2)}_{t|s}$ and $\lambda^{(2)}_{t|s}$ are defined in Proposition 3.2. Then, there exist $\bar{\alpha}_{t|s} \in (0, 1)$ and $b^r_{s,t} > 0$ such that for any probability measures μ_1, μ_2 on $\mathbb{R}^d$,

$$\rho_{b^r_{s,t}}(\mu_1\mathrm{Q}_{t|s}, \mu_2\mathrm{Q}_{t|s}) \le \bar{\alpha}_{t|s}\rho_{b^r_{s,t}}(\mu_1, \mu_2) \,.$$

Proof. Under Assumption 3.1 $\mathrm{Q}_{t|s}$ satisfies the drift and minorization properties of Proposition 3.2 and Proposition 3.3 which are Assumption 1 and 2 of Hairer & Mattingly (2008) adapted to time-inhomogenous Markov transition kernels. The conclusion is then an application of Theorem 1.3 of Hairer & Mattingly (2008). □

Proposition 3.4 establishes the contraction of the backward kernel for a range of metrics with different contraction rates. The choice of ideal metric is directly linked with the choice of r. Note that following Hairer & Mattingly (2008), for any $\alpha_0 \in (0, \bar{\alpha}_{t|s})$ and $\eta_0 \in (\lambda^{(2)}_{t|s} + 2\mathrm{K}^{(2)}_{t|s}/r^2, 1)$, we can choose

$$
b^r_{s,t} = \alpha_0/\mathrm{K}^{(2)}_{t|s} ,
$$

$$
\bar{\alpha}_{t|s} = \left[1 - \left(\varepsilon^{(r)}_{t|s} - \alpha_0\right)\right] \wedge \frac{2 + r^2 b^r_{s,t}\eta_0}{2 + r^2 b^r_{s,t}} ,
$$

where $\varepsilon^{(r)}_{t|s}$ is defined in Proposition 3.3. The explicit derivation of r_c^2 and $\varepsilon^{(r)}_{t|s}$ is given in Appendix F for the variance preserving and variance exploding cases.

Explicit Gaussian contraction. When $\pi_{\mathrm{data}} = \mathcal{N}(\mu, \Sigma)$ the reverse-time SDE (2) has an explicit linear drift and admits closed-form transition kernels (Lemma G.3 and Corollary G.4). This allows to derive explicit contraction rates in the Euclidean norm, and can be used to get sharp estimates in the 2-Wasserstein distance (Lemma G.6). Note that up to a multiplicative constant, the 2-Wasserstein distance is controlled by ρ_b (Lemma B.3). In particular, for any $0 \le s < t \le T$, for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$,

$$
\mathcal{W}_2\big(\delta_{\mathbf{x}}\mathrm{Q}_{t|s}, \delta_{\mathbf{x}'}\mathrm{Q}_{t|s}\big) \le \left\| \mathrm{e}^{-\alpha\int_s^t \bar{\beta}_u \mathrm{d}u} \Sigma_{T-t}\Sigma^{-1}_{T-s} \right\| \|\mathbf{x} - \mathbf{x}'\| ,
$$

with Σ_t defined in Lemma G.1. In Lemma G.5, we show that in the VE case ($\alpha = 0$) the contracting factor is always smaller than 1 and in the VP case ($\alpha > 0$) the same strict contraction holds whenever $\lambda_{\max}(\Sigma)^{-1} > \alpha$, a condition reminiscent of $\gamma_0 > \alpha/2$ in Assumption 3.1 (I). Note that a similar condition was identified in Strasman et al. (2025a).

4. Stability of SGMs generation

Consider a subdivision $\mathcal{T}_N = \{t_0, \dots, t_N\}$. In Appendix F we provide the explicit computations of $\mathrm{K}^{(2)}_{t_{k+1}|t_k}$, $\lambda^{(2)}_{t_{k+1}|t_k}$ and $\varepsilon^{(r)}_{t_{k+1}|t_k}$ for all $0 \le k \le N-1$. Therefore, following Hairer & Mattingly (2008), we may derive $b^r_{t_k,t_{k+1}}$ and $\bar{\alpha}_{t_{k+1}|t_k}$, $0 \le k \le N-1$. Since these constants exist for all $0 \le k \le N-1$, there exist $\bar{\alpha}_* \in (0,1)$ and $b_* > 0$ such that Proposition 3.4 holds uniformly. This allows to establish Theorem 4.3 which uses the uniform geometrically decaying forgetting property to upperbound the sampling error of SGMs. We precede this result by some assumptions and notations.

Assumptions on the generative model. The previous section establishes a quantitative forgetting property for the backward process (2). We now leverage this contraction to control the generation error, *i.e.*, to bound $\rho_b\big(\pi_{\mathrm{data}}, \hat{\pi}^\theta_N\big)$, with $\hat{\pi}^\theta_N$ the distribution of the generative model defined in (5). Since $\rho_b(\cdot,\cdot)$ weights total variation by $1 + b\,\mathrm{V}_2(\cdot)$ with $\mathrm{V}_2(x) = \|x\|^2$, bounding the generation error requires a control of polynomial moments for the approximated discretized chain (4).

Assumption 4.1. Let $(\bar{\mathbf{X}}^\theta_{t_k})_{0\le k\le N}$ be defined by (4). Assume that

$$
\sup_{0\le k\le N} \mathbb{E}\left[\mathrm{V}_{4p+4}(\bar{\mathbf{X}}^\theta_{t_k})\right] < \infty . \tag{9}
$$

Remark 4.2. Assumption 4.1 is mild on a fixed grid. Since $\bar{\mathbf{X}}^\theta_{t_0} \sim \pi_\infty$ is Gaussian, it suffices that the learned score has at most polynomial growth along the grid, *i.e.*, there exist $L_\theta \ge 0$ and $r \ge 1$ such that for all $k \in \{0, \dots, N-1\}$ and all $\mathbf{x} \in \mathbb{R}^d$,

$$
\|\mathrm{s}_\theta(\mathbf{x}, T - t_k)\| \le L_\theta\big(1 + \|\mathbf{x}\|^r\big) .
$$

Under this condition, (9) holds for any p (see Lemma A.4). In particular, the above growth bound is satisfied if $\mathrm{s}_\theta(\cdot, T - t_k)$ is Lipschitz uniformly in k, which is the case for most standard neural architectures.

Stability of SGMs. The forgetting property proved in the previous section can be leverage to analyse the stability of SGMs sampling. Our main stability result follows from a simple principle: forgetting converts local errors into a telescoping, geometrically weighted sum, as proved in Theorem 4.3. For all $1 \le k \le N$, define the score approximation term by

$$
\|\mathcal{E}_k\|_{\mathrm{L}_2(\hat{\pi}^\theta_k)} = \mathbb{E}\left[\left\|\mathrm{S}_{T-t_k}\left(\bar{\mathbf{X}}^\theta_{t_k}\right) - \mathrm{s}_\theta\left(\bar{\mathbf{X}}^\theta_{t_k}, T - t_k\right)\right\|^2\right]^{\frac{1}{2}} .
$$

Theorem 4.3. *Suppose Assumption 3.1 and Assumption 4.1 hold. Let $\Delta_k := \int_{t_k}^{t_{k+1}} \bar{\beta}_u \,\mathrm{d}u$. Then, there exist $b > 0$, a contraction factor $\bar{\alpha} \in (0,1)$, a constant $C^{\mathrm{mix}} > 0$, and nonnegative coefficients $\{C^{\mathrm{discr}}_k\}_{k=0}^{N-1}$, $\{C^{\mathrm{net}}_k\}_{k=0}^{N-1}$ such that the SGM output $\hat{\pi}^\theta$ satisfies*

$$
\rho_b\big(\pi_{\mathrm{data}}, \hat{\pi}^\theta_N\big) \le \bar{\alpha}^N \Lambda(T) C^{\mathrm{mix}} + \sum_{k=1}^N \bar{\alpha}^{N-k}\left(\Delta_k C^{\mathrm{disc}}_{k-1} + \sqrt{\Delta_k} C^{\mathrm{net}}_{k-1} \big\|\mathcal{E}_{k-1}\big\|_{\mathrm{L}_2(\hat{\pi}^\theta_{k-1})}\right)
$$

The regime factor $\Lambda(T)$ *is given by*

$$\Lambda(T) := \frac{1}{2}\left\|\overrightarrow{\mathbf{X}}_0\right\|_{\mathrm{L}_2}\left(\int_0^T \beta_s \mathrm{d}s\right)^{-1/2}$$

in the variance exploding case and by

$$\Lambda(T) := \mathrm{KL}\left(\pi_{\mathrm{data}}\|\pi_\infty\right)^{1/2}\exp\left(-\alpha\int_0^T \beta_s \mathrm{d}s\right)$$

in the variance preserving case.

Proof. Throughout the proof, the time grid is fixed and we adopt the shorthand notation $t_k \equiv k$ whenever no ambiguity arises (e.g. $\mathrm{Q}_{k|\ell}$ stands for $\mathrm{Q}_{t_k|t_\ell}$).

We first decompose the global error into an initialization component and a cumulative approximation component:

$$\begin{aligned}\rho_b(\pi_{\mathrm{data}}, \hat{\pi}_N^\theta) &= \rho_b\big(p_T \mathrm{Q}_T, \pi_\infty \mathbb{Q}_{0:N}^\theta\big)\\ &\leq \rho_b(p_T \mathrm{Q}_T, \pi_\infty \mathrm{Q}_T) + \rho_b(\pi_\infty \mathrm{Q}_T, \pi_\infty \mathbb{Q}_{0:N}^\theta)\end{aligned}$$

Step 1: Initialization error. By applying Proposition 3.4 iteratively across the N transitions of the discretization grid, there exist $b > 0$ and $\bar{\alpha} \in (0,1)$, such that

$$\rho_b(p_T \mathrm{Q}_T, \pi_\infty \mathrm{Q}_T) \leq \bar{\alpha}^N \rho_b(p_T, \pi_\infty)\,.$$

The quantity $\rho_b(p_T, \pi_\infty)$ measures the degree of mixing achieved by the forward diffusion (1) at time T. It is controlled explicitly using a weighted version of Pinsker's inequality (Proposition H.2), combined with an explicit estimate of the Kullback–Leibler decay $\mathrm{KL}\left(p_T\|\pi_\infty\right)$ established in Lemma H.1. Together, these results yield

$$\rho_b\big(p_T \mathrm{Q}_T, \pi_\infty \mathbb{Q}_{0:N}^\theta\big) \leq \bar{\alpha}^N \Lambda(T) C^{\mathrm{mix}}\,,$$

with C^{mix} defined in Proposition H.2.

Step 2: Discretization and approximation errors. We now control the second error term using a telescoping argument along the grid to make appear one step kernel error, e.g. the discrepency between $\mathrm{Q}_{k|k-1}^\theta$ and $\mathrm{Q}_{k|k-1}$. Using the notation $\hat{\pi}_k^\theta$ as defined in (5), for each $k \in \{1, \ldots, N\}$ define the intermediate measures:

$$\eta_k := \hat{\pi}_{k-1}^\theta\, \mathrm{Q}_{k|k-1}\, \mathrm{Q}_{k|N}\,, \quad \tilde{\eta}_k := \hat{\pi}_{k-1}^\theta\, \mathrm{Q}_{k|k-1}^\theta\, \mathrm{Q}_{k|N}\,,$$

so that $\eta_1 = \pi_\infty \mathrm{Q}_N$ and $\tilde{\eta}_N = \pi_\infty \mathbb{Q}_{0:N}^\theta$. We have,

$$\rho_b(\pi_\infty \mathrm{Q}_N,\, \pi_\infty \mathbb{Q}_{0:N}^\theta) \leq \sum_{k=1}^N \rho_b(\eta_k, \tilde{\eta}_k)\,.$$

Using again the forgetting property applied to the tail kernel yields $\mathrm{Q}_{k|N}$,

$$\begin{aligned}&\rho_b\big(\pi_\infty \mathrm{Q}_N, \pi_\infty \mathbb{Q}_{0:N}^\theta\big)\\ &\quad\leq \sum_{k=1}^N \bar{\alpha}^{N-k} \rho_b\Big(\hat{\pi}_{k-1}^\theta \mathrm{Q}_{k|k-1}, \hat{\pi}_{k-1}^\theta \mathrm{Q}_{k|k-1}^\theta\Big)\,.\end{aligned}$$

Moreover, by definition of ρ_b any $k \in \{1, \ldots, N\}$ and $\mathbf{x} \in \mathbb{R}^d$ (Lemma B.2),

$$\begin{aligned}&\rho_b\Big(\hat{\pi}_{k-1}^\theta \mathrm{Q}_{k|k-1}, \hat{\pi}_{k-1}^\theta \mathrm{Q}_{k|k-1}^\theta\Big)\\ &\qquad\leq \int \rho_b\Big(\delta_{\mathbf{x}} \mathrm{Q}_{k|k-1}, \delta_{\mathbf{x}} \mathrm{Q}_{k|k-1}^\theta\Big)\, \hat{\pi}_{k-1}^\theta(\mathrm{d}\mathbf{x})\,.\end{aligned}$$

The remaining task is a Dirac one-step estimate control in ρ_b. This control is provided by Proposition H.3, whose proof relies on a local Girsanov argument (Lemma H.5 and Corollary H.6). Consequently, for each $k \in \{1, \ldots, N\}$,

$$\begin{aligned}&\rho_b\Big(\hat{\pi}_{k-1}^\theta \mathrm{Q}_{t_k|t_{k-1}}, \hat{\pi}_{k-1}^\theta \mathrm{Q}_{t_k|t_{k-1}}^\theta\Big)\\ &\qquad\leq \Delta_k C_{k-1}^{\mathrm{discr}} + \sqrt{\Delta_k} C_{k-1}^{\mathrm{net}} \left\|\mathcal{E}_{k-1}\right\|_{\mathrm{L}_2(\hat{\pi}_{k-1}^\theta)}\,,\end{aligned}$$

with C_{k-1}^{discr} and C_{k-1}^{net} defined in Corollary H.4. Injecting this estimate into the discounted telescoping bound completes the argument. □

Theorem 4.3 establishes a quantitative stability result for SGMs, showing that the reverse diffusion dynamics induces an intrinsic contraction mechanism along the sampling trajectory. The first term, decaying as $\bar{\alpha}^N$ captures the initialization error. The second term accounts for time-discretization error and scales linearly with the local step size Δ_k while the third term quantifies the propagation of score approximation errors and scales as $\sqrt{\Delta_k}\left\|\mathcal{E}_{k-1}\right\|_{L_2(\hat{\pi}_{k-1}^\theta)}$. Crucially, both contributions are geometrically discounted by the factor $\bar{\alpha}^{N-k}$ showing that errors incurred early in the trajectory have a vanishing influence on the final distribution and reveals an intrinsic robustness mechanism of SGMs. In particular, for a uniform discretization grid (for any $k \in \{1, \ldots, N\}$ let $\Delta = (t_k - t_{k-1})\beta_T$), the accumulated error admits, in the VP case, the form

$$\begin{aligned}\rho_b\big(\pi_{\mathrm{data}}, \hat{\pi}_N^\theta\big) &\lesssim \bar{\alpha}^N \mathrm{e}^{-T} + \frac{\Delta}{1-\bar{\alpha}} \sup_k C_k^{\mathrm{discr}}\\ &\quad + \frac{\sqrt{\Delta}}{1-\bar{\alpha}} \sup_k \left\|\mathcal{E}_{k-1}\right\|_{L_2(\hat{\pi}_{k-1}^\theta)}\,.\end{aligned}$$

Moreover, the stability bound is stated in the weighted total variation distance which directly controls other probability distance. Indeed, by Lemma B.3, controls the total variation distance and the 2-Wasserstein distance (up to a multiplicative factors). Consequently, Theorem 4.3 immediately yields the same decomposition and order in those metrics.

5. Numerical illustrations

We propose experiments in controlled settings to illustrate the theoretical *forgetting and stability mechanisms* of Section 3 and Section 4. We therefore focus on distributions for which Assumption 3.1 can be verified and for which the mechanisms predicted by the theory can be meaningfully interpreted and compared to the empirical behavior.

Concretely, we consider both Gaussian targets with several covariance structures (isotropic, ill-conditioned diagonal, and correlated) and Gaussian mixture models, which are nonconvex and multimodal yet satisfy dissipativity. In the main text we present results for a Gaussian mixture model in the Variance-Exploding (VE) setting; additional results for all Gaussian targets in both VE and VP, together with full experimental details, are provided in Appendix I.

All experiments use dimension $d = 50$. For the GMM, we place the 25 component means on a 5×5 grid in the first two coordinates (details in Appendix I); this enables direct 2D visualization by projecting samples onto these coordinates. For the scheduling, we used the scheduler from Karras et al. (2022b, Equation 5) with $N = 100$, $\sigma_{\min} = 0.002$, $\sigma_{\max} = 80$ and $\rho = 3$. All the calculations have been replicated 20 times using different seeds and the reported results consist of the mean and the standard deviations of those 20 replicates.

Sensitivity to initialization. A central prediction of our analysis is that the reverse-time dynamics *forgets* early perturbations: errors introduced earlier along the reverse trajectory should have a weaker impact on the final sample than errors introduced later. To study this mechanism, we consider an experiment that injects a calibrated bias at an intermediate diffusion time $t_{\text{bias}} \in [0, T]$ and measures its effect after completing the reverse sampling up the final time T. We fix $t_{\text{bias}} \in (0, T]$ and compare the ideal reverse transition from time t_{bias} to T, represented by the Markov kernel $\mathrm{Q}_{T|t_{\text{bias}}}$, with its discretized implementation $\bar{\mathrm{Q}}_{T|t_{\text{bias}}}$ (Euler–Maruyama with fixed step size). The goal here is to assess whether the discretized reverse chain exhibits robustness to perturbations. In principle, one would like to quantify the effect of the perturbation by measuring the discrepancy between the target law and the biased, discretized output:

$$\mathcal{W}_2\left(p_{t_{\text{bias}}}\mathrm{Q}_{T|t_{\text{bias}}}, \tilde{p}_{t_{\text{bias}}}\bar{\mathrm{Q}}_{T|t_{\text{bias}}}\right) = \mathcal{W}_2\left(\pi_{\text{data}}, \tilde{p}_{t_{\text{bias}}}\bar{\mathrm{Q}}_{T|t_{\text{bias}}}\right) ,$$

with $\tilde{p}$ the biased distribution. However, on Gaussian mixture targets the quantity above is not available in closed form. We therefore replace $\mathcal{W}_2$ by the *Maximum sliced Wasserstein* distance Max SW (Kolouri et al., 2019, Section 4.3), which is both tractable from samples and provides a meaningful proxy for $\mathcal{W}_2$ distance. To define a *worst-case* perturbation direction, we first estimate a maximizing direction $u_{\max} \in \mathbb{S}^{d-1}$ associated with Max SW using two independent batches of samples from π_{data} (see Appendix I.2 for details). We then construct the biased initialization at time t_{bias} by shifting forward samples $\mathbf{x}^{(i)}_{t_{\text{bias}}} \sim p_{t_{\text{bias}}}$ along $u_{\max}$:

$$\mathbf{x}^{(i)}_{t_{\text{bias}},\lambda} = \mathbf{x}^{(i)}_{t_{\text{bias}}} + \sigma^2_{t_{\text{bias}}|0}\lambda u_{\max} \quad (10)$$

where $\sigma_{t_{\text{bias}}|0}$ is the forward variance at time t_{bias} (see Lemma C.1) and λ is the perturbation magnitude. Starting from $\{\mathbf{x}^{(i)}_{t_{\text{bias}},\lambda}\}_{i=1}^M$, we run the discretized reverse dynamics to time T and report the resulting Max SW discrepancy between the generated samples and π_{data} as a function of t_{bias} and λ (see Figure 3). We also visualize the effect of the perturbation directly in state space by plotting a two-dimensional projection of the GMM samples (data vs. perturbed outputs) in Figure 1. We observe that perturbations injected at small t_{bias} have the largest impact, while perturbations injected at larger t_{bias} are strongly attenuated, in line with the predicted forgetting behavior. A similar qualitative behaviour is observed for Gaussian targets (Figure 5).

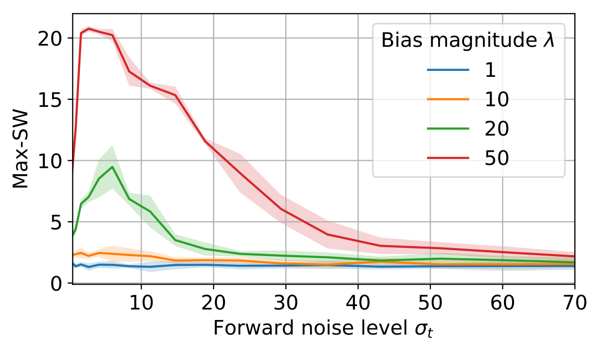

Figure 3. Max SW as a function of the noise level and λ in the case of initialization perturbation. We use the forward-time convention ($t = 0$ is the data distribution).

Local score approximation error. A key qualitative prediction of our stability theory is that local errors are *discounted* when they occur early along the reverse trajectory.

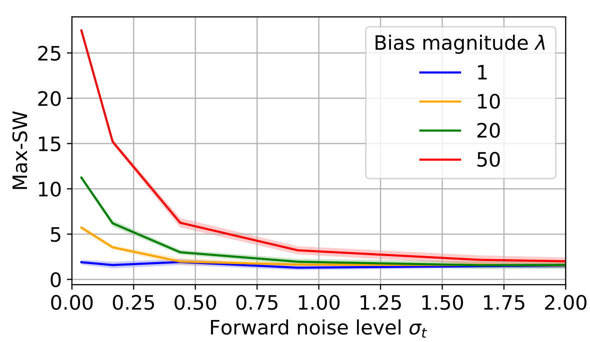

Figure 4. Max SW as a function of the noise level and perturbation magnitude λ for the score-perturbation experiment. We use the forward-time convention ($t = 0$ is the data distribution).

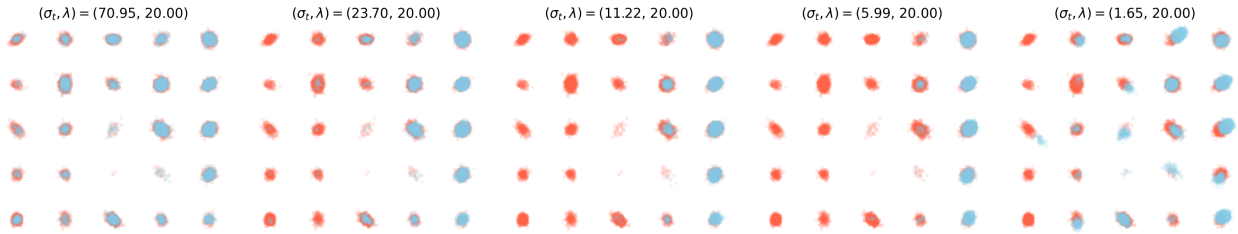

Figure 1. Initialization bias in the GMM case for several noise levels σ_t and with $\lambda = 20$. Red points are samples from π_{data} and blue points the output of the biased initialization experiment.

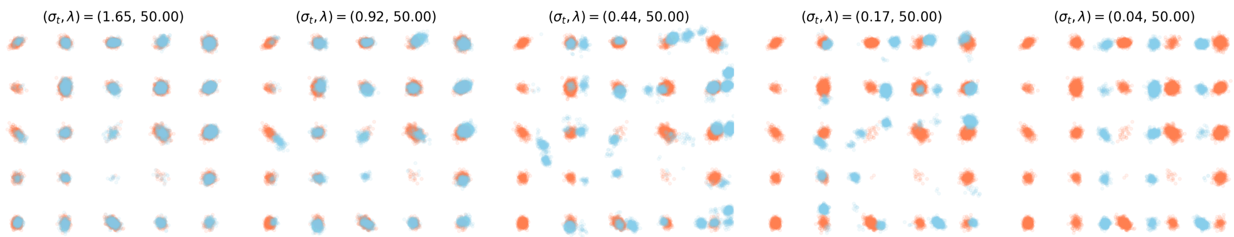

Figure 2. Local perturbation of the score in the GMM case for several noise levels σ_t and with $\lambda = 50$. Red points are samples from π_{data} and blue points the output of the perturbed score experiment.

To probe this local-to-global propagation mechanism, we run the reverse-time sampler over the full horizon $[0, T]$ and introduce a *single* controlled score perturbation at one discretization step. Specifically, consider the Euler–Maruyama reverse sampler on a grid $(t_n)_{n=0}^{N}$. For a chosen error index n_{err}, we modify the score only at that step:

$$\tilde{\mathrm{S}}_{t_{\text{err}}}(\mathbf{x}) = \mathrm{S}_{t_{\text{err}}}(\mathbf{x}) + \frac{\lambda}{\sigma^2_{t_{\text{err}}|0}} u_{\max} \,,$$

with $u_{\max}$ defined as in (10). The perturbation is chosen so that it produces a merely comparable score approximation disturbance across times by accounting for the forward variance terms. Results are reported in Figure 4 and in Figure 2. They support the idea that errors introduced early along the reverse trajectory are geometrically discounted. Moreover, the Gaussian targets exhibit similar trends (Figure 6)

6. Discussion

A widely used heuristic in SGMs is that accuracy near the end of the reverse trajectory matters most. This is also reflected in practice: many implementations rely on adaptive time discretizations that allocate finer step sizes close to the data distribution, *i.e.*, near small noise levels (Karras et al., 2022a).

Our results provide a theoretical framework to understand this phenomenon. Under a dissipativity condition on the data score at time 0 and mild polynomial growth control on its Jacobian, we show that the reverse-time diffusion satisfies a Harris-type stability property. In particular, the Markov semigroup of the backward process contracts a weighted total variation distance ρ_b. This property is essential to understand the robustness of SGMs. At each step, local perturbations (initialization mismatch, discretization, or score approximation error) propagate with a *geometric discount factor*. As a consequence, errors incurred early along the reverse trajectory have a vanishing influence on the final distribution: the reverse diffusion sampling process *forgets*. This perspective connects SGMs with classical stability tools for Markov processes and provides a principled lens to analyze how noise schedules, numerical integrators, and score errors evolve along the sampling trajectory.

Limitations and perspectives. A key limitation is that our minorization argument is localized on a small set. While this yields explicit constants, they can be conservative. In particular, enforcing a single uniform pair $(b, \bar{\alpha})$ over an entire refined grid can lead to pessimistic global bounds. Our framework naturally accommodates time-dependent metrics but optimizing these choices along the trajectory to obtain sharper constants is technically challenging and remains an open direction.

Moreover, in this work, we focus on the quadratic Lyapunov function $\mathrm{V}_2(x) = \|x\|^2$ but other choices could have been explored. This choice allows for an explicit lower bound on the backward kernel, although alternative choices could be investigated to obtain a tighter minorization properties and therefore sharper sampling error bounds. The choice of the Lyapunov function has an impact on the constants in the proposed upper bounds, and investigating the link between these constants and hyperparameter tuning remains an open question for future research.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Bakry, D., Gentil, I., and Ledoux, M. *Analysis and Geometry of Markov Diffusion Operators*, volume 348. Springer, 2014.

Beyler, E. and Bach, F. Convergence of deterministic and stochastic diffusion-model samplers: A simple analysis in wasserstein distance. *arXiv preprint arXiv:2508.03210*, 2025.

Bouchut, F., James, F., and Mancini, S. Uniqueness and weak stability for multi-dimensional transport equations with one-sided lipschitz coefficient. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 4(1):1–25, 2005.

Brigati, G. and Pedrotti, F. Heat flow, log-concavity, and lipschitz transport maps. *Electronic Communications in Probability*, 30:1–12, 2025.

Bruno, S. and Sabanis, S. Wasserstein convergence of score-based generative models under semiconvexity and discontinuous gradients. *arXiv preprint arXiv:2505.03432*, 2025.

Chen, H., Lee, H., and Lu, J. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4735–4763. PMLR, 23–29 Jul 2023a.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=zyLVMgsZ0U_.

Chen, Y., Chewi, S., Salim, A., and Wibisono, A. Improved analysis for a proximal algorithm for sampling. In *Conference on Learning Theory*, pp. 2984–3014. PMLR, 2022.

Dalalyan, A. S. and Karagulyan, A. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12): 5278–5311, 2019.

Durmus, A. and Moulines, E. Non-asymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

Eberle, A. Reflection couplings and contraction rates for diffusions. *Probability theory and related fields*, 166(3): 851–886, 2016.

Erdogdu, M. A. and Hosseinzadeh, R. On the convergence of langevin monte carlo: The interplay between tail growth and smoothness. In *Conference on Learning Theory*, pp. 1776–1822. PMLR, 2021.

Erdogdu, M. A., Hosseinzadeh, R., and Zhang, S. Convergence of langevin monte carlo in chi-squared and rényi divergence. In *International Conference on Artificial Intelligence and Statistics*, pp. 8151–8175. PMLR, 2022.

Gao, X., Nguyen, H. M., and Zhu, L. Wasserstein convergence guarantees for a general class of score-based generative models. *Journal of machine learning research*, 26(43):1–54, 2025a.

Gao, X., Nguyen, H. M., and Zhu, L. Wasserstein convergence guarantees for a general class of score-based generative models. *Journal of Machine Learning Research*, 26(43):1–54, 2025b. URL http://jmlr.org/papers/v26/24-0902.html.

Gentiloni-Silveri, M. and Ocello, A. Beyond log-concavity and score regularity: Improved convergence bounds for score-based generative models in w2-distance. In *Forty-second International Conference on Machine Learning*, 2025.

Hairer, M. and Mattingly, J. C. Yet another look at harris' ergodic theorem for markov chains. *arXiv preprint arXiv:0810.2777*, 2008. URL https://arxiv.org/abs/0810.2777.

Haussmann, U. G. and Pardoux, E. Time reversal of diffusions. *The Annals of Probability*, 14(4):1188–1205, 1986.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6840–6851, 2020.

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

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577. Curran Associates, Inc., 2022a.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the Design Space of Diffusion-Based Generative Models. In *Proc. NeurIPS*, 2022b.

Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., and Rohde, G. Generalized sliced wasserstein distances. In Wallach, H., Larochelle, H., Beygelzimer, A., dAlché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Liptser, R. S. and Shiryaev, A. N. *Statistics of Random Processes I: General Theory*, volume 5 of *Applications of Mathematics*. Springer, Berlin, 2 edition, 2001. doi: 10.1007/978-3-662-13010-8.

Löcherbach, E. Ergodicity and speed of convergence to equilibrium for diffusion processes. Lecture notes, November 2015. URL https://eloecherbach.u-cergy.fr/cours.pdf.

Meyn, S. P. and Tweedie, R. L. Stability of markovian processes iii: Foster–lyapunov criteria for continuous-time processes. *Advances in Applied Probability*, 25(3): 518–548, 1993.

Meyn, S. P. and Tweedie, R. L. *Markov Chains and Stochastic Stability*. Cambridge University Press, 2 edition, 2009. ISBN 978-0-521-73182-9.

Meyn, S. P. and Tweedie, R. L. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

Nutz, M. Introduction to entropic optimal transport. Lecture notes, Columbia University, 2021. URL https://www.math.columbia.edu/~mnutz/docs/EOT_lecture_notes.pdf. Online PDF.

Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pp. 1674–1703. PMLR, 2017.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR.

Song, J., Sohl-Dickstein, J., Kingma, D. P., Kumar, R., Ermon, S., and Poole, B. Solving inverse problems in generative modeling with score-based diffusion. *arXiv preprint arXiv:2201.04600*, 2022.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 11918–11930, 2019.

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

Stéphanovitch, A. Regularity of the score function in generative models. *arXiv preprint arXiv:2506.19559*, 2025.

Strasman, S., Ocello, A., Boyer, C., Corff, S. L., and Lemaire, V. An analysis of the noise schedule for score-based generative models. *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856. URL https://openreview.net/forum?id=BlYIPa0Fx1.

Strasman, S., Surendran, S., Boyer, C., Le Corff, S., Lemaire, V., and Ocello, A. Wasserstein convergence of critically damped langevin diffusions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025b.

Tang, W. and Zhao, H. Contractive diffusion probabilistic models. *arXiv preprint arXiv:2401.13115*, 2024.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009. ISBN 978-0-387-79051-0. doi: 10.1007/978-0-387-79052-7.

Vacher, A., Chehab, O., and Korba, A. Sampling from multimodal distributions with polynomial query complexity in fixed dimension via reverse diffusion. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Villani, C. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-540-71049-3, 978-3-540-71050-9. doi: 10.1007/978-3-540-71050-9.

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

Yu, Y. and Yu, L. Advancing wasserstein convergence analysis of score-based models: Insights from discretization and second-order acceleration. *arXiv preprint arXiv:2502.04849*, 2025.

Zhang, Y., Liang, P., and Charikar, M. A hitting time analysis of stochastic gradient langevin dynamics. In *Conference on Learning Theory*, pp. 1980–2022. PMLR, 2017.

Appendix

Table of contents.

A. Discussion on the assumptions

This section collects useful results derived from Assumption 3.1 and Assumption 4.1 that are discussed in the main paper.

A.1. Regarding the data distribution and its associated score function.

Assumption 3.1 (I) is a dissipativity condition on the data score function $\mathrm{S}_0(\cdot)$; in particular, π_{data} has sub-Gaussian tails and finite moments of all orders (Lemma A.1). Assumption 3.1 (II) controls the Jacobian of the score and implies a polynomial growth bound on the score itself (Lemma A.2).

Lemma A.1. *Suppose that Assumption 3.1 (I) holds. Then, the data distribution π_{data} admits exponential moments of order λ for $\lambda < \gamma_0/2$, i.e.,*

$$\mathbb{E}_{\pi_{\mathrm{data}}}\left[\exp\left(\lambda \|\mathbf{X}\|^2\right)\right] < \infty \,. \tag{11}$$

This implies that π_{data} has sub-Gaussian tails.

Proof. Fix $\lambda < \gamma_0/2$. Let $U(x) = -\log p_0(x)$. Consider the following second-order Taylor expansion:

$$U(\mathbf{x}) = U(0) + \langle \nabla U(0), \mathbf{x}\rangle + \frac{1}{2}\left\langle \mathbf{x}, \nabla^2 U(y)\,\mathbf{x}\right\rangle \,,$$

for some $y \in \{tx : t \in [0,1]\}$. From Assumption 3.1 (I), we have that

$$\langle \mathrm{S}_0(\mathbf{x}), \mathbf{x}\rangle \leq -\gamma_0 \|\mathbf{x}\|^2 + \kappa_0 \,.$$

Fix $\bar\lambda$ such that $\lambda < \bar\lambda < \gamma_0/2$. The previous inequality implies that there exists $R_{\bar\lambda} > 0$ such that

$$-\langle \nabla U(\mathbf{x}), \mathbf{x}\rangle \leq -2\bar\lambda \|\mathbf{x}\|^2 \,, \qquad \text{for } \mathbf{x} \notin \mathrm{B}_{R_{\bar\lambda}}(0), \,.$$

By Lemma 2.2 in Bouchut et al. (2005), the above inequality implies that

$$-\nabla^2 U(\mathbf{x}) \preceq -2\bar\lambda\,\mathbf{I}_d \,, \qquad \text{for } \mathbf{x} \notin \mathrm{B}_{R_{\bar\lambda}}(0) \,.$$

Therefore,

$$\begin{aligned}\int_{\mathbb{R}^d} \mathrm{e}^{\lambda\|\mathbf{x}\|^2} p_0(\mathbf{x})\,\mathrm{d}\mathbf{x} &= \int_{\mathrm{B}_{R_{\bar\lambda}}(0)} \mathrm{e}^{-\lambda\|\mathbf{x}\|^2} p_0(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int_{\mathbb{R}^d\setminus \mathrm{B}_{R_{\bar\lambda}}(0)} \mathrm{e}^{\lambda\|\mathbf{x}\|^2}\mathrm{e}^{-U(\mathbf{x})}\mathrm{d}\mathbf{x} \\ &\leq \int_{\mathrm{B}_{R_{\bar\lambda}}(0)} \mathrm{e}^{-\lambda\|\mathbf{x}\|^2} p_0(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int_{\mathbb{R}^d\setminus \mathrm{B}_{R_{\bar\lambda}}(0)} \mathrm{e}^{\lambda\|\mathbf{x}\|^2}\mathrm{e}^{-U(0)-\langle\nabla U(0),\mathbf{x}\rangle-\bar\lambda\|\mathbf{x}\|^2}\mathrm{d}\mathbf{x} < \infty \,,\end{aligned}$$

which concludes the proof. □

Lemma A.2. *Suppose that Assumption 3.1 (II) holds. Then, there exists $\mathrm{C}_s \in \mathbb{R}_+$ such that*

$$\|\mathrm{S}_0(\mathbf{x})\| \leq \mathrm{C}_s(1 + \|\mathbf{x}\|^{p+1}) \,, \tag{12}$$

where $\mathrm{C}_s = \max\left\{\sqrt{d}\,\mathrm{C}_0, \|\mathrm{S}_0(0)\|\right\}$.

Proof. We have that

$$\|\mathrm{S}_0(\mathbf{x}) - \mathrm{S}_0(0)\| \leq \int_0^1 \|\nabla \mathrm{S}_0(t\mathbf{x})\,\mathbf{x}\|\,\mathrm{d}t \leq \sqrt{d}\,\mathrm{C}_0 \int_0^1 (1 + t^p\|\mathbf{x}\|^{p+1})\mathrm{d}t \,. \tag{13}$$

□

Corollary A.3. *Suppose that Assumption 3.1 (II) holds. Then, there exists $\Gamma_0^s \in \mathbb{R}_+$ such that*

$$\left\|\nabla \mathrm{S}_0(\mathbf{x}) + \mathrm{S}_0(\mathbf{x})^{\otimes 2}\right\|_{\mathrm{F}} \leq \Gamma_0^s\left(1 + \|\mathbf{x}\|^{2(p+1)}\right) \,. \tag{14}$$

Proof. It is enough to note that

$$\left\|\nabla \mathrm{S}_0(\mathbf{x}) + \mathrm{S}_0(\mathbf{x})^{\otimes 2}\right\|_{\mathrm{F}} \leq \|\nabla \mathrm{S}_0(\mathbf{x})\|_{\mathrm{F}} + \left\|\mathrm{S}_0(\mathbf{x})^{\otimes 2}\right\|_{\mathrm{F}} \tag{15}$$

and $\left\|\mathrm{S}_0(\mathbf{x})^{\otimes 2}\right\|_{\mathrm{F}} = \|\mathrm{S}_0(\mathbf{x})\|^2$. □

A.2. Regarding the score approximation and the numerical scheme

We identify a minimal growth condition on the learned score approximation s_θ that guarantees the Euler–type scheme defined in (4) admits finite moments at all discretization times (Lemma A.4). This ensures all quantities appearing in the latter error decomposition are well-defined.

Lemma A.4. *Let $(\bar{\mathbf{X}}^\theta_{t_k})_{0\le k\le N}$ be defined by (4) and recall that $\bar{\mathbf{X}}^\theta_{t_0} \sim \pi_\infty$ is Gaussian. Assume that there exist constants $L_\theta \ge 0$ and $r \ge 1$ such that for all $k \in \{0,\dots,N-1\}$ and all $\mathbf{x} \in \mathbb{R}^d$,*

$$\|\mathrm{s}_\theta\left(\mathbf{x}, T-t_k\right)\| \le L_\theta\left(1 + \|\mathbf{x}\|^r\right) . \tag{16}$$

Then, for every $q \ge 1$,

$$\sup_{0\le k\le N} \mathbb{E}\left\|\bar{\mathbf{X}}^\theta_{t_k}\right\|^q < \infty .$$

Proof. Let $(\xi_k)_{k\ge 0}$ be an i.i.d. sequence of standard Gaussian random vectors in $\mathbb{R}^d$ and write

$$\bar{\mathbf{X}}^\theta_{t_{k+1}} = \bar{\mathbf{X}}^\theta_{t_k} + \Delta_k b_\theta\left(t_k, \bar{\mathbf{X}}^\theta_{t_k}\right) + \sqrt{2\Delta\tau_k}\xi_k ,$$

with $b_\theta(t_k,\mathbf{x}) := \alpha\mathbf{x} + 2\,\mathrm{s}_\theta\left(\mathbf{x}, T-t_k\right)$. By (16) there exists $C_\theta > 0$ such that $\|b_\theta\left(t_k,\mathbf{x}\right)\| \le C_\theta(1+\|\mathbf{x}\|^r)$ for all $k \in \{0,\dots,N-1\}$ and $\mathbf{x}\in\mathbb{R}^d$. Fix $p\ge 1$, we get that there exists C_θ, which may change from line to line, such that

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{\mathbf{X}}^\theta_{t_{k+1}}\right\|^p \middle| \bar{\mathbf{X}}^\theta_{t_k}\right] &\le C_\theta\left(\left\|\bar{\mathbf{X}}^\theta_{t_k}\right\|^p + (\Delta\tau_k)^p\left\|b_\theta(t_k,\bar{\mathbf{X}}^\theta_{t_k})\right\|^p + (\Delta\tau_k)^{p/2}\mathbb{E}\left[\|\xi_k\|^p\right]\right)\\ &\le C_\theta\left(1 + \left\|\bar{\mathbf{X}}^\theta_{t_k}\right\|^{pr}\right) ,\end{aligned}$$

Since $\bar{\mathbf{X}}^\theta_{t_0}$ is Gaussian, it has moments of all orders. The previous inequality shows that for an arbitrary p, if $\mathbb{E}\left\|\bar{\mathbf{X}}^\theta_{t_k}\right\|^{pr} < \infty$ then $\mathbb{E}\left\|\bar{\mathbf{X}}^\theta_{t_{k+1}}\right\|^p < \infty$. By induction over $k=0,\dots,N-1$, we obtain that $\mathbb{E}\left\|\bar{\mathbf{X}}^\theta_{t_k}\right\|^q < \infty$. Taking the supremum over the finite set $\{0,\dots,N\}$ yields the claim. □

Remark A.5. If, for each k, the map $\mathbf{x}\mapsto \mathrm{s}_\theta\left(\mathbf{x}, T-t_k\right)$ is Lipschitz with constant L_θ and $\sup_k \|\mathrm{s}_\theta\left(0, T-t_k\right)\| \le M$, then

$$\|\mathrm{s}_\theta\left(x, T-t_k\right)\| \le \|\mathrm{s}_\theta\left(0, T-t_k\right)\| + L_\theta\|\mathbf{x}\| \le M + L_\theta\|\mathbf{x}\| ,$$

so (16) holds with $r=1$.

B. Probability metrics and useful bounds

We collect here the definitions of the probability distances used throughout the paper, together with a few inequalities relating them. Some results are standard and are included for the sake of clarity. The last subsection is dedicated to properties of ρ_b, which are used to establish the main contraction results.

B.1. Definitions

Weighted total variation distance. For any scale parameter $b>0$ and $\mathrm{V}_2(\mathbf{x}) := \|\mathbf{x}\|^2$, define the weighted supremum norm

$$\|\varphi\|_b := \sup_{\mathbf{x}\in\mathbb{R}^d} \frac{|\varphi(\mathbf{x})|}{1 + b\,\mathrm{V}_2(\mathbf{x})} .$$

The corresponding dual distance on probability measures (Hairer & Mattingly, 2008) is given, for $\mu_1,\mu_2\in\mathcal{P}(\mathbb{R}^d)$, by

$$\rho_b(\mu_1,\mu_2) := \sup_{\|\varphi\|_b\le 1}\left|\int \varphi \mathrm{d}(\mu_1-\mu_2)\right| . \tag{17}$$

This coincides with a weighted total variation distance

$$\rho_b(\mu_1,\mu_2) = \int_{\mathbb{R}^d}(1 + b\,\mathrm{V}_2(\mathbf{x}))|\mu_1-\mu_2|(\mathrm{d}\mathbf{x}) . \tag{18}$$

Total variation distance. The total variation distance between μ_1 and μ_2 is defined by

$$\|\mu_1 - \mu_2\|_{\mathrm{TV}} := \frac{1}{2} \sup_{\|f\|_\infty \le 1} \left| \int f \, \mathrm{d}(\mu_1 - \mu_2) \right| .$$

In particular, if μ_1 and μ_2 admit respectively densities p_1 and p_2 with respect to the Lebesgue measure, then

$$\|\mu_1 - \mu_2\|_{\mathrm{TV}} := \frac{1}{2} \int_{\mathbb{R}^d} |p_1(\mathbf{x}) - p_2(\mathbf{x})| \, \mathrm{d}\mathbf{x}, \quad \mathbf{x} \in \mathbb{R}^d .$$

Kullback–Leibler divergence. The Kullback–Leibler (KL) divergence of μ_1 with respect to μ_2 is defined as

$$\mathrm{KL}\left(\mu_1 \| \mu_2\right) := \int_{\mathbb{R}^d} \log\left(\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}\right) \mathrm{d}\mu_1 ,$$

if $\mu_1 \ll \mu_2$ and $\mathrm{KL}\left(\mu_1 \| \mu_2\right) = +\infty$ otherwise.

Hellinger distance. If μ_1 and μ_2 admit respectively densities p_1 and p_2 with respect to the Lebesgue measure, the Hellinger distance between μ_1 and μ_2 is defined as

$$\mathrm{H}(\mu_1, \mu_2) := \left(\frac{1}{2} \int_{\mathbb{R}^d} \left(\sqrt{p_1(\mathbf{x})} - \sqrt{p_2(\mathbf{x})} \right)^2 \mathrm{d}\mathbf{x} \right)^{1/2} , \quad \mathbf{x} \in \mathbb{R}^d .$$

B.2. Standard inequalities

The Kullback–Leibler divergence and the total variation distance are related by Pinsker's inequality

$$\|\mu_1 - \mu_2\|_{\mathrm{TV}} \le \sqrt{\tfrac{1}{2}\mathrm{KL}\left(\mu_1 \| \mu_2\right)} . \tag{19}$$

For a proof of such inequality see for example Bakry et al. (2014) or Tsybakov (2009, Chapter 2, Distances between probability measures). The Hellinger distance is upper bounded by the Kullback–Leibler divergence

$$\mathrm{H}(\mu_1, \mu_2)^2 \le \tfrac{1}{2} \, \mathrm{KL}\left(\mu_1 \| \mu_2\right) . \tag{20}$$

A proof of such result can be found in Tsybakov (2009, Chapter 2, Distances between probability measures).

B.3. Weighted total variation: properties and bounds

We collect practical inequalities bounding above and below the weighted total variation distance defined in (18) in terms of other probability metrics. These bounds will be used in the subsequent development.

Lemma B.1. *Let $\mu_1, \mu_2 \in \mathcal{P}(\mathbb{R}^d)$ be probability measures admitting Lebesgue densities p_1, p_2 respectively such that*

$$\int_{\mathbb{R}^d} \mathrm{V}_2^2 \, \mathrm{d}\mu_1 < \infty \quad \textit{and} \quad \int_{\mathbb{R}^d} \mathrm{V}_2^2 \, \mathrm{d}\mu_2 < \infty .$$

Then,

$$\rho_b(\mu_1, \mu_2) \le \sqrt{2} \mathrm{H}\left(\mu_1, \mu_2\right) \left(\sqrt{\int_{\mathbb{R}^d} \left(1 + b\, \mathrm{V}_2(\mathbf{x})\right)^2 p_1(\mathbf{x}) \mathrm{d}\mathbf{x}} + \sqrt{\int_{\mathbb{R}^d} \left(1 + b\, \mathrm{V}_2(\mathbf{x})\right)^2 p_2(\mathbf{x}) \mathrm{d}\mathbf{x}} \right) .$$

Proof. Note that

$$|p_1 - p_2| = |\sqrt{p_1} - \sqrt{p_2}| \left(\sqrt{p_1} + \sqrt{p_2}\right) ,$$

and with $w(\mathbf{x}) := 1 + b\, \mathrm{V}_2(\mathbf{x})$, we have,

$$\rho_b(\mu_1, \mu_2) = \int_{\mathbb{R}^d} w(\mathbf{x}) \left| \sqrt{p_1(\mathbf{x})} - \sqrt{p_2(\mathbf{x})} \right| \left(\sqrt{p_1(\mathbf{x})} + \sqrt{p_2(\mathbf{x})} \right) \mathrm{d}\mathbf{x}.$$

By the Cauchy–Schwarz inequlity,

$$\begin{aligned}\rho_b(\mu_1,\mu_2) &\le \left(\int_{\mathbb{R}^d}\left(\sqrt{p_1(\mathbf{x})}-\sqrt{p_2(\mathbf{x})}\right)^2\mathrm{d}\mathbf{x}\right)^{1/2}\left(\int_{\mathbb{R}^d} w(\mathbf{x})^2\left(\sqrt{p_1(\mathbf{x})}+\sqrt{p_2(\mathbf{x})}\right)^2\mathrm{d}\mathbf{x}\right)^{1/2} \\ &= \sqrt{2}\mathrm{H}\left(\mu_1,\mu_2\right)\left(\int_{\mathbb{R}^d} w(\mathbf{x})^2\left(\sqrt{p_1(\mathbf{x})}+\sqrt{p_2(\mathbf{x})}\right)^2\mathrm{d}\mathbf{x}\right)^{1/2}.\end{aligned}$$

Then,

$$\int_{\mathbb{R}^d} w^2(\sqrt{p_1}+\sqrt{p_2})^2\,\mathrm{d}x = \int w^2 p_1\mathrm{d}\mathbf{x} + \int w^2 p_2 \mathrm{d}\mathbf{x} + 2\int w^2\sqrt{p_1p_2}\mathrm{d}\mathbf{x}\,,$$

so that applying the Cauchy–Schwarz to the last term yields

$$\begin{aligned}\int_{\mathbb{R}^d} w^2(\sqrt{p_1}+\sqrt{p_2})^2\mathrm{d}\mathbf{x} &\le \int_{\mathbb{R}^d} w^2p_1\mathrm{d}\mathbf{x} + \int_{\mathbb{R}^d} w^2p_2\mathrm{d}\mathbf{x} + 2\Big(\int_{\mathbb{R}^d} w^2p_1\mathrm{d}x\Big)^{1/2}\Big(\int_{\mathbb{R}^d} w^2p_2\mathrm{d}\mathbf{x}\Big)^{1/2} \\ &= \left(\Big(\int_{\mathbb{R}^d} w^2p_1\mathrm{d}\mathbf{x}\Big)^{1/2} + \Big(\int_{\mathbb{R}^d} w^2p_2\mathrm{d}\mathbf{x}\Big)^{1/2}\right)^2.\end{aligned}$$

Therefore,

$$\rho_b(\mu_1,\mu_2) \le \sqrt{2}\mathrm{H}\left(\mu_1,\mu_2\right)\left(\sqrt{\int_{\mathbb{R}^d} w(\mathbf{x})^2 p_1(\mathbf{x})\mathrm{d}\mathbf{x}} + \sqrt{\int_{\mathbb{R}^d} w(\mathbf{x})^2 p_2(x)\mathrm{d}\mathbf{x}}\right)\,,$$

which concludes the proof. □

Lemma B.2. *Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ and let $K, \tilde{K}$ be two Markov kernels on $\mathbb{R}^d$ such that*

$$\int_{\mathbb{R}^d} \|\mathbf{x}\|^2\,(\mu K)(\mathrm{d}\mathbf{x}) < \infty \quad \textit{and} \quad \int_{\mathbb{R}^d} \|\mathbf{x}\|^2\,(\mu\tilde{K})(\mathrm{d}\mathbf{x}) < \infty.$$

Then,

$$\rho_b\Big(\mu K, \mu\tilde{K}\Big) \le \int_{\mathbb{R}^d} \rho_b\Big(\delta_{\mathbf{x}}K, \delta_{\mathbf{x}}\tilde{K}\Big)\,\mu(\mathrm{d}\mathbf{x})\,.$$

Proof. Fix φ such that $\|\varphi\|_b \le 1$,

$$\begin{aligned}\Big|(\mu K)(\varphi) - (\mu\tilde{K})(\varphi)\Big| &= \Big|\int_{\mathbb{R}^d}\Big((\delta_{\mathbf{x}}K)(\varphi) - (\delta_{\mathbf{x}}\tilde{K})(\varphi)\Big)\mu(\mathrm{d}\mathbf{x})\Big| \\ &\le \int_{\mathbb{R}^d}\big|(\delta_{\mathbf{x}}K)(\varphi) - (\delta_{\mathbf{x}}\tilde{K})(\varphi)\big|\mu(\mathrm{d}\mathbf{x})\,.\end{aligned}$$

For each fixed $\mathbf{x}$, since $\|\varphi\|_b \le 1$,

$$\Big|(\delta_{\mathbf{x}}K)(\varphi) - (\delta_x\tilde{K})(\varphi)\Big| \le \sup_{\|\psi\|_b\le 1}\Big|(\delta_{\mathbf{x}}K)(\psi) - (\delta_{\mathbf{x}}\tilde{K})(\psi)\Big| = \rho_b\Big(\delta_{\mathbf{x}}K, \delta_{\mathbf{x}}\tilde{K}\Big)\,.$$

Therefore,

$$\big|(\mu K)(\varphi) - (\mu\widetilde{K})(\varphi)\big| \le \int_{\mathbb{R}^d}\rho_b\Big(\delta_x K, \delta_x\widetilde{K}\Big)\,\mu(\mathrm{d}x)\,.$$

Taking the supremum over all φ with $\|\varphi\|_b \le 1$ as in (17) gives the result. □

Lemma B.3. *Let $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$. Then,*

$$\|\mu_1 - \mu_2\|_{\mathrm{TV}} \le \frac{1}{2}\rho_b(\mu_1,\mu_2)\,,$$

and

$$\mathcal{W}_2^2(\mu_1,\mu_2) \le \frac{2}{b}\rho_b(\mu_1,\mu_2)\,.$$

Proof. The first inequality follows directly from $1 \leq 1 + b\,\mathrm{V}_2$. For the second, Villani (2009, Theorem 6.15) yields,

$$\mathcal{W}_2^2(\mu_1, \mu_2) \leq 2 \int_{\mathbb{R}^d} \|\mathbf{x}\|^2 \, |\mu_1 - \mu_2|(\mathrm{d}\mathbf{x})\,,$$

and

$$\rho_b(\mu_1, \mu_2) = \int_{\mathbb{R}^d} (1 + b\,\|\mathbf{x}\|^2)|\mu_1 - \mu_2|\mathrm{d}\mathbf{x} \\ \geq \int_{\mathbb{R}^d} b\,\|\mathbf{x}\|^2 \, |\mu_1 - \mu_2|\mathrm{d}\mathbf{x}\,,$$

which concludes the proof. □

C. Stability properties under Gaussian perturbations

This section collects stability and regularity facts for the time-marginals of the forward diffusion, viewed as Gaussian perturbations of the data distribution. The key structural point is that $\mathcal{L}(\overrightarrow{\mathbf{X}}_t)$ is obtained from π_{data} by Gaussian smoothing (in VE) or by a Gaussian smoothing followed by a deterministic down-scaling (in VP). This representation yields explicit identities expressing the score S_t and its Jacobian $\nabla\,\mathrm{S}_t$ in terms of conditional expectations of S_0 and $\nabla\,\mathrm{S}_0$ under the posterior $\mathcal{L}(\overrightarrow{\mathbf{X}}_0 \mid \overrightarrow{\mathbf{X}}_t = \mathbf{x})$. These formulas are the main tool to transfer growth, integrability, and dissipativity properties from π_{data} to $\mathcal{L}(\overrightarrow{\mathbf{X}}_t)$, which will be used later to establish contraction estimates in the metric ρ_b.

C.1. Gaussian representation

Following the forward SDE (1), the marginal law of the process at any time t admits an explicit Gaussian representation as a convolution of the initial distribution. This Gaussian convolution structure plays a central role in the analysis, as it allows us to transfer properties of the initial distribution to the time-marginal laws of the diffusion. We make this representation precise and derive the corresponding expressions in the following result.

Lemma C.1. *The solution of the forward process* (1) *writes as*

$$\overrightarrow{\mathbf{X}}_t \overset{\mathcal{L}}{=} \mathrm{m}_{t|0}\,\overrightarrow{\mathbf{X}}_0 + \sigma_{t|0} G\,, \tag{21}$$

with $\overrightarrow{\mathbf{X}}_0 \sim \pi_{\mathrm{data}}$*,* $G \sim \mathcal{N}(0, \mathbf{I}_d)$ *independent of* $\overrightarrow{\mathbf{X}}_0$*, and*

$$\mathrm{m}_{t|0} = \exp\left(-\alpha \int_0^t \beta_s \mathrm{d}s\right) \quad \text{and} \quad \sigma_{t|0}^2 = \mathrm{e}^{-2\alpha\int_0^t \beta_s \mathrm{d}s} \int_0^t \frac{2\beta_s}{\mathrm{e}^{-2\alpha\int_0^s \beta_u \mathrm{d}u}}\mathrm{d}s\,.$$

In particular, $\overrightarrow{\mathbf{X}}_t \overset{\mathcal{L}}{=} \overrightarrow{\mathbf{X}}_0 + \sigma_{t|0} G$ *in the VE case* ($\alpha = 0$) *and* $\overrightarrow{\mathbf{X}}_t \overset{\mathcal{L}}{=} \mathrm{m}_{t|0}\,\overrightarrow{\mathbf{X}}_0 + (\alpha^{-1}(1 - \mathrm{m}_{t|0}^2))^{1/2}\,G$ *in the VP case* ($\alpha > 0$)*. Furthermore, the probability density of the distribution of* $\mathbf{X}_s$ *given* $(\mathbf{X}_t, \mathbf{X}_0)$ *writes*

$$p_{s|t,0}\left(\mathbf{x}_s|\mathbf{x}_t, \mathbf{x}_0\right) = \mathcal{N}\left(\mathbf{x}_s; \mathrm{m}_{s|t}^0\,\mathbf{x}_0 + \mathrm{m}_{s|t}^t\,\mathbf{x}_t, \gamma_{s|t}^2 \mathbf{I}_d\right)\,, \tag{22}$$

with $\gamma_{s|t,0}^2 := \sigma_{t|s}^2 \sigma_{s|0}^2 / (\sigma_{t|s}^2\,\mathrm{m}_{t|s}^2 + \sigma_{s|0}^2)$ *and* $\mathrm{m}_{s|t,0}^t := \sqrt{(\gamma_{s|t,0}^2 - \sigma_{s|0}^2)/\sigma_{t|0}^2}$ *and* $\mathrm{m}_{s|t,0}^0 := \mathrm{m}_{s|0} - \mathrm{m}_{s|t,0}^t\,\mathrm{m}_{t|0}$.

Proof. Define $Y_t = \mathrm{e}^{\alpha\int_0^t \beta_s \mathrm{d}s}\overrightarrow{\mathbf{X}}_t$ such that, by Itô's lemma,

$$\mathrm{d}Y_t = \sqrt{2\beta_t}\mathrm{e}^{\alpha\int_0^t \beta_s \mathrm{d}s}\mathrm{d}\mathrm{B}_t\,.$$

Therefore,

$$Y_t = Y_0 + \int_0^t \sqrt{2\beta_s}\mathrm{e}^{\alpha\int_0^s \beta_u \mathrm{d}u}\mathrm{d}\mathrm{B}_s\,,$$

and

$$\overrightarrow{\mathbf{X}}_t = \mathrm{e}^{-\alpha\int_0^t \beta_s \mathrm{d}s}\overrightarrow{\mathbf{X}}_0 + \mathrm{e}^{-\alpha\int_0^t \beta_s \mathrm{d}s} \int_0^t \frac{\sqrt{2\beta_s}}{\mathrm{e}^{-\alpha\int_0^s \beta_u \mathrm{d}u}}\mathrm{d}\mathrm{B}_s\,. \tag{23}$$

Note that when $\alpha = 0$ and $\beta_s = \sigma_{s|0}\frac{\mathrm{d}\sigma_{s|0}}{\mathrm{d}s}$, $\int_0^t \sqrt{2\beta_s}\mathrm{dB}_s$ is a centered Gaussian vector with covariance matrix

$$\Sigma_t = \int_0^t 2\frac{\mathrm{d}\sigma_{s|0}}{\mathrm{d}s}\mathrm{d}s\mathbf{I}_d = \int_0^t \frac{\mathrm{d}\sigma^2_{s|0}}{\mathrm{d}s}\mathrm{d}s\mathbf{I}_d = \sigma^2_{t|0}\mathbf{I}_d\,.$$

Hence, for $Z \sim \mathcal{N}(0, \mathbf{I}_d)$ independent of $\overrightarrow{\mathbf{X}}_0$, $\overrightarrow{\mathbf{X}}_t \overset{\mathcal{L}}{=} \overrightarrow{\mathbf{X}}_0 + \sigma_{t|0}Z$. When $\alpha > 0$, the stochastic part of (23) is a centered Gaussian vector with covariance matrix,

$$\Sigma_t = \left(\mathrm{e}^{-2\alpha\int_0^t \beta_s \mathrm{d}s}\int_0^t 2\beta_s \mathrm{e}^{2\alpha\int_0^s \beta_u \mathrm{d}u}\mathrm{d}s\right)\mathbf{I}_d = \frac{1}{\alpha}\left(1 - \mathrm{e}^{-2\int_0^t \alpha\beta_s \mathrm{d}s}\right)\mathbf{I}_d = \frac{1}{\alpha}\left(1 - \mathrm{m}^2_{t|0}\right)\mathbf{I}_d\,,$$

which concludes the proof.

For the conditional distribution of $\mathbf{X}_s$ given $(\mathbf{X}_t, \mathbf{X}_0)$, note that the conditional distributions of $\mathbf{X}_s$ given $\mathbf{X}_0$, $\mathbf{X}_t$ given $\mathbf{X}_0$ and $\mathbf{X}_t$ given $\mathbf{X}_s$ admit a Gaussian density with respect to the Lebesgue measure. Thus, by Bayes theorem,

$$p_{s|t,0}\left(\mathbf{x}_s|\mathbf{x}_t, \mathbf{x}_0\right) = \frac{p_{t|s}\left(\mathbf{x}_t|\mathbf{x}_s\right)p_{s|0}\left(\mathbf{x}_s|\mathbf{x}_0\right)}{p_{t|0}\left(\mathbf{x}_t|\mathbf{x}_0\right)}\,. \tag{24}$$

Therefore, we obtain that

$$p_{s|t,0}\left(\mathbf{x}_s|\mathbf{x}_t, \mathbf{x}_0\right) = \mathcal{N}\left(\mathbf{x}_s; \mathrm{m}^t_{s|t,0}\,\mathbf{x}_t + \mathrm{m}^0_{s|t,0}\,\mathbf{x}_0, \gamma^2_{s|t,0}\mathbf{I}_d\right)\,, \tag{25}$$

with $\gamma^2_{s|t,0} = \sigma^2_{t|s}\sigma^2_{s|0}/(\sigma^2_{t|s}\,\mathrm{m}^2_{t|s} + \sigma^2_{s|0})$, $\mathrm{m}^t_{s|t,0} = \sqrt{(\gamma^2_{s|t,0} - \sigma^2_{s|0})/\sigma^2_{t|0}}$ and $\mathrm{m}^0_{s|t,0} = \mathrm{m}_{s|0} - \mathrm{m}^t_{s|t,0}\,\mathrm{m}_{t|0}$. □

From the Gaussian representation (21) in the case of $\alpha > 0$, we may rewrite

$$\overrightarrow{\mathbf{X}}_t \;\overset{\mathcal{L}}{=}\; \mathrm{m}_{t|0}\left(\overrightarrow{\mathbf{X}}_0 + \varepsilon_t\,G\right), \qquad \varepsilon_t^2 := \frac{\sigma^2_{t|0}}{\mathrm{m}^2_{t|0}},$$

where $G \sim \mathcal{N}\left(0, \mathbf{I}_d\right)$ is independent of $\overrightarrow{\mathbf{X}}_0$. In other words, up to the scaling factor $\mathrm{m}_{t|0}$, the marginal at time t is a Gaussian perturbation of $\overrightarrow{\mathbf{X}}_0$ with mean 0 and covariance matrix $\varepsilon_t^2\mathbf{I}_d = (\sigma^2_{t|0}/\,\mathrm{m}^2_{t|0})\mathbf{I}_d$. Moreover, there exists a simple relation satisfied by $\overrightarrow{\mathbf{X}}_0 + \varepsilon G$ and $m\big(\overrightarrow{\mathbf{X}}_0 + \varepsilon G\big)$ for $m \in (0, 1)$, as stated and proved in the following lemma. For this reason, in the following, we will focus on the VE case, and derive bounds for $\alpha > 0$ afterwards.

Lemma C.2. *Let $m \in (0, 1)$ and q_ϵ (resp. $\bar{q}_\epsilon$) be the density function of the random variable $\mathbf{X}_0 + \epsilon G$ (resp. $m(\mathbf{X}_0 + \epsilon G)$), where $\mathbf{X}_0 \sim p_{\mathrm{data}}$ and $G \sim \mathcal{N}\left(0, \mathbf{I}_d\right)$ is independent of $\mathbf{X}_0$. Then, for all $\mathbf{x} \in \mathbb{R}^d$, the associated score function is given by*

$$\nabla \log q_\epsilon(\mathbf{x}) = \frac{1}{m}\nabla \log \bar{q}_\epsilon\left(\frac{\mathbf{x}}{m}\right)\,. \tag{26}$$

Proof. By construction, the two random variables are one the scaled version of the other. Hence, by a linear change of variables, the probability density function of the above writes as $m^{-d}p_t\left(\mathbf{x}/m\right)$, therefore the desired identity is simply obtained applying the logarithmic function and the derivative. □

Remark C.3. Note that if one uses the same β_t for both VE framework ($\alpha = 0$) and the general case $\alpha > 0$, the relation (26) does not hold for *the same t*. Indeed, for a given t in the VP framework, one must choose $\mathrm{m}_{t|0}$ such that $\mathrm{m}_{t|0} = \exp\left(-\int_0^t \alpha\beta_s\mathrm{d}s\right)$ and at the same find a time $\tau\left(t\right)$ such that

$$\mathrm{m}^2_{t|0}\,\sigma^2_{\tau(t)|0} = \sigma^2_{t|0} = \frac{1}{\alpha}\left(1 - \mathrm{m}^2_{t|0}\right)\,,$$

which implies $\int_0^{\tau(t)} \beta_s\mathrm{d}s = \int_0^t \beta_s \exp(2\alpha\int_0^s \beta_u\mathrm{d}u)\mathrm{d}s$. We shall not refrain from using Lemma C.2 to establish the link between VE and $\alpha > 0$ frameworks, considering the parameter $\sigma^2_{t|0}$ in the VE case as a free parameter.

C.2. Properties of the score under Gaussian perturbation

We study how various properties of the score function are preserved under Gaussian perturbations. These results are crucial for analyzing the forward diffusion process and its associated score functions in both the VE and $\alpha > 0$ frameworks. Due to the previous connection between simply adding Gaussian noise and the VE framework ($\alpha = 0$), we will consider the notations defined for the VE framework for presenting the results. Introduce also the denoiser application $\mathrm{D}_t(\mathbf{x}) = \mathbb{E}\left[\overrightarrow{\mathbf{X}}_0 \middle| \overrightarrow{\mathbf{X}}_t = \mathbf{x}\right]$.

In this section we present only relations between p_t and p_0 thus one must interpret $\sigma_{0|t}$ as a free parameter. We will then rely on the following link between both frameworks to translate results in VE to results for VP.

Lemma C.4. *Assume that Assumption 3.1 (I) holds. Then, for all $\mathbf{x} \in \mathbb{R}^d$,*

$$\begin{aligned} \mathrm{S}_t\left(\mathbf{x}\right) &= -\frac{1}{\sigma_{t|0}^2}\left(\mathbf{x} - \mathrm{D}_t(\mathbf{x})\right) && (27) \\ &= \mathbb{E}\left[\mathrm{S}_0\left(\overrightarrow{\mathbf{X}}_0\right) \middle| \overrightarrow{\mathbf{X}}_t = \mathbf{x}\right] . && (28) \end{aligned}$$

Proof. The density of $\mathbf{X}_t$ is the convolution

$$p_t\left(\mathbf{x}\right) = \int p_0\left(y\right)\, \varphi_t\left(\mathbf{x} - y\right)\, \mathrm{d}y,$$

where φ_t is the density of $\mathcal{N}\left(0, \sigma_{t|0}^2 \mathbf{I}_d\right)$. Differentiating under the integral,

$$\nabla p_t\left(\mathbf{x}\right) = \int p_0\left(y\right)\, \nabla_{\mathbf{x}} \varphi_t\left(\mathbf{x} - y\right)\, \mathrm{d}y = -\frac{1}{\sigma_{t|0}^2} \int \left(\mathbf{x} - y\right) p_0\left(y\right)\, \varphi_t\left(\mathbf{x} - y\right)\, \mathrm{d}y.$$

Hence

$$\mathrm{S}_t\left(\mathbf{x}\right) = \frac{\nabla p_t\left(\mathbf{x}\right)}{p_t\left(\mathbf{x}\right)} = -\frac{1}{\sigma_{t|0}^2}\left(\mathbf{x} - \mathbb{E}\left[\mathbf{X}_0 | \mathbf{X}_t = \mathbf{x}\right]\right),$$

which proves (27). For (28), note that, from Bayes' formula for the posterior distribution of $\mathbf{X}_0$ given $\mathbf{X}_t = \mathbf{x}$,

$$\begin{aligned} \mathbb{E}\left[\mathrm{S}_0\left(\overrightarrow{\mathbf{X}}_0\right) \middle| \overrightarrow{\mathbf{X}}_t = \mathbf{x}\right] &= \frac{1}{p_t\left(\mathbf{x}\right)} \int \mathrm{S}_0\left(y\right)\, p_0\left(y\right)\, \varphi_t\left(\mathbf{x} - y\right)\, \mathrm{d}y \\ &= \frac{1}{p_t\left(\mathbf{x}\right)} \int \nabla_y p_0\left(y\right)\, \varphi_t\left(\mathbf{x} - y\right)\, \mathrm{d}y\, . \end{aligned}$$

Under Assumption 3.1 (I), the sub-Gaussian tail behavior established in Lemma A.1 guarantees the validity of integration by parts with respect to y, with vanishing boundary contributions. This gives

$$\int \nabla p_0\left(y\right)\, \varphi_t\left(\mathbf{x} - y\right)\, \mathrm{d}y = -\int p_0\left(y\right)\, \nabla_y \varphi_t\left(\mathbf{x} - y\right)\, \mathrm{d}y.$$

Noting that $\nabla_y \varphi_t\left(\mathbf{x} - y\right) = -\nabla_{\mathbf{x}} \varphi_t\left(\mathbf{x} - y\right)$, we obtain

$$\int \nabla p_0\left(y\right)\, \varphi_t\left(\mathbf{x} - y\right)\, \mathrm{d}y = \int p_0\left(y\right)\, \nabla_{\mathbf{x}} \varphi_t\left(\mathbf{x} - y\right)\, \mathrm{d}y = \nabla_{\mathbf{x}} p_t\left(x\right).$$

Thus,

$$\mathbb{E}\left[\mathrm{S}_0\left(\mathbf{X}_0\right) | \mathbf{X}_t = \mathbf{x}\right] = \frac{\nabla_{\mathbf{x}} p_t\left(\mathbf{x}\right)}{p_t\left(\mathbf{x}\right)} = \mathrm{S}_t\left(\mathbf{x}\right),$$

which proves (28). □

Lemma C.5. *Assume that Assumption 3.1 (I) holds. Then, for all $\mathbf{x} \in \mathbb{R}^d$,*

$$\nabla \mathrm{S}_t\left(\mathbf{x}\right) + \mathrm{S}_t\left(\mathbf{x}\right) \mathrm{S}_t\left(\mathbf{x}\right)^{\mathrm{T}} = \mathbb{E}\left[\nabla \mathrm{S}_0\left(\overrightarrow{\mathbf{X}}_0\right) + \mathrm{S}_0\left(\overrightarrow{\mathbf{X}}_0\right) \mathrm{S}_0\left(\overrightarrow{\mathbf{X}}_0\right)^{\mathrm{T}} \middle| \overrightarrow{\mathbf{X}}_t = \mathbf{x}\right], \qquad (29)$$

and

$$\nabla \mathrm{S}_t(\mathbf{x}) = \mathrm{Var}\big(\mathrm{S}_0\left(\overrightarrow{\mathbf{X}}_0\right) \mid \overrightarrow{\mathbf{X}}_t = \mathbf{x}\big) + \mathbb{E}\left[\nabla \mathrm{S}_0\left(\overrightarrow{\mathbf{X}}_0\right)\middle|\overrightarrow{\mathbf{X}}_t = \mathbf{x}\right] \tag{30}$$

$$= \frac{1}{\sigma_{t|0}^2}\left(\mathbb{E}\left[\mathrm{S}_0\left(\overrightarrow{\mathbf{X}}_0\right)\overrightarrow{\mathbf{X}}_0^{\mathrm{T}}\middle|\overrightarrow{\mathbf{X}}_t = \mathbf{x}\right] - \mathrm{S}_t(\mathbf{x})\,\mathrm{D}_t(\mathbf{x})^{\mathrm{T}}\right) \tag{31}$$

$$= \frac{1}{\sigma_{t|0}^4}\mathrm{Var}\left(\overrightarrow{\mathbf{X}}_0 \mid \overrightarrow{\mathbf{X}}_t = \mathbf{x}\right) - \frac{1}{\sigma_{t|0}^2}\mathbf{I}_d\,. \tag{32}$$

Proof. Under Assumption 3.1 (I), from Lemma A.1, the Gaussian convolutions are smooth and all integrations by parts below are justified (the boundary terms vanish). First,

$$\nabla \mathrm{S}_t(\mathbf{x}) = \nabla^2 \log p_t(\mathbf{x}) = \frac{\nabla^2 p_t(\mathbf{x})}{p_t(\mathbf{x})} - \mathrm{S}_t(\mathbf{x})\,\mathrm{S}_t(\mathbf{x})^{\mathrm{T}}. \tag{33}$$

Differentiating twice under the integral and integrating by parts in y, we obtain

$$\nabla^2 p_t(\mathbf{x}) = \int \nabla^2 p_0(y)\ \varphi_t(\mathbf{x} - y)\ \mathrm{d}y.$$

Moreover,

$$\nabla^2 p_0(y) = \left(\nabla \mathrm{S}_0(y) + \mathrm{S}_0(y)\,\mathrm{S}_0(y)^{\mathrm{T}}\right) p_0(y)\,,$$

so that we get

$$\frac{\nabla^2 p_t(\mathbf{x})}{p_t(\mathbf{x})} = \mathbb{E}\left[\nabla \mathrm{S}_0\left(\overrightarrow{\mathbf{X}}_0\right) + \mathrm{S}_0\left(\overrightarrow{\mathbf{X}}_0\right)\mathrm{S}_0\left(\overrightarrow{\mathbf{X}}_0\right)^{\mathrm{T}}\middle|\overrightarrow{\mathbf{X}}_t = \mathbf{x}\right].$$

Plugging this into (33) yields (29). Moreover, using (28) from Lemma C.4, we get (30).

To prove (31), recall from (28) that,

$$\mathrm{S}_t(\mathbf{x}) = \mathbb{E}\left[\mathrm{S}_0\left(\overrightarrow{\mathbf{X}}_0\right)\middle|\overrightarrow{\mathbf{X}}_t = \mathbf{x}\right],$$

and write,

$$\mathrm{S}_t(\mathbf{x}) = \frac{N(\mathbf{x})}{p_t(\mathbf{x})}, \qquad N(\mathbf{x}) := \int \mathrm{S}_0(y)\,p_0(y)\ \varphi_t(\mathbf{x} - y)\,\mathrm{d}y\,.$$

It follows that

$$\nabla \mathrm{S}_t(\mathbf{x}) = \frac{\nabla N(\mathbf{x})}{p_t(\mathbf{x})} - \mathrm{S}_t(\mathbf{x})\,\frac{\nabla p_t(\mathbf{x})^{\mathrm{T}}}{p_t(\mathbf{x})}$$
$$= \frac{\nabla N(\mathbf{x})}{p_t(\mathbf{x})} - \mathrm{S}_t(\mathbf{x})\,\mathrm{S}_t(\mathbf{x})^{\mathrm{T}}.$$

Since $\nabla_{\mathbf{x}}\varphi_t(\mathbf{x} - y) = -(\mathbf{x} - y)\sigma_{t|0}^{-2}\varphi_t(\mathbf{x} - y)$, we have

$$\frac{\nabla N(\mathbf{x})}{p_t(\mathbf{x})} = -\frac{1}{\sigma_{t|0}^2}\mathbb{E}\left[\mathrm{S}_0\left(\overrightarrow{\mathbf{X}}_0\right)(\mathbf{x} - \mathbf{X}_0)^{\mathrm{T}}\middle|\overrightarrow{\mathbf{X}}_t = \mathbf{x}\right]$$
$$= \frac{1}{\sigma_{t|0}^2}\left(\mathbb{E}\left[\mathrm{S}_0\left(\overrightarrow{\mathbf{X}}_0\right)\overrightarrow{\mathbf{X}}_0^{\mathrm{T}}\middle|\overrightarrow{\mathbf{X}}_t = \mathbf{x}\right] - \mathbb{E}\left[\mathrm{S}_0\left(\overrightarrow{\mathbf{X}}_0\right)\middle|\overrightarrow{\mathbf{X}}_t = \mathbf{x}\right]\mathbf{x}^{\mathrm{T}}\right)$$
$$= \frac{1}{\sigma_{t|0}^2}\left(\mathbb{E}\left[\mathrm{S}_0\left(\overrightarrow{\mathbf{X}}_0\right)\overrightarrow{\mathbf{X}}_0^{\mathrm{T}}\middle|\overrightarrow{\mathbf{X}}_t = \mathbf{x}\right] - \mathrm{S}_t(\mathbf{x})\,\mathbf{x}^{\mathrm{T}}\right).$$

Then, using (27),

$$\mathrm{S}_t(\mathbf{x})\,\mathbf{x}^{\mathrm{T}} = -\sigma_{t|0}^2\,\mathrm{S}_t(\mathbf{x})\,\mathrm{S}_t(\mathbf{x})^{\mathrm{T}} + \mathrm{S}_t(\mathbf{x})\,\mathrm{D}_t(\mathbf{x})^{\mathrm{T}}.$$

Therefore,

$$\nabla \mathrm{S}_t(\mathbf{x}) = \frac{1}{\sigma_{t|0}^2}\left(\mathbb{E}\left[\mathrm{S}_0\left(\overrightarrow{\mathbf{X}}_0\right)\mathbf{X}_0{}^{\mathrm{T}}\middle|\overrightarrow{\mathbf{X}}_t = \mathbf{x}\right] - \mathrm{S}_t(\mathbf{x})\,\mathrm{D}_t(\mathbf{x})^{\mathrm{T}}\right),$$

which proves (31). Finally, similar computations as for $\nabla \mathrm{S}_t(\mathbf{x})$ yields,

$$\begin{aligned}\nabla \mathrm{D}_t(\mathbf{x}) &= \frac{1}{\sigma_{t|0}^2}\left(\mathbb{E}\left[\overrightarrow{\mathbf{X}}_0\overrightarrow{\mathbf{X}}_0{}^{\mathrm{T}}\middle|\overrightarrow{\mathbf{X}}_t = \mathbf{x}\right] - \mathrm{D}_t(\mathbf{x})\mathrm{D}_t(\mathbf{x})^{\mathrm{T}}\right)\\ &= \frac{1}{\sigma_{t|0}^2}\mathrm{Var}\big(\overrightarrow{\mathbf{X}}_0 \mid \overrightarrow{\mathbf{X}}_t = \mathbf{x}\big),\end{aligned}$$

where the last equality follows from the definition of D_t. Rearranging (27) we get,

$$\mathrm{D}_t(\mathbf{x}) = \sigma_{t|0}^2\,\mathrm{S}_t(\mathbf{x}) + \mathbf{x},$$

which proves (32) and finishes the proof. □

D. Stability of the data assumptions along the diffusion flow and Lyapunov contraction

This section shows that the structural conditions imposed at time 0 remain stable along the diffusion. First, we prove that the dissipativity inequality in Assumption 3.1 (I) propagates to the forward-time score S_t with explicit time-dependent constants. Second, we leverage this propagated dissipativity to establish a Lyapunov drift condition for the backward Markov semigroup $\mathrm{Q}_{t|s}$, which is the key ingredient needed to apply Harris-type theorems (Hairer & Mattingly, 2008). Finally, we show that polynomial growth controls on the score and its Jacobian (as in Assumption 3.1 (II)) are preserved under the forward flow yielding uniform-in-time bounds used throughout the paper.

D.1. Propagation of the dissipativity condition

The properties of the score function established in the previous section allow us to study the stability of the Lyapunov condition stated in Assumption 3.1 (I) through the diffusion dynamics, *i.e.*, from S_0 to S_t. The next proposition makes this transfer explicit by providing time-dependent dissipativity parameters γ_t and κ_t.

Proposition D.1. *Suppose that Assumption 3.1 (I) holds. Then, for all $t > 0$ and $\mathbf{x} \in \mathbb{R}^d$, there exist continuous functions $\gamma_t, \kappa_t : [0, T] \to \mathbb{R}_+$ such that,*

$$\langle \mathrm{S}_t(\mathbf{x}), \mathbf{x}\rangle \le -\gamma_t \|\mathbf{x}\|^2 + \kappa_t, \tag{34}$$

where γ_t and κ_t are defined as follows for $t \in (0, T]$

- *If $\alpha = 0$ (Variance-Exploding case),*

$$\gamma_t = \frac{\gamma_0}{1 + 2\gamma_0\sigma_{t|0}^2}, \qquad \kappa_t = \frac{\kappa_0 + d}{1 + 2\gamma_0\sigma_{t|0}^2}.$$

- *If $\alpha > 0$,*

$$\gamma_t = \frac{\gamma_0\alpha}{\alpha\,\mathrm{m}_{t|0}^2 + 2\gamma_0\left(1 - \mathrm{m}_{t|0}^2\right)}, \qquad \kappa_t = (\kappa_0 + d)\frac{\mathrm{m}_{t|0}^2\,\alpha}{\alpha\,\mathrm{m}_{t|0}^2 + 2\gamma_0\left(1 - \mathrm{m}_{t|0}^2\right)},$$

Proof. *Case $\alpha > 0$.* Computing the trace of (31) yields, for all $\mathbf{x} \in \mathbb{R}^d$, $t > 0$,

$$\mathrm{div}(\mathrm{S}_t(\mathbf{x})) = \frac{1}{\sigma_{t|0}^2}\left(\mathbb{E}\left[\left\langle \mathrm{S}_0\left(\overrightarrow{\mathbf{X}}_0\right), \overrightarrow{\mathbf{X}}_0\right\rangle\middle|\overrightarrow{\mathbf{X}}_t = \mathbf{x}\right] - \langle \mathrm{S}_t(\mathbf{x}), \mathrm{D}_t(\mathbf{x})\rangle\right).$$

Moreover, using (27),

$$\langle \mathrm{S}_t(\mathbf{x}), \mathrm{D}_t(\mathbf{x})\rangle = \langle \mathrm{S}_t(\mathbf{x}), \mathbf{x}\rangle + \sigma_{t|0}^2\|\mathrm{S}_t(\mathbf{x})\|^2,$$

together with the fact that

$$\operatorname{div}(\mathrm{S}_t(\mathbf{x})) + \|\mathrm{S}_t(\mathbf{x})\|^2 = \frac{\Delta p_t(\mathbf{x})}{p_t(\mathbf{x})}, \tag{35}$$

we get

$$\langle \mathrm{S}_t(\mathbf{x}), \mathbf{x} \rangle = \mathbb{E}\left[\left\langle \mathrm{S}_t\left(\overrightarrow{\mathbf{X}}_0\right), \overrightarrow{\mathbf{X}}_0 \right\rangle \middle| \overrightarrow{\mathbf{X}}_t = \mathbf{x} \right] - \sigma_{t|0}^2 \frac{\Delta p_t(\mathbf{x})}{p_t(\mathbf{x})}.$$

It follows from Assumption 3.1 (I) that

$$\langle \mathrm{S}_t(\mathbf{x}), \mathbf{x} \rangle \leq -\gamma_0 \mathbb{E}\left[\left\| \overrightarrow{\mathbf{X}}_0 \right\|^2 \middle| \overrightarrow{\mathbf{X}}_t = \mathbf{x} \right] + \kappa_0 - \sigma_{t|0}^2 \frac{\Delta p_t(\mathbf{x})}{p_t(\mathbf{x})}.$$

Combining $\|a\|^2 = \|b\|^2 + \|a-b\|^2 + 2\langle a-b, b\rangle$ and (27), we have

$$\begin{aligned}
\mathbb{E}\left[\left\| \overrightarrow{\mathbf{X}}_0 \right\|^2 \middle| \overrightarrow{\mathbf{X}}_t = \mathbf{x} \right] &= \|\mathbf{x}\|^2 + \mathbb{E}\left[\left\| \overrightarrow{\mathbf{X}}_0 - \mathbf{x} \right\|^2 \middle| \overrightarrow{\mathbf{X}}_t = \mathbf{x} \right] + 2\sigma_{t|0}^2 \langle \mathrm{S}_t(\mathbf{x}), \mathbf{x} \rangle, \\
&\geq \|\mathbf{x}\|^2 + 2\sigma_{t|0}^2 \langle \mathrm{S}_t(\mathbf{x}), \mathbf{x} \rangle.
\end{aligned}$$

It follows

$$\langle \mathrm{S}_t(\mathbf{x}), \mathbf{x} \rangle \leq -\gamma_0 \|\mathbf{x}\|^2 - 2\gamma_0 \sigma_{t|0}^2 \langle \mathrm{S}_t(\mathbf{x}), \mathbf{x} \rangle + \kappa_0 - \sigma_{t|0}^2 \frac{\Delta p_t(\mathbf{x})}{p_t(\mathbf{x})},$$

which yields

$$\langle \mathrm{S}_t(\mathbf{x}), \mathbf{x} \rangle \leq -\frac{\gamma_0}{1 + 2\gamma_0 \sigma_{t|0}^2} \|\mathbf{x}\|^2 + \frac{1}{1 + 2\gamma_0 \sigma_{t|0}^2} \left(\kappa_0 - \sigma_{t|0}^2 \frac{\Delta p_t(\mathbf{x})}{p_t(\mathbf{x})} \right).$$

Using identity (35),

$$-\sigma_{t|0}^2 \frac{\Delta p_t(\mathbf{x})}{p_t(\mathbf{x})} = -\sigma_{t|0}^2 \left(\operatorname{Tr}(\nabla \mathrm{S}_t)(\mathbf{x}) + \|\mathrm{S}_t(\mathbf{x})\|^2 \right).$$

This with (32) yields

$$-\sigma_{t|0}^2 \frac{\Delta p_t(\mathbf{x})}{p_t(\mathbf{x})} = d - \frac{1}{\sigma_{t|0}^2} \operatorname{Tr}\left(\mathbb{V}\left[\overrightarrow{\mathbf{X}}_0 \middle| \overrightarrow{\mathbf{X}}_t = \mathbf{x} \right] \right) - \sigma_{t|0}^2 \|\mathrm{S}_t(\mathbf{x})\|^2 \leq d,$$

which concludes the proof.

Case $\alpha > 0$. As noted in Remark C.3, the general case $\alpha > 0$ can be deduced from the VE case by a suitable time change. With abuse of notation, write ${\sigma_{t|0}}^{\mathrm{VE}} = \sigma_{t|0} / \mathrm{m}_{t|0}$. By Lemma C.2, we get that

$$\begin{aligned}
\langle \mathrm{S}_t(\mathbf{x}), \mathbf{x} \rangle &= \left\langle \mathrm{S}_t\left(\frac{\mathbf{x}}{\mathrm{m}_{t|0}} \right), \frac{\mathbf{x}}{\mathrm{m}_{t|0}} \right\rangle \\
&\leq \frac{\gamma_0}{1 + 2\gamma_0 (\sigma_{t|0}/\mathrm{m}_{t|0})^2} \left\| \frac{\mathbf{x}}{\mathrm{m}_{t|0}} \right\|^2 + \frac{1}{1 + 2\gamma_0 (\sigma_{t|0}/\mathrm{m}_{t|0})^2} \left(\kappa_0 - (\sigma_{t|0}/\mathrm{m}_{t|0})^2 \frac{\Delta p_t\left(\mathbf{x}/\mathrm{m}_{t|0}\right)}{p_t\left(\mathbf{x}/\mathrm{m}_{t|0}\right)} \right),
\end{aligned}$$

which concludes the proof. □

Remark D.2. From Lemma C.1 and the explicit expressions of $\mathrm{m}_{t|0}$ and $\sigma_{t|0}$, it is straightforward to check that the condition $\gamma_0 > \alpha/2$ propagates in time. Indeed, for all $t \geq 0$,

$$\gamma_t > \frac{\alpha}{2} \iff \gamma_0 > \frac{\alpha}{2} \left(\mathrm{m}_{t|0}^2 + \frac{2\gamma_0}{\alpha} \left(1 - \mathrm{m}_{t|0}^2 \right) \right) \iff \gamma_0 > \frac{\alpha}{2} \mathrm{m}_{t|0}^2 + \gamma_0 \left(1 - \mathrm{m}_{t|0}^2 \right) \iff \left(\gamma_0 - \frac{\alpha}{2} \right) \mathrm{m}_{t|0}^2 > 0,$$

so that the corresponding time-dependent dissipativity parameter remains larger than the threshold $\alpha/2$ uniformly over $t \geq 0$.

D.2. Lyapunov contraction for the backward semigroup (Proposition 3.2)

We now turn to the time-reversed dynamics (2) and its associated semigroup (3). Using the propagated dissipativity of S_{T-t}, we derive a drift inequality for polynomial Lyapunov functions $\mathrm{V}_\ell(x) = \|x\|^\ell$. This is a key ingredient for Harris theorem Hairer & Mattingly (2008) that we use with $\ell = 2$.

Proposition 3.2. (restatement). *Suppose that Assumption 3.1 (I) holds. Let $\mathrm{V}_\ell(\mathbf{x}) := \|\mathbf{x}\|^\ell$, for $\ell \geq 2$. Then, there exist continuous functions $\tilde{\gamma}_{\cdot,\ell}, \tilde{\kappa}_{\cdot,\ell} : [0,T] \to \mathbb{R}_+$ such that, for all $0 \leq s < t \leq T$ and all $\mathbf{x} \in \mathbb{R}^d$,*

$$\mathrm{Q}_{t|s}\,\mathrm{V}_\ell(\mathbf{x}) \leq \lambda^{(\ell)}_{t|s}\,\mathrm{V}_\ell(\mathbf{x}) + \mathrm{K}^{(\ell)}_{t|s}\,.$$

where $\lambda^{(\ell)}_{t|s} := \exp\left(-\int_s^t \tilde{\gamma}_{v,\ell}\mathrm{d}v\right)$ and $\mathrm{K}^{(\ell)}_{t|s} := \int_s^t \exp\left(-\int_u^t \tilde{\gamma}_{v,\ell}\mathrm{d}v\right)\tilde{\kappa}_{u,\ell}\mathrm{d}u$. In particular, when $\ell = 2$,

$$\tilde{\gamma}_{t,2} = 2\bar{\beta}_t\left(2\gamma_{T-t} - \alpha\right), \quad \textit{and} \quad \tilde{\kappa}_{t,2} := 2\bar{\beta}_t\left(2\kappa_{T-t} + d\right)\,.$$

Proof of Proposition 3.2. Note that for all $\mathbf{x} \in \mathbb{R}^d$ and all $t \in [0,T]$,

$$\nabla\,\mathrm{V}_\ell(\mathbf{x}) = \ell\,\|\mathbf{x}\|^{\ell-2}\,\mathbf{x}\,, \qquad \Delta\,\mathrm{V}_\ell(\mathbf{x}) = \ell(\ell-2+d)\,\|\mathbf{x}\|^{\ell-2}\,.$$

The infinitesimal generator associated with (2) satisfies

$$\mathcal{A}_t\,\mathrm{V}_\ell(\mathbf{x}) = \bar{\beta}_t\left\langle \alpha\mathbf{x} + 2\,\mathrm{S}_{T-t}\left(\mathbf{x}\right), \nabla\,\mathrm{V}_\ell(\mathbf{x})\right\rangle + \bar{\beta}_t\Delta\,\mathrm{V}_\ell(\mathbf{x})\,.$$

From Proposition D.1 together with Assumption 3.1 (I) there exist continuous functions γ, κ: $[0,T] \to \mathbb{R}_+$ such that, for all $t \in [0,T]$ and all $\mathbf{x} \in \mathbb{R}^d$,

$$\left\langle \mathrm{S}_t\left(\mathbf{x}\right), \mathbf{x}\right\rangle \leq -\gamma_t\,\|\mathbf{x}\|^2 + \kappa_t\,. \tag{36}$$

As a consequence,

$$\mathcal{A}_t\,\mathrm{V}_\ell(\mathbf{x}) \leq -a_t\,\mathrm{V}_\ell(\mathbf{x}) + b_t\,\|\mathbf{x}\|^{\ell-2}$$

where

$$a_t := \ell\bar{\beta}_t\left(2\gamma_{T-t} - \alpha\right), \quad b_t := \ell\bar{\beta}_t\left(2\kappa_{T-t} + \ell - 2 + d\right),$$

which finished the proof for $\ell = 2$. Moreover, if $\ell > 2$, for any $\eta_t > 0$, we can apply Young's inequality with conjugated exponents $\bar{p}, \bar{q}$ such that $1/\bar{p} = (\ell-2)/\ell$ and $1/\bar{q} = 2/\ell$. This gives

$$\|\mathbf{x}\|^{\ell-2} = \eta_t^{\frac{\ell-2}{\ell}}\,\|\mathbf{x}\|^{\ell-2}\,\eta_t^{-\frac{\ell-2}{\ell}} \leq \frac{\ell-2}{\ell}\eta_t\,\|\mathbf{x}\|^\ell + \frac{2}{\ell}\,\eta_t^{-\frac{\ell-2}{2}}\,.$$

Take $\eta_t > 0$ such that $a_t - \frac{\ell-2}{\ell}\eta_t b_t > a_t/2$. Thus,

$$\mathcal{A}_t\,\mathrm{V}_\ell(\mathbf{x}) \leq -\left(a_t - \frac{\ell-2}{\ell}\eta_t b_t\right)\mathrm{V}_\ell(\mathbf{x}) + b_t\,\frac{2}{\ell}\,\eta_t^{-\frac{\ell-2}{2}} \leq -\frac{a_t}{2}\,\mathrm{V}_\ell(\mathbf{x}) + b_t\,\frac{2}{\ell}\,\eta_t^{-\frac{\ell-2}{2}}\,.$$

It follows from Dynkin's formula that, for all $0 \leq s < t \leq T$,

$$\mathrm{Q}_{t|s}\,\mathrm{V}_\ell(\mathbf{x}) \leq \mathrm{V}_\ell(\mathbf{x}) + \int_s^t \left(-\tilde{\gamma}_{v,\ell}\mathrm{Q}_{v|s}\,\mathrm{V}_\ell(\mathbf{x}) + \tilde{\kappa}_{v,\ell}\right)\mathrm{d}v\,.$$

where

$$\tilde{\gamma}_{t,\ell} := \frac{a_t}{2}\,, \qquad \tilde{\kappa}_{t,\ell} := b_t\,\frac{2}{\ell}\,\eta_t^{-\frac{\ell-2}{2}}\,.$$

The claimed bound then follows from Grönwall's lemma. □

D.3. Propagation of the growth condition

We aim at showing the propagation of Assumption 3.1 (II) along the flow. To this end, we first study the time stability of the quantity $\nabla \mathrm{S}_t(x) + \mathrm{S}_t(x)\,\mathrm{S}_t(x)^{\mathrm{T}}$ in the VE setting, following an approach similar to that developed in Section C.2. The term plays a central role in the Lyapunov stability analysis. Indeed, as a direct consequence of Lemma C.5 and Corollary A.3, this combination naturally appears when controlling the drift and growth properties of the dynamics, and thus governs the stability estimates. The following lemma further clarifies its interpretation by showing that this quantity is tightly connected to the conditional variance of the denoiser $\overrightarrow{\mathbf{X}}_0 \mid \overrightarrow{\mathbf{X}}_t$, providing a probabilistic meaning to the key terms involved in the Lyapunov analysis.

Lemma D.3. *Assume that Assumption 3.1 (I) holds. Then, for all $\mathbf{x} \in \mathbb{R}^d$,*

$$\nabla \mathrm{S}_t(\mathbf{x}) + \mathrm{S}_t(\mathbf{x})\,\mathrm{S}_t(\mathbf{x})^{\mathrm{T}} = \frac{1}{\sigma_{t|0}^4}\operatorname{Var}\Big(\overrightarrow{\mathbf{X}}_0 \mid \overrightarrow{\mathbf{X}}_t = \mathbf{x}\Big) - \frac{1}{\sigma_{t|0}^2}\mathbf{I}_d\,. \tag{37}$$

Proof. Let φ_t be the density of $\mathcal{N}\left(0, \sigma_{t|0}^2 \mathbf{I}_d\right)$. Then,

$$\nabla \varphi_t(\mathbf{x}) = -\frac{1}{\sigma_{t|0}^2}\mathbf{x}\,\varphi_t(\mathbf{x})\ , \qquad \nabla^2 \varphi_t(\mathbf{x}) = \left(\frac{1}{\sigma_{t|0}^4}\mathbf{x}\mathbf{x}^{\mathrm{T}} - \frac{1}{\sigma_{t|0}^2}\mathbf{I}_d\right)\varphi_t(\mathbf{x})\ .$$

Since

$$p_t(\mathbf{x}) = \int_{\mathbb{R}^d} p_0(y)\ \varphi_t(\mathbf{x} - y)\ dy\,,$$

As in Lemma C.4, from Assumption 3.1 (I) and Lemma A.1, we can differentiate under the integral sign to get

$$\nabla p_t(\mathbf{x}) = \int p_0(y)\ \nabla \varphi_t(\mathbf{x} - y)\ dy, \qquad \nabla^2 p_t(\mathbf{x}) = \int p_0(y)\ \nabla^2 \varphi_t(\mathbf{x} - y)\ dy.$$

Similarly to the computations in Lemma C.4, applying Bayes' formula together with the previous integration by parts formulae yields

$$\frac{\nabla^2 p_t(\mathbf{x})}{p_t(\mathbf{x})} = \mathbb{E}\left[\left(\frac{1}{\sigma_{t|0}^4}(\mathbf{x} - \overrightarrow{\mathbf{X}}_0)(\mathbf{x} - \overrightarrow{\mathbf{X}}_0)^{\mathrm{T}} - \frac{1}{\sigma_{t|0}^2}\mathbf{I}_d\right)\middle|\,\overrightarrow{\mathbf{X}}_t = \mathbf{x}\right]. \tag{38}$$

Using $\mathrm{S}_t(\mathbf{x}) = \nabla \log p_t(\mathbf{x}) = \nabla p_t(\mathbf{x})/p_t(\mathbf{x})$ and the formula $\nabla \mathrm{S}_t(\mathbf{x}) = \nabla^2 \log p_t(\mathbf{x}) = \nabla^2 p_t(\mathbf{x})/p_t(\mathbf{x}) - \nabla \log p_t(\mathbf{x})\,\nabla \log p_t(\mathbf{x})^{\mathrm{T}}$, we obtain

$$\nabla \mathrm{S}_t(\mathbf{x}) + \mathrm{S}_t(\mathbf{x})\,\mathrm{S}_t(\mathbf{x})^{\mathrm{T}} = \mathbb{E}\left[\left(\frac{1}{\sigma_{t|0}^4}(\mathbf{x} - \overrightarrow{\mathbf{X}}_0)(\mathbf{x} - \overrightarrow{\mathbf{X}}_0)^{\mathrm{T}} - \frac{1}{\sigma_{t|0}^2}\mathbf{I}_d\right)\middle|\,\mathbf{X}_t = \mathbf{x}\right].$$

Finally, since $\mathbb{E}\left[\mathbf{x} - \mathbf{X}_0 | \mathbf{X}_t = \mathbf{x}\right] = \mathbf{x} - \mathbb{E}\left[\mathbf{X}_0 | \mathbf{X}_t = \mathbf{x}\right]$, a direct expansion yields

$$\mathbb{E}\left[\left(\frac{1}{\sigma_{t|0}^4}(\mathbf{x} - \overrightarrow{\mathbf{X}}_0)(\mathbf{x} - \overrightarrow{\mathbf{X}}_0)^{\mathrm{T}}\right)\middle|\,\overrightarrow{\mathbf{X}}_t = \mathbf{x}\right] = \operatorname{Var}(\overrightarrow{\mathbf{X}}_0 \mid \overrightarrow{\mathbf{X}}_t = x),$$

which plugged into (38), gives (37). □

Proposition D.4. *Suppose that Assumption 3.1 (II) holds and $\alpha = 0$. Then, there exists a continuous function $[0, T] \ni t \mapsto \mathrm{C}_{g,t} \in \mathbb{R}_+$ such that, for all $\mathbf{x} \in \mathbb{R}^d$,*

$$\|\mathrm{S}_t(\mathbf{x})\| \leq \sqrt{\mathrm{C}_{g,t}}\big(1 + \|\mathbf{x}\|^{p+1}\big) \tag{39}$$

$$\left\|\nabla \mathrm{S}_t(\mathbf{x}) + \mathrm{S}_t(\mathbf{x})\,\mathrm{S}_t(\mathbf{x})^{\mathrm{T}}\right\|_{\mathrm{F}} \leq \mathrm{C}_{g,t}\left(1 + \|\mathbf{x}\|^{2p+2}\right). \tag{40}$$

Proof. First, we prove (39). Taking the norm together with Jensen's inequality to (28), we get

$$\|\mathrm{S}_t(\mathbf{x})\| \leq \mathbb{E}\left[\left\|\mathrm{S}_0\left(\overrightarrow{\mathbf{X}}_0\right)\right\| \middle| \overrightarrow{\mathbf{X}}_t = \mathbf{x}\right] .$$

Using Lemma A.2, it follows

$$\|\mathrm{S}_t(\mathbf{x})\| \leq \mathrm{C}_s\left(1 + \mathbb{E}\left[\left\|\overrightarrow{\mathbf{X}}_0\right\|^{p+1} \middle| \overrightarrow{\mathbf{X}}_t = \mathbf{x}\right]\right) .$$

Applying Proposition 3.2 with $\ell = p+1$ yields

$$\|\mathrm{S}_t(\mathbf{x})\| \leq \Gamma_0^s\left(1 + \lambda_{0|t}^{(p+1)}\|\mathbf{x}\|^{p+1} + \mathrm{K}_{0|t}^{(p+1)}\right) .$$

Therefore, we can find $\mathrm{C}_{g,t}$ such that (39) holds.

Similarly, we focus on (40). Applying the Frobenius norm together with Jensen's inequality to (29), we obtain

$$\left\|\nabla \mathrm{S}_t(\mathbf{x}) + \mathrm{S}_t(\mathbf{x})^{\otimes 2}\right\|_{\mathrm{F}} \leq \mathbb{E}\left[\left\|\nabla \mathrm{S}_0\left(\overrightarrow{\mathbf{X}}_0\right) + \mathrm{S}_0\left(\overrightarrow{\mathbf{X}}_0\right)^{\otimes 2}\right\|_{\mathrm{F}} \middle| \overrightarrow{\mathbf{X}}_t = \mathbf{x}\right] .$$

Thus, from Corollary A.3,

$$\left\|\nabla \mathrm{S}_t(\mathbf{x}) + \mathrm{S}_t(\mathbf{x})^{\otimes 2}\right\|_{\mathrm{F}} \leq \Gamma_0^s\left(1 + \mathbb{E}\left[\left\|\overrightarrow{\mathbf{X}}_0\right\|^{2p+2} \middle| \overrightarrow{\mathbf{X}}_t = \mathbf{x}\right]\right) .$$

Applying Proposition 3.2 with $\ell = 2p+2$, it follows

$$\left\|\nabla \mathrm{S}_t(\mathbf{x}) + \mathrm{S}_t(\mathbf{x})^{\otimes 2}\right\|_{\mathrm{F}} \leq \Gamma_0^s\left(1 + \lambda_{0|t}^{(2p+2)}\|\mathbf{x}\|^{2p+2} + \mathrm{K}_{0|t}^{(2p+2)}\right) .$$

This means that, eventually increasing its value, we can find $\mathrm{C}_{g,t}$ such that (40) holds. The continuity of the function $[0,T] \ni t \mapsto \mathrm{C}_{H,t} \in \mathbb{R}_+$ follows immediately from the definition of the $t \mapsto (\lambda_{0|t}^{(\ell)}, \mathrm{K}_{0|t}^{(\ell)})$ from Proposition 3.2, for $\ell \geq 2$. □

We are now in a position to leverage the previous result to establish the propagation in time of Assumption 3.1 (II). This argument applies to both the VE and VP cases, thereby ensuring that the corresponding regularity and stability properties of the score hold uniformly in time.

Corollary D.5. *Suppose that Assumption 3.1 holds. Then, there exists $[0,T] \ni t \mapsto \mathrm{C}_{H,t} \in \mathbb{R}_+$ such that, for all $\mathbf{x} \in \mathbb{R}^d$, it holds that*

$$\|\nabla \mathrm{S}_t(\mathbf{x})\|_{\mathrm{F}} \leq \mathrm{C}_{H,t}\left(1 + \|\mathbf{x}\|^{2p+2}\right), \qquad \|\mathrm{S}_t(\mathbf{x})\| \leq \mathrm{C}_{H,t}\left(1 + \|\mathbf{x}\|^{2p+3}\right) .$$

Proof. Consider first the VE case. By combining Propositions D.1 and D.4 and using the fact that for all $v, w \in \mathbb{R}^d$, $\left\|vw^{\mathrm{T}}\right\|_{\mathrm{F}} = |\langle v, w\rangle|$, we have that

$$\|\nabla \mathrm{S}_t(\mathbf{x})\|_{\mathrm{F}} \leq \left\|\nabla \mathrm{S}_t(\mathbf{x}) + \mathrm{S}_t(\mathbf{x})^{\otimes 2}\right\|_{\mathrm{F}} + \|\mathrm{S}_t(\mathbf{x})\|^2 \leq 2\,\mathrm{C}_{g,t}\left(1 + \|\mathbf{x}\|^{2p+2}\right) .$$

Applying Lemma C.2, we recover the bound (D.5) in the variance–preserving case as well.

To establish the polynomial growth of the score function $x \mapsto \mathrm{S}_t(x)$, we proceed as in Lemma A.2, using the intermediate value theorem together with the previously derived bound on the Jacobian of the score function. □

E. Localized Doeblin minorization condition for the backward process (Proof of Proposition 3.3)

This section proves a localized Doeblin (minorization) condition for the backward Markov kernel $\mathrm{Q}_{t|s}$ on bounded sets that is informally stated in Proposition 3.3. Concretely, we show that for any radius $r > 0$ there exist a constant $\varepsilon_{s|t}(r) \in (0,1)$ and a probability measure $\upsilon_{s|t}$ such that $\mathrm{Q}_{t|s}(\mathbf{x}, \cdot)$ dominates $\varepsilon_{s|t}(r)\upsilon_{s|t}(\cdot)$ uniformly over $\mathbf{x} \in B(0,r)$. Such result is the second ingredient of Harris theorem Hairer & Mattingly (2008).

Proposition E.1 (Formal statement of Proposition 3.3)**.** *Let $0 \leq s < t \leq T$. Assume that Assumption 3.1 hold. Then, for all $r > 0$, there exist a constant $\varepsilon_{t|s}^{(r)} \in (0,1)$ and a probability measure $\upsilon_{t|s}$ on $\mathbb{R}^d$ such that, for all $\mathbf{x} \in \mathrm{B}_r(0)$, $\mathrm{A} \in \mathcal{B}(\mathbb{R}^d)$,*

$$\mathrm{Q}_{t|s}(\mathbf{x}; \mathrm{A}) \geq \varepsilon_{t|s}^{(r)} \upsilon_{t|s}(\mathrm{A}) ,$$

where

$$\varepsilon_{t|s}^{(r)} = \frac{(\pi\sigma_{T-s|0}^2)^{d/2}}{(\mathrm{m}_{T-s|T-t}\,\mathrm{m}_{T-t|0})^d \max\left\{(2\pi\sigma_{T-s|0}^2)^{d/2}\,\mathrm{m}_{T-s|0}^{-d}\left\|p_{\mathrm{data}}\right\|_\infty , 1\right\}} \exp\left(-\frac{r^2}{\sigma_{T-s|T-t}^2}\right)\mathrm{I}_{t|s} ,$$

with

$$\mathrm{I}_{t|s} := \mathbb{E}\left[\mathcal{N}\left(\mathbf{X}_0; 0, \frac{\sigma_{T-s|T-t}^2 + 2\sigma_{T-t|0}^2\,\mathrm{m}_{T-s|T-t}^2}{2\,\mathrm{m}_{T-t|0}^2\,\mathrm{m}_{T-s|T-t}^2}\mathbf{I}_d\right)\right] .$$

Proof. Note that for all $\mathbf{x}_t \in \mathbb{R}^d$, using the notations of Lemma C.1,

$$\begin{aligned} p_t(\mathbf{x}_t) &= \int p_{\mathrm{data}}(\mathbf{x}_0)\,\mathcal{N}\left(\mathbf{x}_t; \mathrm{m}_{t|0}\,\mathbf{x}_0, \sigma_{t|0}^2\mathbf{I}_d\right)\mathrm{d}\mathbf{x}_0 \\ &= \frac{1}{\mathrm{m}_{t|0}^d}\int p_{\mathrm{data}}\left(\frac{u}{\mathrm{m}_{t|0}}\right)\mathcal{N}\left(u; \mathbf{x}_t, \sigma_{t|0}^2\mathbf{I}_d\right)\mathrm{d}u \\ &\leq \max\left\{\mathrm{m}_{t|0}^{-d}\left\|p_{\mathrm{data}}\right\|_\infty , (2\pi\sigma_{t|0}^2)^{-d/2}\right\} \\ &\leq (2\pi\sigma_{t|0}^2)^{-d/2}\max\left\{(2\pi\sigma_{t|0}^2)^{d/2}\,\mathrm{m}_{t|0}^{-d}\left\|p_{\mathrm{data}}\right\|_\infty , 1\right\} . \end{aligned} \tag{41}$$

Let $r > 0$ and assume that $\mathbf{x}_t \in \mathrm{B}_r(0)$. Using that for all $a, b \in \mathbb{R}$, $(a+b)^2 \leq 2a^2 + 2b^2$, we have

$$\begin{aligned} p_{t|s}(\mathbf{x}_t|\mathbf{x}_s) &\geq (2\pi\sigma_{t|s}^2)^{-d/2}\exp\left(-\frac{\|\mathrm{m}_{t|s}\,\mathbf{x}_s\|^2}{\sigma_{t|s}^2}\right)\exp\left(-\frac{\|\mathbf{x}_t\|^2}{\sigma_{t|s}^2}\right) \\ &\geq (\sqrt{2}\,\mathrm{m}_{t|s})^{-d}\mathcal{N}\left(\mathbf{x}_s; 0, \frac{\sigma_{t|s}^2}{2\,\mathrm{m}_{t|s}^2}\mathbf{I}_d\right)\exp\left(-\frac{r^2}{\sigma_{t|s}^2}\right) . \end{aligned} \tag{42}$$

For all $0 < s < t < T$, $\mathrm{A} \in \mathcal{B}(\mathbb{R}^d)$, and $\mathbf{x}_t \in \mathbb{R}^d$,

$$\begin{aligned} \mathrm{Q}_{T-s|T-t}(\mathbf{x}_t; \mathrm{A}) &= \mathbb{E}\left[\mathbb{E}\left[1_{\mathrm{A}}\left(\overrightarrow{\mathbf{X}}_s\right)\middle|\overrightarrow{\mathbf{X}}_t = \mathbf{x}_t, \overrightarrow{\mathbf{X}}_0\right]\middle|\overrightarrow{\mathbf{X}}_t = \mathbf{x}_t\right] \\ &= \int 1_{\mathrm{A}}(\mathbf{x}_s)\,p_{\mathrm{data}}(\mathbf{x}_0)\,p_{s|t,0}(\mathbf{x}_s|\mathbf{x}_t, \mathbf{x}_0)\,\frac{p_{t|0}(\mathbf{x}_t|\mathbf{x}_0)}{p_t(\mathbf{x}_t)}\mathrm{d}\mathbf{x}_0\mathrm{d}\mathbf{x}_s . \end{aligned}$$

By Markovianity of the forward process, we have, for all $\mathbf{x}_0, \mathbf{x}_s, \mathbf{x}_t \in \mathbb{R}^d$,

$$p_{s|t,0}(\mathbf{x}_s|\mathbf{x}_t, \mathbf{x}_0)\,p_{t|0}(\mathbf{x}_t|\mathbf{x}_0) = p_{s|0}(\mathbf{x}_s|\mathbf{x}_0)\,p_{t|s}(\mathbf{x}_t|\mathbf{x}_s) .$$

Therefore, by (42),

$$\mathrm{Q}_{T-s|T-t}(\mathbf{x}_t; \mathrm{A}) \geq p_t(\mathbf{x}_t)^{-1}(\sqrt{2}\,\mathrm{m}_{t|s})^{-d}\exp\left(-\frac{r^2}{\sigma_{t|s}^2}\right)\int 1_{\mathrm{A}}(\mathbf{x}_s)\,p_{s|0}(\mathbf{x}_s|\mathbf{x}_0)\,\mathcal{N}\left(\mathbf{x}_s; 0, \frac{\sigma_{t|s}^2}{2\,\mathrm{m}_{t|s}^2}\mathbf{I}_d\right)p_{\mathrm{data}}(\mathbf{x}_0)\,\mathrm{d}\mathbf{x}_0\mathrm{d}\mathbf{x}_s .$$

By Gaussian conjugation, for all $\mathbf{x}_0, \mathbf{x}_s \in \mathbb{R}^d$, we have

$$p_{s|0}(\mathbf{x}_s|\mathbf{x}_0)\,\mathcal{N}\left(\mathbf{x}_s; 0, \frac{\sigma_{t|s}^2}{2\,\mathrm{m}_{t|s}^2}\mathbf{I}_d\right) = \mathrm{m}_{s|0}^{-d}\,\mathcal{N}\left(\mathbf{x}_s; \frac{\sigma_{t|s}^2\,\mathrm{m}_{s|0}\,\mathbf{x}_0}{\sigma_{t|0}^2 + \sigma_{s|0}^2\,\mathrm{m}_{t|s}^2}, \frac{\sigma_{t|s}^2\sigma_{s|0}^2}{\sigma_{t|0}^2 + \sigma_{s|0}^2\,\mathrm{m}_{t|s}^2}\mathbf{I}_d\right)\mathcal{N}\left(\mathbf{x}_0; 0, \frac{\sigma_{t|s}^2 + 2\sigma_{s|0}^2\,\mathrm{m}_{t|s}^2}{2\,\mathrm{m}_{s|0}^2\,\mathrm{m}_{t|s}^2}\mathbf{I}_d\right) .$$

Thus, if we define

$$\mathrm{I}_{T-s|T-t} := \mathbb{E}\left[\mathcal{N}\left(\overrightarrow{\mathbf{X}}_0; 0, \frac{\sigma_{t|s}^2 + 2\sigma_{s|0}^2 \mathrm{m}_{t|s}^2}{2\,\mathrm{m}_{s|0}^2 \mathrm{m}_{t|s}^2}\mathbf{I}_d\right)\right],$$

$$\mu_{0,s,t}(\mathrm{A}) := \mathbb{E}\left[1_{\mathrm{A}}\left(\overrightarrow{\mathbf{X}}_0\right)\mathcal{N}\left(\overrightarrow{\mathbf{X}}_0; 0, \frac{\sigma_{t|s}^2 + 2\sigma_{s|0}^2 \mathrm{m}_{t|s}^2}{2\,\mathrm{m}_{s|0}^2 \mathrm{m}_{t|s}^2}\mathbf{I}_d\right)\right] / \mathrm{I}_{T-s|T-t},$$

$$\upsilon_{T-s|T-t}(\mathrm{A}) := \int 1_{\mathrm{A}}(\mathbf{x}_s)\,\mathcal{N}\left(\mathbf{x}_s; \frac{\sigma_{t|s}^2 \mathrm{m}_{s|0}\mathbf{x}_0}{\sigma_{t|0}^2 + \sigma_{s|0}^2 \mathrm{m}_{t|s}^2}, \frac{\sigma_{t|s}^2\sigma_{s|0}^2}{\sigma_{t|0}^2 + \sigma_{s|0}^2 \mathrm{m}_{t|s}^2}\mathbf{I}_d\right)\mu_{0,s,t}(\mathrm{d}\mathbf{x}_0)\mathrm{d}\mathbf{x}_s\,.$$

we can complete the proof by (41) with

$$\varepsilon_{T-s|T-t} = \frac{(\pi\sigma_{t|0}^2)^{d/2}}{(\mathrm{m}_{t|s}\,\mathrm{m}_{s|0})^d \max\left\{(2\pi\sigma_{t|0}^2)^{d/2}\,\mathrm{m}_{t|0}^{-d}\,\|p_{\mathrm{data}}\|_\infty, 1\right\}}\exp\left(-\frac{r^2}{\sigma_{t|s}^2}\right)\mathrm{I}_{T-s|T-t}\,.$$

□

F. Explicit constants in Propositions 3.2-3.4

By Proposition E.1, for all $0 \le s < t \le T$, note that using $\mathrm{I}_{T-s|T-t} \le \|p_{\mathrm{data}}\|_\infty$ yields

$$\varepsilon_{T-s|T-t}^{(r)} \le \frac{(\pi\sigma_{t|0}^2)^{d/2}\,\|p_{\mathrm{data}}\|_\infty}{(\mathrm{m}_{t|s}\,\mathrm{m}_{s|0})^d \max\left\{(2\pi\sigma_{t|0}^2)^{d/2}\,\mathrm{m}_{t|0}^{-d}\,\|p_{\mathrm{data}}\|_\infty, 1\right\}}\exp\left(-\frac{r^2}{\sigma_{t|s}^2}\right).$$

Note that

$$\frac{(\pi\sigma_{t|0}^2)^{d/2}\,\|p_{\mathrm{data}}\|_\infty}{(\mathrm{m}_{t|s}\,\mathrm{m}_{s|0})^d \max\left\{(2\pi\sigma_{t|0}^2)^{d/2}\,\mathrm{m}_{t|0}^{-d}\,\|p_{\mathrm{data}}\|_\infty, 1\right\}} = \min\left\{\frac{(\pi\sigma_{t|0}^2)^{d/2}\,\|p_{\mathrm{data}}\|_\infty}{(\mathrm{m}_{t|s}\,\mathrm{m}_{s|0})^d}, \frac{\mathrm{m}_{t|0}^d}{2^{d/2}(\mathrm{m}_{t|s}\,\mathrm{m}_{s|0})^d}\right\},$$

which leads to

$$\varepsilon_{T-s|T-t}^{(r)} \le \min\left\{\frac{(\pi\sigma_{t|0}^2)^{d/2}\,\|p_{\mathrm{data}}\|_\infty}{(\mathrm{m}_{t|s}\,\mathrm{m}_{s|0})^d}, \frac{\mathrm{m}_{t|0}^d}{2^{d/2}(\mathrm{m}_{t|s}\,\mathrm{m}_{s|0})^d}\right\}\exp\left(-\frac{r^2}{\sigma_{t|s}^2}\right).$$

Variance Exploding case. Consider now the VE case, *i.e.*, $\alpha = 0$. Note that

$$\lambda_{t|s}^{(2)} = \exp\left(-\int_s^t 2\bar{\beta}_v\left(\frac{2\gamma_0}{1 + 2\gamma_0\sigma_{T-v|0}^2}\right)\mathrm{d}v\right) = \exp\left(\int_s^t \frac{\mathrm{d}\sigma_{T-v|0}^2}{\mathrm{d}v}\left(\frac{2\gamma_0}{1 + 2\gamma_0\sigma_{T-v|0}^2}\right)\mathrm{d}v\right)$$
$$= \left(\frac{1 + 2\gamma_0\sigma_{T-t|0}^2}{1 + 2\gamma_0\sigma_{T-s|0}^2}\right)$$

and

$$
\begin{aligned}
\mathrm{K}^{(2)}_{t|s} &= \int_s^t \exp\left(-\int_u^t 2\bar{\beta}_v(2\gamma_{T-v})\mathrm{d}v\right) 2\bar{\beta}_u(2\kappa_{T-u}+d)\mathrm{d}u \\
&= -\int_s^t \left(\frac{1+2\gamma_0\sigma^2_{T-t|0}}{1+2\gamma_0\sigma^2_{T-u|0}}\right)\frac{\mathrm{d}\sigma^2_{T-u|0}}{\mathrm{d}u}\frac{2(\kappa_0+d)}{1+2\gamma_0\sigma^2_{T-u|0}}\mathrm{d}u - d(1+2\gamma_0\sigma^2_{T-t|0})\int_s^t\left(\frac{1}{1+2\gamma_0\sigma^2_{T-u|0}}\right)\frac{\mathrm{d}\sigma^2_{T-u|0}}{\mathrm{d}u}\mathrm{d}u \\
&= -2\left(1+2\gamma_0\sigma^2_{T-t|0}\right)(\kappa_0+d)\int_s^t\left(\frac{1}{1+2\gamma_0\sigma^2_{T-u|0}}\right)^2\frac{\mathrm{d}\sigma^2_{T-u|0}}{\mathrm{d}u}\mathrm{d}u + d\frac{(1+2\gamma_0\sigma^2_{T-t|0})}{2\gamma_0}\log\left(\frac{1+2\gamma_0\sigma^2_{T-s|0}}{1+2\gamma_0\sigma^2_{T-t|0}}\right) \\
&= \left(1+2\gamma_0\sigma^2_{T-t|0}\right)\frac{\kappa_0+d}{\gamma_0}\int_s^t\frac{\mathrm{d}}{\mathrm{d}u}\left(\frac{1}{1+2\gamma_0\sigma^2_{T-u|0}}\right)\mathrm{d}u + d\frac{(1+2\gamma_0\sigma^2_{T-t|0})}{2\gamma_0}\log\left(\frac{1+2\gamma_0\sigma^2_{T-s|0}}{1+2\gamma_0\sigma^2_{T-t|0}}\right) \\
&= \left(1+2\gamma_0\sigma^2_{T-t|0}\right)\frac{\kappa_0+d}{\gamma_0}\left(\frac{1}{1+2\gamma_0\sigma^2_{T-t|0}}-\frac{1}{1+2\gamma_0\sigma^2_{T-s|0}}\right) + d\frac{(1+2\gamma_0\sigma^2_{T-t|0})}{2\gamma_0}\log\left(\frac{1+2\gamma_0\sigma^2_{T-s|0}}{1+2\gamma_0\sigma^2_{T-t|0}}\right) \\
&= 2(\kappa_0+d)\frac{\sigma^2_{T-s|T-t}}{1+2\gamma_0\sigma^2_{T-s|0}} + d\frac{(1+2\gamma_0\sigma^2_{T-t|0})}{2\gamma_0}\log\left(\frac{1+2\gamma_0\sigma^2_{T-s|0}}{1+2\gamma_0\sigma^2_{T-t|0}}\right),
\end{aligned}
$$

where we have used that $\sigma^2_{T-s|T-t} = \sigma^2_{T-s|0} - \sigma^2_{T-t|0}$. Therefore, we have that

$$
\frac{2\mathrm{K}^{(2)}_{t|s}}{1-\lambda^{(2)}_{t|s}} = 2\frac{\kappa_0+d}{\gamma_0} + d(1+2\gamma_0\sigma^2_{T-t|0})\log\left(\frac{1+2\gamma_0\sigma^2_{T-s|0}}{1+2\gamma_0\sigma^2_{T-t|0}}\right)\frac{1+2\gamma_0\sigma^2_{T-s|0}}{2\gamma_0\sigma^2_{T-t|T-s}}. \tag{43}
$$

Variance Preserving case. For the VP case ($\alpha > 0$), note that, since $\sigma^2_{u|0} = \alpha^{-1}\left(1-\mathrm{m}^2_{u|0}\right)$ for any $u \geq 0$, we have

$$
\begin{aligned}
\lambda^{(\ell)}_{t|s} &= \exp\left(-\int_s^t \tilde{\gamma}_{v,\ell}\,\mathrm{d}v\right) = \exp\left(-\int_s^t 2\bar{\beta}_t\big(2\gamma_{T-v}-\alpha\big)\,\mathrm{d}v\right) \\
&= \exp\left(\int_s^t 2\bar{\alpha}\bar{\beta}_v\,\mathrm{m}^2_{T-v|0}\left(\frac{1-\frac{2\gamma_0}{\bar{\alpha}}}{(1-\frac{2\gamma_0}{\bar{\alpha}})\,\mathrm{m}^2_{T-v|0}+\frac{2\gamma_0}{\bar{\alpha}}}\right)\mathrm{d}v\right) \\
&= \exp\left(\int_s^t \frac{\mathrm{d}\,\mathrm{m}^2_{T-v|0}}{\mathrm{d}v}\left(\frac{1-\frac{2\gamma_0}{\bar{\alpha}}}{(1-\frac{2\gamma_0}{\bar{\alpha}})\,\mathrm{m}^2_{T-v|0}+\frac{2\gamma_0}{\bar{\alpha}}}\right)\mathrm{d}v\right) \\
&= \frac{(1-\frac{2\gamma_0}{\bar{\alpha}})\,\mathrm{m}^2_{T-t|0}+\frac{2\gamma_0}{\bar{\alpha}}}{(1-\frac{2\gamma_0}{\bar{\alpha}})\,\mathrm{m}^2_{T-s|0}+\frac{2\gamma_0}{\bar{\alpha}}} = \frac{\mathrm{m}^2_{T-t|0}+2\gamma_0\sigma^2_{T-t|0}}{\mathrm{m}^2_{T-s|0}+2\gamma_0\sigma^2_{T-s|0}}
\end{aligned}
$$

and

$$
\begin{aligned}
\mathrm{K}^{(2)}_{t|s} &= \int_s^t \exp\left(-\int_u^t 2\bar{\beta}_v(2\gamma_{T-v}-\alpha)\mathrm{d}v\right)2\bar{\beta}_u(2\kappa_{T-u}+d)\mathrm{d}u \\
&= \int_s^t 2\bar{\beta}_v\left(\frac{\mathrm{m}^2_{T-t|0}+2\gamma_0\sigma^2_{T-t|0}}{\mathrm{m}^2_{T-v|0}+2\gamma_0\sigma^2_{T-v|0}}\right)\left(\frac{2(\kappa_0+d)\,\mathrm{m}^2_{T-v|0}}{\mathrm{m}^2_{T-v|0}+2\gamma_0\sigma^2_{T-v|0}}+d\right)\mathrm{d}u \\
&= \int_s^t 2\bar{\beta}_v\left(\frac{(1-\frac{2\gamma_0}{\alpha})\,\mathrm{m}^2_{T-t|0}+\frac{2\gamma_0}{\alpha}}{(1-\frac{2\gamma_0}{\alpha})\,\mathrm{m}^2_{T-v|0}+\frac{2\gamma_0}{\alpha}}\right)\left(\frac{2(\kappa_0+d)\,\mathrm{m}^2_{T-v|0}}{(1-\frac{2\gamma_0}{\alpha})\,\mathrm{m}^2_{T-v|0}+\frac{2\gamma_0}{\alpha}}+d\right)\mathrm{d}u \\
&= \int_s^t \frac{1}{\alpha\,\mathrm{m}^2_{T-u|0}}\frac{\mathrm{d}\,\mathrm{m}^2_{T-u|0}}{\mathrm{d}u}\left(\frac{(1-\frac{2\gamma_0}{\alpha})\,\mathrm{m}^2_{T-t|0}+\frac{2\gamma_0}{\alpha}}{(1-\frac{2\gamma_0}{\alpha})\,\mathrm{m}^2_{T-v|0}+\frac{2\gamma_0}{\alpha}}\right)\left(\frac{2(\kappa_0+d)\,\mathrm{m}^2_{T-v|0}}{(1-\frac{2\gamma_0}{\alpha})\,\mathrm{m}^2_{T-v|0}+\frac{2\gamma_0}{\alpha}}+d\right)\mathrm{d}u\,,
\end{aligned}
$$

where we used the definition of $\mathrm{m}_{t|0}^2 = \exp\left(-2\alpha \int_0^t \beta_v \mathrm{d}v\right)$ together with $\sigma_{t|0}^2 = \frac{1}{\alpha}\left(1 - \mathrm{m}_{t|0}^2\right)$, for $t \in [0, T]$, when $\alpha > 0$. Using the change of variable $y(u) = \mathrm{m}_{T-u|0}^2$, we can rewrite the last integral as

$$
\begin{aligned}
\mathrm{K}_{t|s}^{(2)} &= \left(\left(1 - \frac{2\gamma_0}{\alpha}\right)\mathrm{m}_{T-t|0}^2 + \frac{2\gamma_0}{\alpha}\right) \int_{\mathrm{m}_{T-s|0}^2}^{\mathrm{m}_{T-t|0}^2} \left(\frac{2(\kappa_0 + d)/\alpha}{\left(\left(1 - \frac{2\gamma_0}{\alpha}\right) y + \frac{2\gamma_0}{\alpha}\right)^2} + \frac{d/\alpha}{y\left(\left(1 - \frac{2\gamma_0}{\alpha}\right) y + \frac{2\gamma_0}{\alpha}\right)} \right) \mathrm{d}y \\
&= \frac{2(\kappa_0 + d)}{\alpha - 2\gamma_0} \left(1 - \frac{\left(1 - \frac{2\gamma_0}{\alpha}\right)\mathrm{m}_{T-t|0}^2 + \frac{2\gamma_0}{\alpha}}{\left(1 - \frac{2\gamma_0}{\alpha}\right)\mathrm{m}_{T-s|0}^2 + \frac{2\gamma_0}{\alpha}}\right) \\
&\quad + \frac{d}{2\gamma_0}\left(\left(1 - \frac{2\gamma_0}{\alpha}\right)\mathrm{m}_{T-t|0}^2 + \frac{2\gamma_0}{\alpha}\right) \log\left(\frac{\mathrm{m}_{T-t|0}^2\left(\left(1 - \frac{2\gamma_0}{\alpha}\right)\mathrm{m}_{T-s|0}^2 + \frac{2\gamma_0}{\alpha}\right)}{\mathrm{m}_{T-s|0}^2\left(\left(1 - \frac{2\gamma_0}{\alpha}\right)\mathrm{m}_{T-t|0}^2 + \frac{2\gamma_0}{\alpha}\right)} \right) \\
&= 2\alpha(\kappa_0 + d)\left(\frac{\mathrm{m}_{T-s|0}^2 - \mathrm{m}_{T-t|0}^2}{\mathrm{m}_{T-s|0}^2 + 2\gamma_0 \sigma_{T-s|0}^2}\right) + \frac{d}{2\gamma_0}\left(\mathrm{m}_{T-t|0}^2 + 2\gamma_0\sigma_{T-t|0}^2\right) \log\left(\frac{\mathrm{m}_{T-t|0}^2\left(\mathrm{m}_{T-s|0}^2 + 2\gamma_0\sigma_{T-s|0}^2\right)}{\mathrm{m}_{T-s|0}^2\left(\mathrm{m}_{T-t|0}^2 + 2\gamma_0\sigma_{T-t|0}^2\right)}\right) .
\end{aligned}
$$

G. Gaussian framework: explicit backward kernel and contraction

This section provides a closed-form analysis of the backward dynamics in the Gaussian setting. When the data distribution is Gaussian, the forward marginals remain Gaussian and the score admits an explicit linear form. This makes the backward SDE linear with time-dependent coefficients, yielding an explicit representation of both the backward flow and its transition probabilities.

G.1. Closed-form score and backward transition

Lemma G.1. *Consider the forward diffusion (1) and assume $\pi_{\mathrm{data}} = \mathcal{N}(\mu, \Sigma)$, with $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ symmetric positive definite. Then, for every $t \in [0, T]$, using the notations of Lemma C.1, the score function is given by*

$$
\mathrm{S}_t(\mathbf{x}) = -\Sigma_t^{-1}\left(\mathbf{x} - \mathrm{m}_{t|0}\,\mu\right) , \tag{44}
$$

with

$$
\Sigma_t := \mathrm{m}_{t|0}^2\,\Sigma + \sigma_{t|0}^2 \mathbf{I}_d .
$$

Moreover,

$$
\partial_t \Sigma_{T-t} = 2\beta_{T-t}\left(\alpha\,\mathrm{m}_{T-t|0}^2\,\Sigma - \left(1 - \alpha\sigma_{T-t|0}^2\right)\mathbf{I}_d\right) . \tag{45}
$$

Proof. Using Lemma C.1

$$
\overrightarrow{\mathbf{X}}_t \overset{\mathcal{L}}{=} \mathrm{m}_{t|0}\,\overrightarrow{\mathbf{X}}_0 + \sigma_{t|0} G ,
$$

with $\overrightarrow{\mathbf{X}}_0 \sim \mathcal{N}(\mu, \Sigma)$, $\overrightarrow{\mathbf{X}}_t$ is Gaussian with $\mathrm{m}_{t|0}\,\mu$ and covariance is $\mathrm{m}_{t|0}^2\,\Sigma + \sigma_{t|0}^2\mathbf{I}_d$. The score of a Gaussian $\mathcal{N}(m, C)$ is $-C^{-1}(\mathbf{x} - m)$, giving the claimed expression for (44). It remains to compute the time derivative. We have

$$
\partial_t\,\mathrm{m}_{t|0} = -\alpha\beta_t\,\mathrm{m}_{t|0} \quad \Longrightarrow \quad \partial_t\,\mathrm{m}_{t|0}^2 = -2\alpha\beta_t\,\mathrm{m}_{t|0}^2 ,
$$

and

$$
\partial_t \sigma_{t|0}^2 = 2\beta_t - 2\alpha\beta_t\sigma_{t|0}^2 = 2\beta_t(1 - \alpha\sigma_{t|0}^2) .
$$

Differentiating $\Sigma_t = \mathrm{m}_{t|0}^2\,\Sigma + \sigma_{t|0}^2\mathbf{I}_d$ and using $\partial_t\Sigma_{T-t} = -\left.\frac{\mathrm{d}}{\mathrm{d}s}\Sigma_s\right|_{s=T-t}$ gives (45). □

Remark G.2. For any $s, t \in [0, T]$ the matrices Σ_s and Σ_t commute.

Lemma G.3. *Assume that $\pi_{\mathrm{data}} = \mathcal{N}(\mu, \Sigma)$. Then, for all $t \in [0, T]$,*

$$\begin{aligned}\overleftarrow{\mathbf{X}}_t &= \mathrm{e}^{-\alpha \int_0^t \bar{\beta}_s \, \mathrm{d}s} \Sigma_{T-t} \Sigma_T^{-1} \overleftarrow{\mathbf{X}}_0 \\ &\quad + \mathrm{e}^{-\alpha \int_0^t \bar{\beta}_s \mathrm{d}s} \Sigma_{T-t} \int_0^t \mathrm{e}^{\alpha \int_0^s \bar{\beta}_u \mathrm{d}u} \sqrt{2\bar{\beta}_s} \Sigma_{T-s}^{-1} \mathrm{d}\mathrm{B}_s \\ &\quad + 2 \left(\mathrm{e}^{-\alpha \int_0^t \bar{\beta}_s \mathrm{d}s} \Sigma_{T-t} \int_0^t \bar{\beta}_s \mathrm{e}^{\alpha \int_0^s \bar{\beta}_u \mathrm{d}u} \Sigma_{T-s}^{-2} \, \mathrm{m}_{T-s|0} \, \mathrm{d}s \right) \mu \,,\end{aligned}$$

where $\Sigma_t = \mathrm{m}_{t|0}^2 \, \Sigma + \sigma_{t|0}^2 \mathbf{I}_d$ (as in Lemma G.1).

Proof. By Lemma G.1, for $u \in [0, T]$, $\mathrm{S}_u(\mathbf{x}) = -\Sigma_u^{-1}(\mathbf{x} - \mathrm{m}_{u|0} \, \mu)$. Plugging $u = T - t$ into (2) gives

$$\mathrm{d}\overleftarrow{\mathbf{X}}_t = \left(\alpha \bar{\beta}_t \overleftarrow{\mathbf{X}}_t - 2\bar{\beta}_t \Sigma_{T-t}^{-1} \left(\overleftarrow{\mathbf{X}}_t - \mathrm{m}_{T-t|0} \, \mu \right) \right) \mathrm{d}t + \sqrt{2\bar{\beta}_t} \mathrm{d}\mathrm{B}_t \ .$$

Define

$$Z_t := \Sigma_{T-t}^{-1} \mathrm{e}^{\alpha \int_0^t \bar{\beta}_s \mathrm{d}s} \overleftarrow{\mathbf{X}}_t \,.$$

Using Itô's formula and (45) we obtain

$$\mathrm{d}Z_t = 2\bar{\beta}_t \mathrm{e}^{\alpha \int_0^t \bar{\beta}_s \mathrm{d}s} \Sigma_{T-t}^{-2} \, \mathrm{m}_{T-t|0} \, \mu \mathrm{d}t + \mathrm{e}^{\alpha \int_0^t \bar{\beta}_s \mathrm{d}s} \sqrt{2\bar{\beta}_t} \Sigma_{T-t}^{-1} \mathrm{d}\mathrm{B}_t \ .$$

Integrating from 0 to t yields

$$Z_t = \Sigma_T^{-1} \overleftarrow{\mathbf{X}}_0 + 2 \int_0^t \bar{\beta}_s \mathrm{e}^{\alpha \int_0^s \bar{\beta}_u \, \mathrm{d}u} \Sigma_{T-s}^{-2} \, \mathrm{m}_{T-s|0} \, \mu \mathrm{d}s + \int_0^t \mathrm{e}^{\alpha \int_0^s \bar{\beta}_u \mathrm{d}u} \sqrt{2\bar{\beta}_s} \Sigma_{T-s}^{-1} \mathrm{d}\mathrm{B}_s \,,$$

and multiplying by $\mathrm{e}^{-\alpha \int_0^t \bar{\beta}_s \, \mathrm{d}s} \Sigma_{T-t}$ gives the desired expression for $\overleftarrow{\mathbf{X}}_t$. □

Corollary G.4. *Assume that $\pi_{\mathrm{data}} = \mathcal{N}(\mu, \Sigma)$ and let $0 \leq s < t \leq T$. Then, the conditional distribution of $\overleftarrow{\mathbf{X}}_t$ given $\overleftarrow{\mathbf{X}}_s = \mathbf{x}$ is Gaussian with*

$$\begin{aligned}\mathbb{E}\left[\overleftarrow{\mathbf{X}}_t \middle| \overleftarrow{\mathbf{X}}_s = \mathbf{x} \right] &= \mathrm{e}^{-\alpha \int_s^t \bar{\beta}_u \, \mathrm{d}u} \Sigma_{T-t} \Sigma_{T-s}^{-1} x \\ &\quad + 2 \left(\mathrm{e}^{-\alpha \int_s^t \bar{\beta}_u \mathrm{d}u} \Sigma_{T-t} \int_s^t \bar{\beta}_r \mathrm{e}^{\alpha \int_s^r \bar{\beta}_u \mathrm{d}u} \Sigma_{T-r}^{-2} \, \mathrm{m}_{T-r|0} \, \mathrm{d}r \right) \mu \,, \\ \mathbb{V}\left[\overleftarrow{\mathbf{X}}_t \middle| \overleftarrow{\mathbf{X}}_s = \mathbf{x} \right] &= \Sigma_{T-t} - \mathrm{e}^{-2\alpha \int_s^t \bar{\beta}_u \mathrm{d}u} \Sigma_{T-t} \Sigma_{T-s}^{-1} \Sigma_{T-t} \,,\end{aligned}$$

in particular the conditional covariance does not depend on $\mathbf{x}$.

Proof. Starting from Lemma G.3 (applied at times t and s) and using the identity

$$\Sigma_T^{-1} \overleftarrow{\mathbf{X}}_0 + 2 \int_0^s \bar{\beta}_r \mathrm{e}^{\alpha \int_0^r \bar{\beta}_u \, \mathrm{d}u} \Sigma_{T-r}^{-2} \, \mathrm{m}_{T-r|0} \, \mu \mathrm{d}r + \int_0^s \mathrm{e}^{\alpha \int_0^r \bar{\beta}_u \, \mathrm{d}u} \sqrt{2\bar{\beta}_r} \Sigma_{T-r}^{-1} \, \mathrm{d}\mathrm{B}_r = \mathrm{e}^{\alpha \int_0^s \bar{\beta}_u \mathrm{d}u} \Sigma_{T-s}^{-1} \overleftarrow{\mathbf{X}}_s,$$

one obtains a decomposition of $\overleftarrow{\mathbf{X}}_t$ given $\overleftarrow{\mathbf{X}}_s$ into: a linear term in $\overleftarrow{\mathbf{X}}_s$, a stochastic integral over $[s, t]$, and a deterministic μ-shift over $[s, t]$. Taking the expectation and using that the stochastic part has mean 0 yields the conditional mean. The conditional covariance is given by Itô's isometry and the same calculation as in Lemma G.3 (now on $[s, t]$ instead of $[0, t]$). □

G.2. Explicit contraction rates for the Euclidean norm

We exploit this structure of the transitions derived previously to obtain explicit contraction bounds: first in Euclidean operator norm (through $\|A_{s:t}\|$), and, as a consequence, in $\mathcal{W}_2$ distance.

Lemma G.5. *Following Corollary G.4, in the Gaussian case, the conditional mean satisfies for any $s < t$*

$$\mathbb{E}\left[\overleftarrow{\mathbf{X}}_t \middle| \overleftarrow{\mathbf{X}}_s = \mathbf{x}\right] = A_{s:t}\mathbf{x} + b_{s:t}\,, \qquad A_{s:t} := \mathrm{e}^{-\alpha\int_s^t \bar{\beta}_u \mathrm{d}u}\Sigma_{T-t}\Sigma_{T-s}^{-1}\,,$$

for some deterministic offset $b_{s:t}$ (explicit in Corollary G.4). In particular, the dependence on the initial condition over one step is entirely carried by the matrix $A_{s:t}$, which is a strict contraction $\|A_{s:t}\| < 1$ for the Euclidean (operator) norm in the VE case and whenever $\alpha^{-1} \geq \lambda_{\max}(\Sigma)$ in the VP case.

Proof. Recall from Lemma G.1 that

$$\Sigma_t = \mathrm{m}_{t|0}^2\,\Sigma + \sigma_{t|0}^2\mathbf{I}_d.$$

Let $\lambda_1, \ldots, \lambda_d$ be the eigenvalues of Σ. Then,

$$\|A_{s:t}\| = \mathrm{e}^{-\alpha\int_s^t \bar{\beta}_u\,\mathrm{d}u} \max_{1\leq i\leq d} \frac{\mathrm{m}_{T-t|0}^2\,\lambda_i + \sigma_{T-t|0}^2}{\mathrm{m}_{T-s|0}^2\,\lambda_i + \sigma_{T-s|0}^2}\,.$$

- **VE case** ($\alpha = 0$). By definition, $\mathrm{m}_{t|0} \equiv 1$ and $\sigma_{t|0}^2$ is strictly increasing in t. Since $T - t < T - s$, we have $\sigma_{T-t|0}^2 < \sigma_{T-s|0}^2$ and therefore for every i,

  $$\frac{\lambda_i + \sigma_{T-t|0}^2}{\lambda_i + \sigma_{T-s|0}^2} < 1,$$

  which implies $\|A_{s:t}\| < 1$. Hence the one–step conditional mean is a strict contraction in Euclidean norm.

- **VP case** ($\alpha > 0$). The prefactor $\mathrm{e}^{-\alpha\int_{t_k}^{t_{k+1}} \bar{\beta}_u\,\mathrm{d}u} < 1$ is always contractive. Moreover, in the VP case, for any $s \in [0, T]$ $\sigma_{s|0}^2 = \alpha^{-1}\left(1 - \mathrm{m}_{s|0}^2\right)$, hence, for any $1 \leq i \leq d$,

  $$\frac{\mathrm{m}_{T-t|0}^2\,\lambda_i + \sigma_{T-t|0}^2}{\mathrm{m}_{T-s|0}^2\,\lambda_i + \sigma_{T-s|0}^2} = \frac{\alpha^{-1} + \mathrm{m}_{T-t|0}^2\left(\lambda_i - \alpha^{-1}\right)}{\alpha^{-1} + \mathrm{m}_{T-s|0}^2\left(\lambda_i - \alpha^{-1}\right)}\,.$$

  In particular, using that $\mathrm{m}_{t|0}$ is strictly decreasing in t,

  - if $\lambda_i \leq \alpha^{-1}$, then

    $$\frac{\mathrm{m}_{T-t|0}^2\,\lambda_i + \sigma_{T-t|0}^2}{\mathrm{m}_{T-s|0}^2\,\lambda_i + \sigma_{T-s|0}^2} < 1,$$

    so the covariance ratio is contractive along this direction. Moreover, if $\lambda_{\max} \leq \alpha^{-1}$, then for all i, $\lambda_i \leq \alpha^{-1}$.
  - if $\lambda_i \geq \alpha^{-1}$, then the ratio is maybe be larger and and this eigendirection may expand if not compensated by the prefactor term.

  It follows that the one–step conditional mean in the VP case is also a strict contraction in Euclidean norm when $\lambda_{\max} < \alpha^{-1}$.

□

Lemma G.6. *Assume that $\pi_{\mathrm{data}} = \mathcal{N}(\mu, \Sigma)$ and let $0 \leq s < t \leq T$. Then, for any $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$,*

$$\mathcal{W}_2\big(\delta_{\mathbf{x}}\mathrm{Q}_{t|s}, \delta_{\mathbf{x}'}\mathrm{Q}_{t|s}\big) \leq \left\|\mathrm{e}^{-\alpha\int_s^t \bar{\beta}_u \mathrm{d}u}\Sigma_{T-t}\Sigma_{T-s}^{-1}\right\| \|\mathbf{x} - \mathbf{x}'\|\,.$$

Proof. In the Gaussian case, for any $0 \leq s < t \leq T$ the conditional law of $\overleftarrow{\mathbf{X}}_t$ given $\overleftarrow{\mathbf{X}}_s = \mathbf{x}$ is Gaussian with covariance independent of $\mathbf{x}$ (see Corollary G.4) and conditional mean

$$\mathbb{E}\left[\overleftarrow{\mathbf{X}}_t \middle| \overleftarrow{\mathbf{X}}_s = \mathbf{x}\right] = A_{s:t}\mathbf{x} + b_{s:t}\,, \qquad A_{s:t} := \mathrm{e}^{-\alpha \int_s^t \bar{\beta}_u \mathrm{d}u} \Sigma_{T-t}\Sigma_{T-s}^{-1}\,.$$

As a consequence, using the closed-form formula for Gaussian random variables (Bures metric)

$$\begin{aligned}\mathcal{W}_2\big(\delta_{\mathbf{x}}\mathrm{Q}_{t|s}, \delta_{\mathbf{x}'}\mathrm{Q}_{t|s}\big) &= \left\| \mathrm{e}^{-\alpha \int_s^t \bar{\beta}_u \,\mathrm{d}u} \Sigma_{T-t}\Sigma_{T-s}^{-1}(\mathbf{x} - \mathbf{x}') \right\| \\ &\leq \left\| \mathrm{e}^{-\alpha \int_s^t \bar{\beta}_u \,\mathrm{d}u} \Sigma_{T-t}\Sigma_{T-s}^{-1} \right\| \|\mathbf{x} - \mathbf{x}'\|\,.\end{aligned}$$

□

H. Proof of the main stability bound (Theorem 4.3)

A proof sketch is provided in the main text. This appendix completes the proof. We collect the intermediate estimates used in the proof sketch from the main text.

H.1. Initialization error: mixing properties of the forward process

We bound the discrepancy between the terminal forward law p_T and the reference Gaussian measure π_∞ used to initialize the backward chain. In the VE case, the KL divergence to the Gaussian reference is controlled via a convexity argument (Lemma H.1). In the VP case, KL contracts exponentially fast along the Ornstein–Uhlenbeck flow (Lemma H.1). These KL bounds are then converted into weighted total variation bounds through Pinsker/Hellinger estimates and moment controls.

Lemma H.1. *Let p_T denote the law of $\overrightarrow{\mathbf{X}}_T$ and suppose that Assumption 3.1 (I) holds.*

- *If $\alpha = 0$ (Variance Exploding) and $\pi_\infty = \mathcal{N}\left(0, (2\int_0^T \beta_s \,\mathrm{d}s)\mathbf{I}_d\right)$, then*

$$\mathrm{KL}\left(p_T \| \pi_\infty\right) \leq \frac{\mathbb{E}\left[\left\|\overrightarrow{\mathbf{X}}_0\right\|^2\right]}{4\int_0^T \beta_s \mathrm{d}s}\,. \tag{46}$$

- *If $\alpha > 0$ (Variance Preserving) and $\pi_\infty = \mathcal{N}\left(0, \alpha^{-1}\mathbf{I}_d\right)$, then*

$$\mathrm{KL}\left(p_T \| \pi_\infty\right) \leq \mathrm{KL}\left(\pi_{\mathrm{data}} \| \pi_\infty\right) \mathrm{e}^{-2\alpha \int_0^T \beta_s \mathrm{d}s}\,. \tag{47}$$

Proof. In the VE case, recall that, for $\mathbf{x} \in \mathbb{R}^d$,

$$p_T(\mathbf{x}) = \int_{\mathbb{R}^d} p_0(y)(2\pi\sigma_{T|0}^2)^{-d/2} \exp\left\{-\frac{\|\mathbf{x} - y\|^2}{2\sigma_{T|0}^2}\right\} \mathrm{d}y\,.$$

Using convexity of the Kullback-Leibler in its first argument yields

$$\mathrm{KL}\left(p_T \| \pi_\infty\right) \leq \int_{\mathbb{R}^d} p_0(y) \mathrm{KL}\left(\mathcal{N}\left(y, (2\int_0^T \beta_s \,\mathrm{d}s)\mathbf{I}_d\right) \middle\| \mathcal{N}\left(0, (2\int_0^T \beta_s \mathrm{d}s)\mathbf{I}_d\right)\right) \mathrm{d}y\,.$$

Moreover,

$$\mathrm{KL}\left(\mathcal{N}\left(y, (2\int_0^T \beta_s \,\mathrm{d}s)\mathbf{I}_d\right) \middle\| \mathcal{N}\left(0, (2\int_0^T \beta_s \,\mathrm{d}s)\mathbf{I}_d\right)\right) = \frac{\|y\|^2}{4\int_0^T \beta_s \mathrm{d}s}\,,$$

which proves (46). In the VP case, note that it follows from Assumption 3.1 (I) together with Lemma A.1 that

$$\mathrm{KL}\left(\pi_{\mathrm{data}} \| \pi_\infty\right) \leq \log \|p_0\|_\infty + \frac{1}{2\alpha^{-1}} \mathbb{E}\left[\left\|\overrightarrow{\mathbf{X}}_0\right\|^2\right] + \frac{d}{2} \log\left(2\pi\alpha^{-1}\right) < \infty.$$

It follows from Lemma B.1 of Strasman et al. (2025a), after a suitable time change and an adjustment of the invariant measure, that (47) holds. □

Proposition H.2. *Suppose that Assumption 3.1 (I) holds and let $\overrightarrow{\mathbf{X}}_\infty \sim \pi_\infty$. Then, the mixing time for SDE (1) in the weighted total variation distance is upper bounded by*

$$\rho_b(p_T, \pi_\infty) \leq \left(\frac{1}{\sqrt{2}} + b\sqrt{2\left(\mathbb{E}\left[V^2(\overrightarrow{\mathbf{X}}_T)\right] + \mathbb{E}\left[V^2(\overrightarrow{\mathbf{X}}_\infty)\right]\right)}\right) \frac{\left\|\overrightarrow{\mathbf{X}}_0\right\|_{L_2}}{2\sqrt{\int_0^T \beta_s \mathrm{d}s}}\,,$$

in the VE case ($\alpha = 0$), and by

$$\rho_b(p_T, \pi_\infty) \leq \left(\frac{1}{\sqrt{2}} + b\sqrt{2\left(\mathbb{E}\left[V^2(\overrightarrow{\mathbf{X}}_T)\right] + \mathbb{E}\left[V^2(\overrightarrow{\mathbf{X}}_\infty)\right]\right)}\right) \sqrt{\mathrm{KL}\left(\pi_{\mathrm{data}} \| \pi_\infty\right)} \mathrm{e}^{-\alpha \int_0^T \beta_s \mathrm{d}s}\,,$$

in the VP case ($\alpha > 0$).

Proof. Using that the both distributions are absolutely continuous with respect to the Lebesgue measure:

$$\begin{aligned}\rho_b(p_T, \pi_\infty) &= \int (1 + bV(\mathbf{x}))|p_T(\mathbf{x}) - p_\infty(\mathbf{x})|\mathrm{d}\mathbf{x} \\ &= \|p_T - p_\infty\|_{\mathrm{TV}} + b\int V(\mathbf{x})|p_T(\mathbf{x}) - p_\infty(\mathbf{x})|\mathrm{d}\mathbf{x}\,.\end{aligned}$$

The left hand-side is controlled using Pinsker's inequality (19),

$$\|p_T - p_\infty\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2}\mathrm{KL}\left(p_T \| \pi_\infty\right)}$$

It remains to control the right-hand side. Write

$$|p_T - p_\infty| = |\sqrt{p_T} - \sqrt{p_\infty}|\,(\sqrt{p_T} + \sqrt{p_\infty}).$$

By Cauchy–Schwarz,

$$\int V(\mathbf{x})|p_T(\mathbf{x}) - p_\infty(\mathbf{x})|\mathrm{d}\mathbf{x} \leq \left(\int V^2(\mathbf{x})\left(\sqrt{p_T(\mathbf{x})} + \sqrt{p_\infty(\mathbf{x})}\right)^2 \mathrm{d}\mathbf{x}\right)^{1/2} \left(\int \left(\sqrt{p_T(\mathbf{x})} - \sqrt{p_\infty(\mathbf{x})}\right)^2 \mathrm{d}\mathbf{x}\right)^{1/2}.$$

On one hand, using $(\sqrt{a} + \sqrt{b})^2 \leq 2(a+b)$,

$$\begin{aligned}\int V^2(\mathbf{x})(\sqrt{p_T(\mathbf{x})} + \sqrt{p_\infty(\mathbf{x})})^2\mathrm{d}\mathbf{x} &\leq 2\int V^2(\mathbf{x})(p_T(\mathbf{x}) + p_\infty(\mathbf{x}))\mathrm{d}\mathbf{x} \\ &= 2\left(\mathbb{E}\left[V^2(\overrightarrow{\mathbf{X}}_T)\right] + \mathbb{E}\left[V^2(\overrightarrow{\mathbf{X}}_\infty)\right]\right)\,,\end{aligned}$$

with $\overrightarrow{\mathbf{X}}_\infty \sim \pi_\infty$. On the other hand, using inequality (20)

$$\int \left(\sqrt{p_T(\mathbf{x})} - \sqrt{p_\infty(\mathbf{x})}\right)^2 \mathrm{d}x = 2\mathrm{H}(\mu_1, \mu_2)^2 \leq \mathrm{KL}\left(p_T \| \pi_\infty\right)\,.$$

It follows that,

$$\rho_b(p_T, \pi_\infty) \leq \left(\frac{1}{\sqrt{2}} + b\sqrt{2\left(\mathbb{E}\left[V^2(\overrightarrow{\mathbf{X}}_T)\right] + \mathbb{E}\left[V^2(\overrightarrow{\mathbf{X}}_\infty)\right]\right)}\right) \sqrt{\mathrm{KL}\left(p_T \| \pi_\infty\right)}\,,$$

since $\mathbb{E}\left[V^2(\overrightarrow{\mathbf{X}}_T)\right] < \infty$ thanks to Lemma A.1 and recall that $\overrightarrow{\mathbf{X}}_\infty \sim \pi_\infty$ is a Gaussian measure so that $\mathbb{E}\left[V^2(\overrightarrow{\mathbf{X}}_\infty)\right] < \infty$. Finally, conclusion follows from Lemma H.1. □

H.2. One-step discretization and approximation error for the backward kernel

We quantify the one-step discrepancy between the exact backward transition $\mathrm{Q}_{t|s}$ and its learned/discretized counterpart $\mathrm{Q}^{\theta}_{t|s}$. The bound is expressed in a weighted total variation distance and splits into two components: a discretization/freezing error (due to keeping the drift frozen over $[s,t]$) and a score approximation error. The proof proceeds by controlling a local KL divergence via a Girsanov argument and converting it to a weighted bound through Hellinger/Pinsker inequalities and Lyapunov moment estimates. In this section we use the shorthand notations

$$\overline{\lambda}^{(\ell)}_{t|s} := \sup_{u\in[s,t]} \lambda^{(\ell)}_{u|s}, \qquad \overline{\mathrm{K}}^{(\ell)}_{t|s} := \sup_{u\in[s,t]} \mathrm{K}^{(\ell)}_{u|s}, \qquad \overline{\mathrm{C}}_{H,t,s} := \sup_{u\in[s,t]} \mathrm{C}_{H,T-u}\,.$$

for the constants of Proposition 3.2 and Corollary D.5,

Proposition H.3. *Let $0 \le s < t \le T$ such that $s < t$ and $\mathbf{x}\in\mathbb{R}^d$. Suppose that Assumption 3.1 hold. Set $\Delta := \int_s^t \bar{\beta}_u \mathrm{d}u > 0$ and*

$$\mathcal{E}^{\mathbf{x}}_s := \mathrm{S}_{T-s}\left(\mathbf{x}\right) - \mathrm{s}_{\theta}\left(\mathbf{x}, T-s\right)\,.$$

Then,

$$\begin{aligned}\rho_b\Big(\delta_{\mathbf{x}}\mathrm{Q}_{t|s}, \delta_{\mathbf{x}}\mathrm{Q}^{\theta}_{t|s}\Big) &\le \sqrt{\Delta^2\big(\overline{\Gamma}^{(0)}_{s:t} + \overline{\Gamma}^{(1)}_{s:t}\left\|\mathbf{x}\right\|^{4p+4}\big) + 3\Delta\left\|\mathcal{E}^{\mathbf{x}}_s\right\|^2} H_s(\mathbf{x}) \\ &\le \Delta\sqrt{\overline{\Gamma}^{(0)}_{s:t} + \overline{\Gamma}^{(1)}_{s:t}\left\|\mathbf{x}\right\|^{4p+4}} H_{s:t}(\mathbf{x}) + \sqrt{3\Delta}\left\|\mathcal{E}^{\mathbf{x}}_s\right\| H_{s:t}(\mathbf{x})\,,\end{aligned}$$

where

$$\overline{\Gamma}^{(0)}_{s:t} := \tfrac{3}{4}\Delta\alpha^4\left(\overline{\lambda}^{(2)}_{t|s} + \overline{\mathrm{K}}^{(2)}_{t|s}\right) + 3\left(\alpha^2 d + 8\overline{\mathrm{C}}^2_{H,t,s}\left(1 + \overline{\mathrm{K}}^{(4p+4)}_{t|s}\right)\right)\,,$$

$$\overline{\Gamma}^{(1)}_{s:t} := 24\overline{\mathrm{C}}^2_{H,t,s}\overline{\lambda}^{(4p+4)}_{t|s} + \tfrac{3}{4}\Delta\alpha^4\overline{\lambda}^{(2)}_{t|s}\,,$$

and

$$H_{s:t}(\mathbf{x}) := \left(\sqrt{\mathrm{Q}_{t|s}(1+b\,\mathrm{V}_2(\mathbf{x}))^2} + \sqrt{\mathrm{Q}^{\theta}_{t|s}(1+b\,\mathrm{V}_2(\mathbf{x}))^2}\right)\,.$$

Proof. Using Lemma B.1 and inequality (20) relating the Hellinger distance and Kullback–Leibler divergence, we obtain for $\mathbf{x}\in\mathbb{R}^d$,

$$\begin{aligned}\rho_b\Big(\delta_{\mathbf{x}}\mathrm{Q}_{t|s}, \delta_{\mathbf{x}}\mathrm{Q}^{\theta}_{t|s}\Big) &\le \sqrt{2}\mathrm{H}\left(\delta_{\mathbf{x}}\mathrm{Q}_{t|s}, \delta_{\mathbf{x}}\mathrm{Q}^{\theta}_{t|s}\right) \\ &\quad\times\left(\mathbb{E}\left[\left(1+b\,\mathrm{V}_2\Big(\overleftarrow{\mathbf{X}}_t\Big)\right)^2\middle|\overleftarrow{\mathbf{X}}_s=\mathbf{x}\right]^{1/2} + \mathbb{E}\left[\left(1+b\,\mathrm{V}_2(\bar{\mathbf{X}}^{\theta}_t)\right)^2\middle|\bar{\mathbf{X}}^{\theta}_s=\mathbf{x}\right]^{1/2}\right) \\ &\le \sqrt{\mathrm{KL}\left(\delta_{\mathbf{x}}\mathrm{Q}_{t|s}\middle\|\delta_{\mathbf{x}}\mathrm{Q}^{\theta}_{t|s}\right)} \\ &\quad\times\left(\mathbb{E}\left[\left(1+b\,\mathrm{V}_2\Big(\overleftarrow{\mathbf{X}}_t\Big)\right)^2\middle|\overleftarrow{\mathbf{X}}_s=\mathbf{x}\right]^{1/2} + \mathbb{E}\left[\left(1+b\,\mathrm{V}_2(\bar{\mathbf{X}}^{\theta}_t)\right)^2\middle|\bar{\mathbf{X}}^{\theta}_s=\mathbf{x}\right]^{1/2}\right)\,.\end{aligned}$$

Since $(1+b\,\mathrm{V}_2)^2 \le 2(1+b^2\,\mathrm{V}_2^2)$, applying Proposition 3.2 for V_4, we have that $\mathbb{E}\left[\mathrm{V}_2\Big(\overleftarrow{\mathbf{X}}_t\Big)^2\middle|\overleftarrow{\mathbf{X}}_s=\mathbf{x}\right] < \infty$. Moreover, since $\|\mathrm{s}_{\theta}\left(\mathbf{x}, s\right)\|$ is deterministic then $\mathbb{E}\left[\mathrm{V}_2\big(\bar{\mathbf{X}}^{\theta}_t\big)^2\middle|\bar{\mathbf{X}}^{\theta}_s=\mathbf{x}\right] < \infty$.

Step 1: KL bound via L_2 drift bound. Consider the continuous-time interpolation $(\bar{\mathbf{X}}^{\theta}_u)_{u\in[s,t]}$ of the Markov chain (4), defined for $u\in[s,t]$, as

$$\bar{\mathbf{X}}^{\theta}_u = \mathbf{x} + \int_s^u \bar{\beta}_{\ell}\mathrm{d}\ell\left(\alpha\mathbf{x} + 2\,\mathrm{s}_{\theta}\left(\mathbf{x}, T-s\right)\right) + \int_s^u \sqrt{2\bar{\beta}_{\ell}}\mathrm{d}\mathrm{B}_{\ell}\,, \tag{48}$$

Applying Lemma H.5 and Corollary H.6, together with the data processing inequality (Nutz, 2021, Lemma 1.6), the Kullback-Leibler divergence between the one–step kernels can be bounded in terms of the L^2 drift mismatch between (2) and (48):

$$\mathrm{KL}\left(\delta_{\mathbf{x}}\mathrm{Q}_{t|s}\middle\|\delta_{\mathbf{x}}\mathrm{Q}^{\theta}_{t|s}\right) \leq \frac{1}{4}\int_s^t \bar{\beta}_u \mathbb{E}\left[\left\|\alpha\left(\overleftarrow{\mathbf{X}}_u - \mathbf{x}\right) + 2\left(\mathrm{S}_{T-u}\left(\overleftarrow{\mathbf{X}}_u\right) - \mathrm{s}_\theta\left(\mathbf{x}, T-s\right)\right)\right\|^2 \middle| \overleftarrow{\mathbf{X}}_s = \mathbf{x}\right] \mathrm{d}u\,. \tag{49}$$

Step 2: error decomposition. The score mismatch decomposes as the sum of a freezing error and a score approximation error that writes as

$$\begin{aligned}
&\mathrm{S}_{T-u}\left(\overleftarrow{\mathbf{X}}_u\right) - \mathrm{s}_\theta\left(\mathbf{x}, T-s\right) \\
&= \left(\mathrm{S}_{T-u}\left(\overleftarrow{\mathbf{X}}_u\right) - \mathrm{S}_{T-s}\left(\mathbf{x}\right)\right) + \left(\mathrm{S}_{T-s}\left(\mathbf{x}\right) - \mathrm{s}_\theta\left(\mathbf{x}, T-s\right)\right) \\
&= \mathbf{Y}^{\mathbf{x}}_u - \mathbf{Y}^{\mathbf{x}}_s + \mathcal{E}^{\mathbf{x}}_s\,,
\end{aligned}$$

where we used the notation of Lemma H.7 *i.e.*, $\mathbf{Y}^{\mathbf{x}}_u := \mathrm{S}_{T-u}\left(\overleftarrow{\mathbf{X}}_u\right)$ for $u \in [s,t]$. From Lemma H.7, with $\mathbf{Z}^{\mathbf{x}}_u = \nabla \mathrm{S}_{T-u}\left(\overleftarrow{\mathbf{X}}^{\mathbf{x}}_u\right)$ for $u \in [s,t]$, we have that

$$\mathbf{Y}^{\mathbf{x}}_t - \mathbf{Y}^{\mathbf{x}}_s = -\alpha\int_s^t \bar{\beta}_u \mathbf{Y}^{\mathbf{x}}_u \mathrm{d}u + \int_s^t \sqrt{2\bar{\beta}_u}\mathbf{Z}^{\mathbf{x}}_u \mathrm{dB}_u\,.$$

By definition of the backward process we have

$$\overleftarrow{\mathbf{X}}^{\mathbf{x}}_t - \mathbf{x} = \int_s^t \left(\alpha\bar{\beta}_u\overleftarrow{\mathbf{X}}^{\mathbf{x}}_u + 2\bar{\beta}_u\mathbf{Y}^{\mathbf{x}}_u\right)\mathrm{d}u + \int_s^t \sqrt{2\bar{\beta}_u}\mathrm{dB}_u\,.$$

Combining these cancels the $\mathbf{Y}$-integrals and yields the following identity

$$\alpha(\overleftarrow{\mathbf{X}}^{\mathbf{x}}_t - \mathbf{x}) + 2(\mathbf{Y}^{\mathbf{x}}_t - \mathbf{Y}^{\mathbf{x}}_s) = \alpha^2\int_s^t \bar{\beta}_u\,\overleftarrow{\mathbf{X}}^{\mathbf{x}}_u\,\mathrm{d}u + \int_s^t \sqrt{2\bar{\beta}_u}(\alpha\mathbf{I}_d + 2\mathbf{Z}^{\mathbf{x}}_u)\mathrm{dB}_u\,.$$

Step 3: bounding the conditional error terms. Plugging the above decomposition in (49) and using the triangle inequality yields

$$\begin{aligned}
&\mathbb{E}\left[\left\|\alpha\left(\overleftarrow{\mathbf{X}}_t - \mathbf{x}\right) + 2\left(\mathrm{S}_{T-t}\left(\overleftarrow{\mathbf{X}}_t\right) - \mathrm{s}_\theta\left(\mathbf{x}, T-s\right)\right)\right\|^2\middle|\overleftarrow{\mathbf{X}}_s = \mathbf{x}\right] \\
&\quad\leq 3\mathbb{E}\left[\left\|\alpha^2\int_s^t \bar{\beta}_u\overleftarrow{\mathbf{X}}_u\mathrm{d}u\right\|^2\middle|\overleftarrow{\mathbf{X}}_s = \mathbf{x}\right] + 3\mathbb{E}\left[\left\|\int_s^t\sqrt{2\bar{\beta}_u}\left(\alpha\mathbf{I}_d + 2\mathbf{Z}^{\mathbf{x}}_u\right)\mathrm{dB}_u\right\|^2\middle|\overleftarrow{\mathbf{X}}_s = \mathbf{x}\right] + 12\left\|\mathcal{E}^{\mathbf{x}}_s\right\|^2\,.
\end{aligned}$$

Let $\nu(\mathrm{d}u) := (\bar{\beta}_u/\Delta)\mathrm{d}u$ be a probability measure over $[s,t]$. Then, using Jensen's inequality

$$\begin{aligned}
\mathbb{E}\left[\left\|\int_s^t \bar{\beta}_u\overleftarrow{\mathbf{X}}_u\mathrm{d}u\right\|^2\middle|\overleftarrow{\mathbf{X}}_s = \mathbf{x}\right] &= \Delta^2\mathbb{E}\left[\left\|\int_s^t \overleftarrow{\mathbf{X}}_u\nu(\mathrm{d}u)\right\|^2\middle|\overleftarrow{\mathbf{X}}_s = \mathbf{x}\right] \\
&\leq \Delta^2\mathbb{E}\left[\int_s^t\left\|\overleftarrow{\mathbf{X}}_u\right\|^2\nu(\mathrm{d}u)\middle|\overleftarrow{\mathbf{X}}_s = \mathbf{x}\right] \\
&= \Delta\mathbb{E}\left[\int_s^t\bar{\beta}_u\left\|\overleftarrow{\mathbf{X}}_u\right\|^2\mathrm{d}u\middle|\overleftarrow{\mathbf{X}}_s = \mathbf{x}\right]\,.
\end{aligned}$$

By Proposition 3.2,

$$\mathbb{E}\left[\left\|\overleftarrow{\mathbf{X}}_u\right\|^2\middle|\overleftarrow{\mathbf{X}}_s = \mathbf{x}\right] \leq \lambda^{(2)}_{u|s}\|\mathbf{x}\|^2 + \mathrm{K}^{(2)}_{u|s}\,.$$

Hence,

$$
\begin{aligned}
\mathbb{E}\left[\left\|\alpha^2\int_s^t \bar\beta_u \overleftarrow{\mathbf{X}}_u \mathrm{d}u\right\|^2 \middle| \overleftarrow{\mathbf{X}}_s = \mathbf{x}\right] &\leq \alpha^4\Delta\mathbb{E}\left[\int_s^t \bar\beta_u \left\|\overleftarrow{\mathbf{X}}_u\right\|^2 \mathrm{d}u \middle| \overleftarrow{\mathbf{X}}_s = \mathbf{x}\right] \\
&\leq \alpha^4\Delta\left(\|\mathbf{x}\|^2 \int_s^t \bar\beta_u \lambda_{u|s}^{(2)} \mathrm{d}u + \int_s^t \bar\beta_u \mathrm{K}_{u|s}^{(2)} \mathrm{d}u\right) \\
&\leq \alpha^4\Delta^2\left(\overline{\lambda}_{t|s}^{(2)} \|\mathbf{x}\|^2 + \overline{\mathrm{K}}_{t|s}^{(2)}\right) . \qquad (50)
\end{aligned}
$$

By Itô's isometry,

$$
\begin{aligned}
\mathbb{E}\left[\left\|\int_s^t \sqrt{2\bar\beta_u}\,(\alpha\mathbf{I}_d + 2\mathbf{Z}_u^{\mathbf{x}})\,\mathrm{dB}_u\right\|^2 \middle| \overleftarrow{\mathbf{X}}_s = \mathbf{x}\right] &= 2\int_s^t \bar\beta_u \mathbb{E}\left[\|\alpha\mathbf{I}_d + 2\mathbf{Z}_u^{\mathbf{x}}\|_{\mathrm{F}}^2 \middle| \overleftarrow{\mathbf{X}}_s = \mathbf{x}\right] \mathrm{d}u \\
&\leq 4\alpha^2 d\Delta + 16\int_s^t \bar\beta_u \mathbb{E}\left[\|\mathbf{Z}_u^{\mathbf{x}}\|_{\mathrm{F}}^2 \middle| \overleftarrow{\mathbf{X}}_s = \mathbf{x}\right] \mathrm{d}u\,.
\end{aligned}
$$

Since $\mathbf{Z}_u^{\mathbf{x}} := \nabla\,\mathrm{S}_{T-u}\left(\overleftarrow{\mathbf{X}}_u\right)$, Corollary D.5 applied at time $T-u$ and point $\overleftarrow{\mathbf{X}}_u$ yields

$$
\|\mathbf{Z}_u^{\mathbf{x}}\|_{\mathrm{F}} = \left\|\nabla\,\mathrm{S}_{T-u}\left(\overleftarrow{\mathbf{X}}_u\right)\right\|_{\mathrm{F}} \leq \mathrm{C}_{H,T-u}\left(1 + \left\|\overleftarrow{\mathbf{X}}_u\right\|^{2p+2}\right) .
$$

Squaring and using $(1+a)^2 \leq 2(1+a^2)$, we obtain

$$
\mathbb{E}\left[\|\mathbf{Z}_u^{\mathbf{x}}\|_{\mathrm{F}}^2 \middle| \overleftarrow{\mathbf{X}}_s = \mathbf{x}\right] \leq 2\;\mathrm{C}_{H,T-u}^2\left(1 + \mathbb{E}\left[\left\|\overleftarrow{\mathbf{X}}_u\right\|^{4p+4} \middle| \overleftarrow{\mathbf{X}}_s = \mathbf{x}\right]\right) .
$$

Therefore,

$$
\int_s^t \bar\beta_u \mathbb{E}\left[\|\mathbf{Z}_u^{\mathbf{x}}\|_{\mathrm{F}}^2 \middle| \overleftarrow{\mathbf{X}}_s = \mathbf{x}\right] \mathrm{d}u \leq 2\int_s^t \bar\beta_u\,\mathrm{C}_{H,T-u}^2\left(1 + \mathbb{E}\left[\left\|\overleftarrow{\mathbf{X}}_u\right\|^{4p+4} \middle| \overleftarrow{\mathbf{X}}_s = \mathbf{x}\right]\right)\mathrm{d}u\,.
$$

Finally, applying Proposition 3.2 with $\ell = 4p+4$ gives, for all $u \in [s,t]$,

$$
\mathbb{E}\left[\left\|\overleftarrow{\mathbf{X}}_u\right\|^{4p+4} \middle| \overleftarrow{\mathbf{X}}_s = \mathbf{x}\right] \leq \lambda_{u|s}^{(4p+4)} \|\mathbf{x}\|^{4p+4} + \mathrm{K}_{u|s}^{(4p+4)} ,
$$

and thus

$$
\begin{aligned}
\int_s^t \bar\beta_u \mathbb{E}\left[\|\mathbf{Z}_u^{\mathbf{x}}\|_{\mathrm{F}}^2 \middle| \overleftarrow{\mathbf{X}}_s = \mathbf{x}\right] \mathrm{d}u &\leq 2\int_s^t \bar\beta_u\,\mathrm{C}_{H,T-u}^2\left(1 + \lambda_{u|s}^{((4p+4))} \|\mathbf{x}\|^{4p+4} + \mathrm{K}_{u|s}^{((4p+4))}\right)\mathrm{d}u \\
&\leq 2\Delta\overline{\mathrm{C}}_{H,t,s}^2\left(1 + \overline{\lambda}_{t|s}^{((4p+4))} \|\mathbf{x}\|^{4p+4} + \overline{\mathrm{K}}_{t|s}^{((4p+4))}\right) .
\end{aligned}
$$

Hence,

$$
\mathbb{E}\left[\left\|\int_s^t \sqrt{2\bar\beta_u}\,(\alpha\mathbf{I}_d + 2\mathbf{Z}_u)\,\mathrm{dB}_u\right\|_{\mathrm{F}}^2 \middle| \overleftarrow{\mathbf{X}}_s = \mathbf{x}\right] \leq 4\Delta\left(\alpha^2 d + 8\overline{\mathrm{C}}_{H,t,s}^2\left(1 + \overline{\lambda}_{t|s}^{((4p+4))} \|\mathbf{x}\|^{4p+4} + \overline{\mathrm{K}}_{t|s}^{((4p+4))}\right)\right) . \qquad (51)
$$

Combining (50) with (51) yields,

$$
\begin{aligned}
&3\mathbb{E}\left[\left\|\alpha^2\int_s^t \bar\beta_u \overleftarrow{\mathbf{X}}_u \mathrm{d}u\right\|^2 \middle| \overleftarrow{\mathbf{X}}_s = \mathbf{x}\right] + 3\mathbb{E}\left[\left\|\int_s^t \sqrt{2\bar\beta_u}\,(\alpha\mathbf{I}_d + 2\mathbf{Z}_u^{\mathbf{x}})\,\mathrm{dB}_u\right\|^2 \middle| \overleftarrow{\mathbf{X}}_s = \mathbf{x}\right] \\
&\quad\leq 3\alpha^4\Delta^2\left(\overline{\lambda}_{t|s}^{(2)} \|\mathbf{x}\|^2 + \overline{\mathrm{K}}_{t|s}^{(2)}\right) \\
&\qquad + 12\Delta\left(\alpha^2 d + 8\overline{\mathrm{C}}_{H,t,s}^2\left(1 + \overline{\lambda}_{t|s}^{(4p+4)} \|\mathbf{x}\|^{4p+4} + \overline{\mathrm{K}}_{t|s}^{(4p+4)}\right)\right) .
\end{aligned}
$$

Hence, using that $\|\mathbf{x}\|^2 \leq 1 + \|\mathbf{x}\|^{4p+4}$,

$$3\mathbb{E}\left[\left\|\alpha^2 \int_s^t \bar{\beta}_u \overleftarrow{\mathbf{X}}_u \mathrm{d}u\right\|^2 \middle| \overleftarrow{\mathbf{X}}_s = \mathbf{x}\right] + 3\mathbb{E}\left[\left\|\int_s^t \sqrt{2\bar{\beta}_u}\,(\alpha \mathbf{I}_d + 2\mathbf{Z}_u^{\mathbf{x}})\,\mathrm{dB}_u\right\|^2 \middle| \overleftarrow{\mathbf{X}}_s = \mathbf{x}\right]$$
$$\leq \Delta\left(\mathrm{D}_{s:t}^{(0)} + \mathrm{D}_{s:t}^{(1)} \|\mathbf{x}\|^{4p+4}\right),$$

with

$$\mathrm{D}_{s:t}^{(0)} := 3\Delta\alpha^4\left(\overline{\lambda}_{t|s}^{(2)} + \overline{\mathrm{K}}_{t|s}^{(2)}\right) + 12\left(\alpha^2 d + 8\overline{\mathrm{C}}_{H,t,s}^2\left(1 + \overline{\mathrm{K}}_{t|s}^{(4p+4)}\right)\right),$$

and

$$\mathrm{D}_{s:t}^{(1)} := 96\overline{\mathrm{C}}_{H,t,s}^2\,\overline{\lambda}_{t|s}^{(4p+4)} + 3\Delta\alpha^4\overline{\lambda}_{t|s}^{(2)}.$$

Step 4: combining all the bounds. It follows that the one-step discretization error is upper bounded by,

$$\begin{aligned}
\mathrm{KL}\left(\delta_{\mathbf{x}}\mathrm{Q}_{t|s}\middle\|\delta_{\mathbf{x}}\mathrm{Q}_{t|s}^{\theta}\right) &\leq \frac{1}{4}\int_s^t \bar{\beta}_u \mathbb{E}\left[\left\|\alpha\left(\overleftarrow{\mathbf{X}}_u - \mathbf{x}\right) + 2\left(\mathrm{S}_{T-u}\left(\overleftarrow{\mathbf{X}}_u\right) - \mathrm{s}_\theta(\mathbf{x}, T-s)\right)\right\|^2 \middle| \overleftarrow{\mathbf{X}}_s = \mathbf{x}\right]\mathrm{d}u \\
&\leq \frac{1}{4}\int_s^t \bar{\beta}_u\left(\Delta(\mathrm{D}_{s:t}^{(0)} + \mathrm{D}_{s:t}^{(1)}\|\mathbf{x}\|^{4p+4}) + 12\|\mathcal{E}_s^{\mathbf{x}}\|^2\right)\mathrm{d}u \\
&= \frac{\Delta^2}{4}(\mathrm{D}_{s:t}^{(0)} + \mathrm{D}_{s:t}^{(1)}\|\mathbf{x}\|^{4p+4}) + 3\Delta\|\mathcal{E}_s^{\mathbf{x}}\|^2,
\end{aligned}$$

absorbing $\frac{1}{4}$ in $\mathrm{D}_{s:t}^{(0)}$ and $\mathrm{D}_{s:t}^{(1)}$ yields the exact constants in the statement of the Lemma. □

Corollary H.4. *Let $k \in \{1, \ldots, N\}$ and set $\Delta_k := \int_{t_{k-1}}^{t_k} \bar{\beta}_u\,\mathrm{d}u$. Assume the hypotheses of Proposition H.3 hold, and recall that $\hat{\pi}_k^{\theta} := \pi_\infty \mathbb{Q}_{0:k}^{\theta}$. Then*

$$\rho_b\left(\hat{\pi}_{k-1}^{\theta}\mathrm{Q}_{t_k|t_{k-1}}, \hat{\pi}_{k-1}^{\theta}\mathrm{Q}_{t_k|t_{k-1}}^{\theta}\right) \leq \Delta_k C_{k-1}^{\mathrm{discr}} + \sqrt{\Delta_k}C_{k-1}^{\mathrm{net}}\left\|\mathcal{E}_{k-1}\right\|_{L_2(\hat{\pi}_{k-1}^{\theta})},$$

with

$$\left\|\mathcal{E}_{k-1}\right\|_{L_2(\hat{\pi}_{k-1}^{\theta})} = \mathbb{E}\left[\left\|\mathrm{S}_{T-t_{k-1}}\left(\bar{\mathbf{X}}_{t_{k-1}}^{\theta}\right) - \mathrm{s}_\theta\left(\bar{\mathbf{X}}_{t_{k-1}}^{\theta}, T - t_{k-1}\right)\right\|^2\right]^{1/2},$$

$$C_{k-1}^{\mathrm{discr}} := \sqrt{\overline{\Gamma}_{t_{k-1}:t_k}^{(0)} + \overline{\Gamma}_{t_{k-1}:t_k}^{(1)}\mathbb{E}\left[\left\|\bar{\mathbf{X}}_{t_{k-1}}^{\theta}\right\|^{4p+4}\right]}\sqrt{\mathbb{E}\left[H_k(\bar{\mathbf{X}}_{t_{k-1}}^{\theta})^2\right]}$$

and

$$C_{k-1}^{\mathrm{net}} = \sqrt{3\mathbb{E}\left[H_k(\bar{\mathbf{X}}_{t_{k-1}}^{\theta})^2\right]}$$

and where H_k (with $s = t_{k-1}$ and $t = t_k$), $\overline{\Gamma}_{t_{k-1}:t_k}^{(0)}$, and $\overline{\Gamma}_{t_{k-1}:t_k}^{(1)}$ are defined in Proposition H.3.

Proof. Using Lemma B.2 we have that

$$\rho_b\left(\hat{\pi}_{k-1}^{\theta}\mathrm{Q}_{t_k|t_{k-1}}, \hat{\pi}_{k-1}^{\theta}\mathrm{Q}_{t_k|t_{k-1}}^{\theta}\right) \leq \int \rho_b\left(\delta_{\mathbf{x}}\mathrm{Q}_{t_k|t_{k-1}}, \delta_{\mathbf{x}}\mathrm{Q}_{t_k|t_{k-1}}^{\theta}\right)\hat{\pi}_{k-1}^{\theta}(\mathrm{d}\mathbf{x}).$$

Let $\bar{\mathbf{X}}_{t_{k-1}}^{\theta} \sim \hat{\pi}_{k-1}^{\theta}$. Then, by Cauchy–Schwarz

$$\begin{aligned}
\int \left\|\mathcal{E}_{t_{k-1}}^{\mathbf{x}}\right\| H_k(\mathbf{x})\hat{\pi}_{k-1}^{\theta}(\mathrm{d}\mathbf{x}) &= \mathbb{E}\left[\left\|\mathcal{E}_{t_{k-1}}^{\bar{\mathbf{X}}_{t_{k-1}}^{\theta}}\right\| H_k(\bar{\mathbf{X}}_{t_{k-1}}^{\theta})\right] \\
&\leq \sqrt{\mathbb{E}\left[\left\|\mathcal{E}_{t_{k-1}}^{\bar{\mathbf{X}}_{t_{k-1}}^{\theta}}\right\|^2\right]}\sqrt{\mathbb{E}\left[H_k(\bar{\mathbf{X}}_{t_{k-1}}^{\theta})^2\right]},
\end{aligned}$$

and similarly,

$$\int \sqrt{\overline{\Gamma}^{(0)}_{s:t} + \overline{\Gamma}^{(1)}_{s:t} \|\mathbf{x}\|^{4p+4}} H_k(\mathbf{x}) \hat{\pi}^\theta_{k-1}(\mathrm{d}\mathbf{x}) \le \sqrt{\overline{\Gamma}^{(0)}_{s:t} + \overline{\Gamma}^{(1)}_{s:t} \mathbb{E}\left[\left\|\bar{\mathbf{X}}^\theta_{t_{k-1}}\right\|^{4p+4}\right]} \sqrt{\mathbb{E}\left[H_k(\bar{\mathbf{X}}^\theta_{t_{k-1}})^2\right]}\,.$$

Multiplying by Δ_k and $\sqrt{3\Delta_k}$ yields the claimed bound. It remains to check the finiteness of the expectations on the right-hand side. In particular, let $G(\mathbf{x}) := (1 + b\,\mathrm{V}_2(\mathbf{x}))^2$ and

$$H_k(\mathbf{x}) := \sqrt{\mathrm{Q}_{t_k|t_{k-1}}\, G(\mathbf{x})} + \sqrt{\mathrm{Q}^\theta_{t_k|t_{k-1}}\, G(\mathbf{x})}\,.$$

Then, using $(a+b)^2 \le 2(a^2+b^2)$,

$$\begin{aligned}\mathbb{E}\left[H_k\big(\bar{\mathbf{X}}^\theta_{t_{k-1}}\big)^2\right] &\le 2(\hat{\pi}^\theta_{k-1}\mathrm{Q}_{t_k|t_{k-1}})[G] + 2(\hat{\pi}^\theta_{k-1}\mathrm{Q}^\theta_{t_k|t_{k-1}})[G] \\ &\le 2(\pi_\infty \mathbb{Q}^\theta_{0:k-1}\mathrm{Q}_{t_k|t_{k-1}})\,[G] + 2\pi_\infty \mathbb{Q}^\theta_{1:k}\,[G]\;.\end{aligned}$$

In particular, since $G(\mathbf{x}) \lesssim 1 + \mathrm{V}_4(\mathbf{x})$, Assumption 4.1 yields $\pi_\infty \mathbb{Q}^\theta_{1:k}\,[G] < \infty$. Moreover, applying Proposition 3.2 with $\ell = 4$ and using again Assumption 4.1,

$$(\pi_\infty \mathbb{Q}^\theta_{1:k-1}\mathrm{Q}_{t_k|t_{k-1}})\,[\mathrm{V}_4(\mathbf{x})] \le \lambda^{(4)}_{t_k|t_{k-1}} \pi_\infty \mathbb{Q}^\theta_{0:k-1}\,[\mathrm{V}_4(\mathbf{x})] + \mathrm{K}^{(4)}_{t_k|t_{k-1}} < \infty\,,$$

and therefore

$$\mathbb{E}\left[H_k\big(\bar{\mathbf{X}}^\theta_{t_{k-1}}\big)^2\right] < \infty\,.$$

□

H.3. Technical lemmas for the main proof

First, this section provides the pathwise change-of-measure identity used to control the KL divergence between the exact backward SDE over one step and the continuous-time interpolation of the (discrete) SGM dynamics (Lemma H.5 and Corollary H.6). Second, under the uniform Hessian control from Corollary D.5 we write the SDE representation of the score function that is used in Proposition H.3.

One-step Girsanov type bound

Lemma H.5. *Fix a discretization grid $0 = t_1 \le \cdots \le t_N = T$, an index $k \in \{1, \cdots, N-1\}$ and a point $\mathbf{x} \in \mathbb{R}^d$. Let $(\overleftarrow{\mathbf{X}}_t)_{t\in[t_k,t_{k+1}]}$ be solution to the backward SDE*

$$\mathrm{d}\overleftarrow{\mathbf{X}}_t = \bar{\beta}_t b(t, \overleftarrow{\mathbf{X}}_t)\mathrm{d}t + \sqrt{2\bar{\beta}_t}\mathrm{dB}_t\,, \qquad \overleftarrow{\mathbf{X}}_{t_k} = \mathbf{x}\,, \tag{52}$$

with $b(t,y) := \alpha y + 2\,\mathrm{S}_{T-t}\,(y)$. Denote by $\mathbb{P}^{\mathbf{x}}_{t_k}$ the distribution of (52) on $(\overleftarrow{\mathbf{X}}_t)_{t\in[t_k,t_{k+1}]}$. Similarly, let $(\bar{\mathbf{X}}^\theta_t)_{t\in[t_k,t_{k+1}]}$ be the continuous-time interpolation of the generative model defined as solution to

$$\mathrm{d}\bar{\mathbf{X}}^\theta_t = \bar{\beta}_t b_\theta(t_k, \mathbf{x})\mathrm{d}t + \sqrt{2\bar{\beta}_t}\mathrm{dB}_t\,, \qquad \bar{\mathbf{X}}^\theta_{t_k} = \mathbf{x}\,, \tag{53}$$

with $b_\theta(t_k,\mathbf{x}) := \alpha\mathbf{x} + 2\,\mathrm{s}_\theta\,(\mathbf{x}, T - t_k)$. Denote by $\mathbb{P}^{\mathbf{x},\theta}_{t_k}$ the distribution of (53) on $(\bar{\mathbf{X}}^\theta_t)_{[t_k,t_{k+1}]}$. Suppose that Assumption 3.1 (I) and Assumption 3.1 (II) hold. Then, it follows that

$$\begin{aligned}&\mathbb{E}\left[\log\left(\frac{\mathrm{d}\mathbb{P}^{\mathbf{x}}_{t_k}}{\mathrm{d}\mathbb{P}^{\mathbf{x},\theta}_{t_k}}\left((\overleftarrow{\mathbf{X}}_t)_{t\in[t_k,t_{k+1}]}\right)\right)\right] \\ &\qquad = \mathbb{E}\left[\int_{t_k}^{t_{k+1}} \sqrt{\frac{\bar{\beta}_t}{2}}\big(b(t,\overleftarrow{\mathbf{X}}_t) - b_\theta(t_k,\mathbf{x})\big)^\top \mathrm{d}B_t + \frac{1}{4}\int_{t_k}^{t_{k+1}} \bar{\beta}_t \left\|b(t,\overleftarrow{\mathbf{X}}_t) - b_\theta(t_k,\mathbf{x})\right\|^2 \mathrm{d}t\right],\end{aligned}$$

where $(B_t)_{t\in[t_k,t_{k+1}]}$ is a Brownian motion under $\mathbb{P}^{\mathbf{x}}_{t_k}$.

Proof. Let $\Omega := C([t_k, t_{k+1}], \mathbb{R}^d)$ and $\mathbf{X} = (\mathbf{X}_t)_{t\in[t_k,t_{k+1}]}$ be the canonical process on Ω. Introduce the following reference process

$$\mathrm{d}\mathbf{Z}_t = \sqrt{2\bar{\beta}_t}\mathrm{dB}_t\,, \qquad \mathbf{Z}_{t_k} = \mathbf{x}\,, \tag{54}$$

and denote by $\mathbb{W}^{\mathbf{x}}_{t_k}$ the distribution of (54) on Ω.

Step 1: Time-changed processes.

Our goal is to apply Theorem 7.7 of Liptser & Shiryaev (2001), which is stated for diffusion processes with unit diffusion coefficient. We therefore introduce the deterministic time-change

$$\tau(t) := \int_{t_k}^{t} 2\bar{\beta}_s \mathrm{d}s\,,\ \text{for } t \in [t_k, t_{k+1}]\,,$$

and denote by $\tau^{-1} : [0, \tau(t_{k+1})] \to [t_k, t_{k+1}]$ its inverse. By the inverse function theorem,

$$(\tau^{-1})'(u) = \frac{1}{2\bar{\beta}_{\tau^{-1}(u)}}\,.$$

Define the time-changed path space $\widehat{\Omega} := C([0, \tau(t_{k+1})], \mathbb{R}^d)$ and introduce the bijective time-change map $\widehat{T} : \Omega \to \widehat{\Omega}$ by

$$(\widehat{T}\omega)(u) := \omega(\tau^{-1}(u)), \qquad u \in [0, \tau(t_{k+1})]\,.$$

Its inverse $\widehat{T}^{-1} : \widehat{\Omega} \to \Omega$ is given by $(\widehat{T}^{-1}\hat{\omega})(t) = \hat{\omega}(\tau(t))$ for $t \in [t_k, t_{k+1}]$. For $\widehat{\mathbf{Z}}_u := \mathbf{Z}_{\tau^{-1}(u)}$, we get that $(\widehat{\mathbf{Z}}_u)_{u\in[0,\tau(t_{k+1})]}$ is a standard d-dimensional Brownian motion starting from $\mathbf{x}$, and we denote by $\widehat{\mathbb{W}}^{\mathbf{x}}_{t_k} := \mathbb{W}^{\mathbf{x}}_{t_k} \circ \widehat{T}^{-1}$ its law on $\widehat{\Omega}$. Similarly, define the time-changed processes

$$\widehat{\overleftarrow{\mathbf{X}}}_u := \overleftarrow{\mathbf{X}}_{\tau^{-1}(u)} \qquad (\text{resp. } \hat{\bar{\mathbf{X}}}^{\theta}_u := \bar{\mathbf{X}}^{\theta}_{\tau^{-1}(u)})\,,$$

satisfying

$$\mathrm{d}\widehat{\overleftarrow{\mathbf{X}}}_u = \tfrac{1}{2}b(\tau^{-1}(u), \widehat{\overleftarrow{\mathbf{X}}}_u)\,\mathrm{d}u + \mathrm{d}\widehat{\mathbf{Z}}_u\,, \qquad \mathrm{d}\hat{\bar{\mathbf{X}}}^{\theta}_u = \tfrac{1}{2}b_\theta(t_k, \mathbf{x})\,\mathrm{d}u + \mathrm{d}\widehat{\mathbf{Z}}_u\,.$$

We denote by $\widehat{\mathbb{P}}^{\mathbf{x}}_{t_k} = \mathbb{P}^{\mathbf{x}}_{t_k} \circ \widehat{T}^{-1}$ and $\widehat{\mathbb{P}}^{\mathbf{x},\theta}_{t_k} = \mathbb{P}^{\mathbf{x},\theta}_{t_k} \circ \widehat{T}^{-1}$ their respective laws on $\widehat{\Omega}$.

Step 2: density processes.

Let $\widehat{\mathbf{X}} = (\widehat{\mathbf{X}}_u)_{u\in[0,\tau(t_{k+1})]}$ be the canonical process on $\widehat{\Omega}$. Under $\widehat{\mathbb{W}}^{\mathbf{x}}_{t_k}$, $\widehat{\mathbf{X}}$ is a standard d-dimensional Brownian motion with $\widehat{\mathbf{X}}_0 = \mathbf{x}$. Moreover, under $\widehat{\mathbb{P}}^{\mathbf{x}}_{t_k}$ (resp. $\widehat{\mathbb{P}}^{\mathbf{x},\theta}_{t_k}$) it has the same distribution as $(\widehat{\overleftarrow{\mathbf{X}}}_u)_{u\in[0,\tau(t_{k+1})]}$ (resp. $(\hat{\bar{\mathbf{X}}}^{\theta}_u)_{u\in[0,\tau(t_{k+1})]}$). It follows from Proposition D.4 and Proposition 3.2 that

$$\mathbb{E}\left[\int_0^{\tau(t_{k+1})} \frac{1}{4}\left\|b(\tau^{-1}(u), \widehat{\overleftarrow{\mathbf{X}}}_u)\right\|^2 \mathrm{d}u\right] = \mathbb{E}\left[\int_0^{\tau(t_{k+1})} \frac{1}{4}\left\|\alpha\widehat{\overleftarrow{\mathbf{X}}}_u + 2\,\mathrm{S}_{T-\tau^{-1}(u)}\left(\widehat{\overleftarrow{\mathbf{X}}}_u\right)\right\|^2 \mathrm{d}u\right] < \infty$$

From Proposition D.4, we also get that

$$\mathbb{E}\left[\int_0^{\tau(t_{k+1})} \frac{1}{4}\left\|b(\tau^{-1}(u), \widehat{\mathbf{Z}}_u)\right\|^2 \mathrm{d}u\right] < \infty\,.$$

Since the integrands are nonnegative, these imply the corresponding a.s. finiteness conditions required in Theorem 7.7 in Liptser & Shiryaev (2001). Then, by Theorem 7.7 in Liptser & Shiryaev (2001), the path measures are equivalent $\widehat{\mathbb{P}}^{\mathbf{x}}_{t_k} \sim \widehat{\mathbb{W}}^{\mathbf{x}}_{t_k}$ and

$$\frac{\mathrm{d}\widehat{\mathbb{P}}^{\mathbf{x}}_{t_k}}{\mathrm{d}\widehat{\mathbb{W}}^{\mathbf{x}}_{t_k}}(\widehat{\mathbf{X}}) = \exp\left\{\int_0^{\tau(t_{k+1})} \tfrac{1}{2}\, b(\tau^{-1}(u), \widehat{\mathbf{X}}_u)^\top \mathrm{d}\widehat{\mathbf{X}}_u - \frac{1}{8}\int_0^{\tau(t_{k+1})} \left\|b(\tau^{-1}(u), \widehat{\mathbf{X}}_u)\right\|^2 \mathrm{d}u\right\} \quad \widehat{\mathbb{W}}^{\mathbf{x}}_{t_k} \text{ a.s.}$$

Moreover, since $b_\theta(t_k,\mathbf{x})$ is deterministic, the same theorem yields $\widehat{\mathbb{P}}^{\mathbf{x},\theta}_{t_k} \sim \widehat{\mathbb{W}}^{\mathbf{x}}_{t_k}$ and

$$\frac{\mathrm{d}\widehat{\mathbb{P}}^{\mathbf{x},\theta}_{t_k}}{\mathrm{d}\widehat{\mathbb{W}}^{\mathbf{x}}_{t_k}}(\widehat{\mathbf{X}}) = \exp\left\{\int_0^{\tau(t_{k+1})} \tfrac{1}{2}\, b_\theta(t_k,\mathbf{x})^\top \mathrm{d}\widehat{\mathbf{X}}_u - \frac{1}{8}\int_0^{\tau(t_{k+1})} \left\|b_\theta(t_k,\mathbf{x})\right\|^2 \mathrm{d}u\right\} \quad \widehat{\mathbb{W}}^{\mathbf{x}}_{t_k} \text{ a.s.}$$

Therefore, the ratio of the above densities yields $\widehat{\mathbb{P}}^{\mathbf{x},\theta}_{t_k}$ almost surely,

$$\frac{\mathrm{d}\widehat{\mathbb{P}}^{\mathbf{x}}_{t_k}}{\mathrm{d}\widehat{\mathbb{P}}^{\mathbf{x},\theta}_{t_k}}(\widehat{\mathbf{X}}) = \exp\left\{\int_0^{\tau(t_{k+1})} \tfrac{1}{2}\Big(b(\tau^{-1}(u),\widehat{\mathbf{X}}_u) - b_\theta(t_k,\mathbf{x})\Big)^\top \mathrm{d}\widehat{\mathbf{X}}_u \right.$$
$$\left. - \frac{1}{8}\int_0^{\tau(t_{k+1})}\left(\left\|b(\tau^{-1}(u),\widehat{\mathbf{X}}_u)\right\|^2 - \|b_\theta(t_k,\mathbf{x})\|^2\right)\mathrm{d}u\right\}. \tag{55}$$

By equivalence of the measures, equation (55) also holds $\widehat{\mathbb{P}}^{\mathbf{x}}_{t_k}$ almost surely.

Under $\widehat{\mathbb{P}}^{\mathbf{x}}_{t_k}$ the process

$$\widehat{\mathrm{B}}_u := \widehat{\mathbf{X}}_u - \mathbf{x} - \int_0^u \tfrac{1}{2} b(\tau^{-1}(s),\widehat{\mathbf{X}}_s)\mathrm{d}s, \qquad \text{for } u \in [0,\tau(t_{k+1})]\,,$$

is a d-dimensional standard Brownian motion. Consequently, under $\widehat{\mathbb{P}}^{\mathbf{x}}_{t_k}$, $\mathrm{d}\widehat{\mathbf{X}}_u = \frac{1}{2}b(\tau^{-1}(u),\widehat{\mathbf{X}}_u)\mathrm{d}u + \mathrm{d}\widehat{\mathrm{B}}_u$. As a consequence, $\widehat{\mathbb{P}}^{\mathbf{x}}_{t_k}$ almost surely,

$$\frac{\mathrm{d}\widehat{\mathbb{P}}^{\mathbf{x}}_{t_k}}{\mathrm{d}\widehat{\mathbb{P}}^{\mathbf{x},\theta}_{t_k}}(\widehat{\mathbf{X}}) = \exp\left\{\int_0^{\tau(t_{k+1})} \tfrac{1}{2}\Big(b(\tau^{-1}(u),\widehat{\mathbf{X}}_u) - b_\theta(t_k,\mathbf{x})\Big)^\top \mathrm{d}\widehat{\mathrm{B}}_u \right.$$
$$\left. + \frac{1}{8}\int_0^{\tau(t_{k+1})}\left(\left\|b(\tau^{-1}(u),\widehat{\mathbf{X}}_u) - b_\theta(t_k,\mathbf{x})\right\|^2\right)\mathrm{d}u\right\}.$$

Step 3: density process for the original-time laws.

Let $\widehat{T} : \Omega \to \widehat{\Omega}$ be the deterministic time-change map defined in Step 1. By construction of the time-changed processes, recall that

$$\widehat{\mathbb{P}}^{\mathbf{x}}_{t_k} = \mathbb{P}^{\mathbf{x}}_{t_k} \circ \widehat{T}^{-1}, \qquad \widehat{\mathbb{P}}^{\mathbf{x},\theta}_{t_k} = \mathbb{P}^{\mathbf{x},\theta}_{t_k} \circ \widehat{T}^{-1}\,.$$

In particular, we have equivalence of the measures, *i.e.*, $\mathbb{P}^{\mathbf{x}}_{t_k} \sim \mathbb{P}^{\mathbf{x},\theta}_{t_k}$ and for any test function $\psi : \Omega \to \mathbb{R}$,

$$\int_\Omega \psi(\mathbf{X})\mathrm{d}\mathbb{P}^{\mathbf{x}}_{t_k} = \int_{\widehat{\Omega}} \psi(\widehat{T}^{-1}\widehat{\mathbf{X}})\widehat{\mathbb{P}}^{\mathbf{x}}_{t_k} = \int_{\widehat{\Omega}} \psi(\widehat{T}^{-1}\widehat{\mathbf{X}})\frac{\mathrm{d}\widehat{\mathbb{P}}^{\mathbf{x}}_{t_k}}{\mathrm{d}\widehat{\mathbb{P}}^{\mathbf{x},\theta}_{t_k}}(\widehat{\mathbf{X}})\mathrm{d}\widehat{\mathbb{P}}^{\mathbf{x},\theta}_{t_k} = \int_\Omega \psi(\mathbf{X})\frac{\mathrm{d}\widehat{\mathbb{P}}^{\mathbf{x}}_{t_k}}{\mathrm{d}\widehat{\mathbb{P}}^{\mathbf{x},\theta}_{t_k}}(\widehat{T}\mathbf{X})\mathrm{d}\mathbb{P}^{\mathbf{x},\theta}_{t_k}\,,$$

so that

$$\frac{\mathrm{d}\mathbb{P}^{\mathbf{x}}_{t_k}}{\mathrm{d}\mathbb{P}^{\mathbf{x},\theta}_{t_k}}(\mathbf{X}) = \frac{\mathrm{d}\widehat{\mathbb{P}}^{\mathbf{x}}_{t_k}}{\mathrm{d}\widehat{\mathbb{P}}^{\mathbf{x},\theta}_{t_k}}(\widehat{T}\mathbf{X}), \qquad \mathbb{P}^{\mathbf{x},\theta}_{t_k}\text{-a.s.}$$

In particular, using that $\mathbb{P}^{\mathbf{x},\theta}_{t_k} \sim \mathbb{P}^{\mathbf{x}}_{t_k}$,

$$\mathbb{E}_{\mathbb{P}^{\mathbf{x}}_{t_k}}\left[\log\left(\frac{\mathrm{d}\mathbb{P}^{\mathbf{x}}_{t_k}}{\mathrm{d}\mathbb{P}^{\mathbf{x},\theta}_{t_k}}(\mathbf{X})\right)\right]$$
$$= \mathbb{E}_{\mathbb{P}^{\mathbf{x}}_{t_k}}\left[\int_{t_k}^{t_{k+1}} \sqrt{\tfrac{\bar{\beta}_t}{2}}\Big(b(t,\mathbf{X}_t) - b_\theta(t_k,\mathbf{x})\Big)\mathrm{d}\mathrm{B}_t + \frac{1}{4}\int_{t_k}^{t_{k+1}} \bar{\beta}_t\Big(\|b(t,\mathbf{X}_t) - b_\theta(t_k,\mathbf{x})\|^2\Big)\mathrm{d}t\right]\,,$$

which concludes the proof. □

Corollary H.6. *Under Assumption 3.1,*

$$\mathbb{E}\left[\int_{t_k}^{t_{k+1}} \frac{\bar{\beta}_t}{2}\left\|b(t,\overleftarrow{\mathbf{X}}_t) - b_\theta(t_k,\mathbf{x})\right\|^2 \mathrm{d}t\right] < \infty\,.$$

Thus, we get

$$\mathrm{KL}\left(\mathbb{P}^{\mathbf{x}}_{t_k}\middle\|\mathbb{P}^{\mathbf{x},\theta}_{t_k}\right) = \frac{1}{4}\mathbb{E}\left[\int_{t_k}^{t_{k+1}} \bar{\beta}_t\,\|b(t,\overleftarrow{\mathbf{X}}_t) - b_\theta(t_k,\mathbf{x})\|^2\,\mathrm{d}t\right]\,.$$

Proof. Note that,

$$\mathbb{E}\left[\int_{t_k}^{t_{k+1}} \frac{\bar{\beta}_t}{2}\left\|b(t,\overleftarrow{\mathbf{X}}_t) - b_\theta(t_k,\mathbf{x})\right\|^2 \mathrm{d}t\right] \le \mathbb{E}\left[\int_{t_k}^{t_{k+1}} \bar{\beta}_t\left\|b(t,\overleftarrow{\mathbf{X}}_t)\right\|^2 dt\right] + \|b_\theta(t_k,\mathbf{x})\|^2\int_{t_k}^{t_{k+1}}\bar{\beta}_t \mathrm{d}t$$

Moreover, using Proposition D.4, there exists a universal constant C_k (depending on $[t_k, t_{k+1}]$), such that

$$\mathbb{E}\left[\int_{t_k}^{t_{k+1}} \bar{\beta}_t\left\|b(t,\overleftarrow{\mathbf{X}}_t)\right\|^2 \mathrm{d}t\right] \le C_k\int_{t_k}^{t_{k+1}}\bar{\beta}_t\mathbb{E}\left[1+\left\|\overleftarrow{\mathbf{X}}_t\right\|^{2p+2}\right]\mathrm{d}t\,.$$

Using Proposition 3.2 with $\ell = 2p+2$, we have $\sup_{t\in[t_k,t_{k+1}]}\mathbb{E}[\|\overleftarrow{\mathbf{X}}_t\|^{2p+2}] < \infty$. Therefore,

$$\mathbb{E}\left[\int_{t_k}^{t_{k+1}} \bar{\beta}_t\left\|b(t,\overleftarrow{\mathbf{X}}_t)\right\|^2 \mathrm{d}t\right] < \infty\,,$$

which implies

$$\mathbb{E}\left[\log\left(\frac{\mathrm{d}\mathbb{P}^{\mathbf{x}}_{t_k}}{\mathrm{d}\mathbb{P}^{\mathbf{x},\theta}_{t_k}}\left((\overleftarrow{\mathbf{X}}_t)_{t\in[t_k,t_{k+1}]}\right)\right)\right] = \frac{1}{4}\mathbb{E}\left[\int_{t_k}^{t_{k+1}}\bar{\beta}_t\,\left\|b(t,\overleftarrow{\mathbf{X}}_t) - b_\theta(t_k,\mathbf{x})\right\|^2\mathrm{d}t\right]\,.$$

□

Backward evolution of the score. Consider the process obtained by evaluating the score function along the backward trajectory, namely

$$\mathbf{Y}_t := \mathrm{S}_{T-t}\left(\overleftarrow{\mathbf{X}}_t\right)\,,$$

where $(\overleftarrow{\mathbf{X}}_t)_{t\ge0}$ is a weak solution to (2). This process can be shown to satisfy a SDE on the entire interval [0,T], as detailed in the following result. The well-posedness of this SDE, as well as the validity of the representation (56), rely critically on Corollary D.5.

Lemma H.7. *Suppose that Assumption 3.1 hold. Then, for any $\mathbf{x}\in\mathbb{R}^d$, we have*

$$\mathrm{d}\mathbf{Y}_t = -\alpha\bar{\beta}_t\mathbf{Y}_t\mathrm{d}t + \sqrt{2\bar{\beta}_t}\nabla^2\,\mathrm{S}_{T-t}\left(\overleftarrow{\mathbf{X}}_t\right)\mathrm{dB}_t\,. \tag{56}$$

Proof. Consider the Fokker–Planck equation associated with the forward process (1), *i.e.*,

$$\partial_t p_t(\mathbf{x}) = \alpha\beta_t\,\mathrm{div}(\mathbf{x}p_t(\mathbf{x})) + \beta_t\Delta p_t(\mathbf{x})\,, \tag{57}$$

for $\mathbf{x}\in\mathbb{R}^d, t\in(0,T]$. We manipulate the preceding equation to derive the partial differential equation satisfied by the function $(t,x)\mapsto\log p_t(x)$. Thus,

$$\partial_t\log p_t(\mathbf{x}) = \alpha\beta_t\frac{\mathrm{div}(\mathbf{x}p_t(\mathbf{x}))}{p_t(\mathbf{x})} + \beta_t\frac{\Delta p_t(\mathbf{x})}{p_t(\mathbf{x})}\,.$$

Introducing $\Delta p_t/p_t = \Delta\log p_t + \|\nabla\log p_t\|^2$ and expanding the previous equation yields

$$\partial_t\log p_t(\mathbf{x}) = \alpha\beta_t(d + \langle\mathbf{x},\nabla\log p_t\rangle) + \beta_t\left(\Delta\log p_t + \|\nabla\log p_t\|^2\right)\,. \tag{58}$$

Using Itô's Lemma, $Y_t = \nabla\phi_t(\overleftarrow{\mathbf{X}}_t)$, with $\phi_t(x) = \log p_{T-t}(x)$ we get

$$
\begin{aligned}
&\mathrm{d}Y_t \\
&= \left[\partial_t \nabla \phi_t\left(\overleftarrow{\mathbf{X}}_t\right) + \nabla^2\phi_t\left(\overleftarrow{\mathbf{X}}_t\right)\left(\alpha\bar{\beta}_t\overleftarrow{\mathbf{X}}_t + 2\bar{\beta}_t\nabla\phi_t\left(\overleftarrow{\mathbf{X}}_t\right)\right) + \bar{\beta}_t\Delta\nabla\phi_t\left(\overleftarrow{\mathbf{X}}_t\right)\right]\mathrm{d}t + \sqrt{2\bar{\beta}_t}\nabla^2\phi_t\left(\overleftarrow{\mathbf{X}}_t\right)\mathrm{dB}_t \\
&= \left[\nabla\left(\partial_t\phi_t\left(\overleftarrow{\mathbf{X}}_t\right) + \bar{\beta}_t\left(\Delta\phi_t\left(\overleftarrow{\mathbf{X}}_t\right) + \left\|\nabla\phi_t\left(\overleftarrow{\mathbf{X}}_t\right)\right\|^2\right)\right) + \alpha\bar{\beta}_t\nabla^2\phi_t\left(\overleftarrow{\mathbf{X}}_t\right)\overleftarrow{\mathbf{X}}_t\right]\mathrm{d}t + \sqrt{2\bar{\beta}_t}\nabla^2\phi_t\left(\overleftarrow{\mathbf{X}}_t\right)\mathrm{dB}_t\,,
\end{aligned}
$$

where we used that, for $\mathbf{x} \in \mathbb{R}^d$ $2\nabla^2\phi_t(\mathbf{x})\nabla\phi_t(\mathbf{x}) = \nabla\left\|\nabla\phi_t(\mathbf{x})\right\|^2$. By (58),

$$
\partial_t\phi_t(\mathbf{x}) = -\alpha\beta_{T-t}(d + \langle\mathbf{x}, \nabla\log p_t\rangle) - \bar{\beta}_t\left(\Delta\log p_t + \|\nabla\log p_t\|^2\right)\,,
$$

which implies

$$
\mathrm{d}Y_t = \left[-\alpha\bar{\beta}_t\left(\nabla\phi_t(\overleftarrow{\mathbf{X}}_t) + \nabla^2\phi_t(\overleftarrow{\mathbf{X}}_t)\overleftarrow{\mathbf{X}}_t\right) + \alpha\bar{\beta}_t\nabla^2\phi_t(\overleftarrow{\mathbf{X}}_t)\overleftarrow{\mathbf{X}}_t\right]\mathrm{d}t + \sqrt{2\bar{\beta}_t}\nabla^2\phi_t\left(\overleftarrow{\mathbf{X}}_t\right)\mathrm{dB}_t\;.
$$

which implies (56). Note that the Fokker–Planck equation (57) is only known to hold for $t \in (0, T]$. Therefore, the validity of this SDE representation cannot be extended directly up to the terminal time T. To overcome this limitation, it is necessary to establish additional regularity properties of the stochastic terms involved. In particular, we must show that the quadratic variation of the martingale component in (56) remains uniformly bounded on $[0, T]$, which allows the SDE to be extended to the full time horizon and ensures the well-posedness of the backward dynamics.

Applying Corollary D.5, there exists a constant $C > 0$, independent of $t \in [0, T)$, that bounds the quadratic variation of (56) as follows

$$
\mathbb{E}\left[\int_0^t 2\bar{\beta}_s\left\|\nabla\,\mathrm{S}_{T-s}\left(\overleftarrow{\mathbf{X}}_s\right)\right\|_{\mathrm{F}}^2\mathrm{d}s \le\right]\mathbb{E}\left[\int_0^t 4\bar{\beta}_s\,\mathrm{C}_{H,T-s}^2\left(1 + \left\|\overleftarrow{\mathbf{X}}_s\right\|^{4p+4}\right)\mathrm{d}s\right] \le C\,,
$$

where in the last inequality we used the continuity of $t \mapsto \mathrm{C}_{H,t}$ from Corollary D.5, together with the Gaussian representation (21). Indeed, Lemma C.1 shows that the law of $\overleftarrow{\mathbf{X}}_t$, equal to the law of $\overrightarrow{\mathbf{X}}_{T-t}$, is the product of a continuous factor times a Gaussian convolution. As both the normal distribution and p_0 admit any order moment, from Lemma A.1, we have that the previous bound is uniform in time and does not explode for $t \to T$. □

I. Numerical illustration of forgetting in controlled settings

This appendix describes the numerical protocols used to illustrate the *forgetting/stability mechanism* predicted by the Harris-type stability result established in the main text.

I.1. Synthetic datasets

We study the following probability distributions that satisfy Assumption 3.1.

Gaussian random vectors. We set dimension to $d = 50$ and consider Gaussian data distributions of the form

$$
\pi_{\text{data}} = \mathcal{N}(\mu, \Sigma), \qquad \mu := \mathbf{1}_d \in \mathbb{R}^d,
$$

where $\mathbf{1}_d$ denotes the vector of ones. We study three covariance structures:

1. **Isotropic.** We take an isotropic covariance

   $$
   \Sigma^{(\text{iso})} := \sigma^2\mathbf{I}_d, \qquad \sigma^2 = 0.1,
   $$

   so that all coordinates have the same variance and are uncorrelated.

2. **Anisotropic (Heteroscedastic).** We consider a diagonal covariance with two variance levels:

   $$
   \Sigma^{(\text{heterosc})} := \mathrm{diag}(v_1, \ldots, v_d), \qquad v_j := \begin{cases} 1, & 1 \le j \le 5, \\ 10^{-3}, & 6 \le j \le d, \end{cases}
   $$

   so that the first five coordinates have much larger variance than the remaining ones (no cross-correlation).

3. **Correlated.** We use a dense covariance matrix with unit marginal variances and slowly decaying off-diagonal correlations:

$$\Sigma^{(\mathrm{corr})}_{jj} := 1, \qquad \Sigma^{(\mathrm{corr})}_{jj'} := \frac{1}{\sqrt{|j-j'|+1}} \quad \text{for } j \neq j',$$

for $1 \leq j, j' \leq d$.

Gaussian Mixture Model. We consider a Gaussian mixture model in $\mathbb{R}^{50}$ defined by

$$p_{\mathrm{data}}(\mathbf{x}) = \sum_{i=1}^{25} \omega_i \mathcal{N}(\mu_i, \Sigma_i) \ ,$$

where

- $\{\omega_i\}_{i=1}^{25}$ are sampled i.i.d. from a ξ^2 distributions with 3 degrees of freedom and re-normalized to sum to 1,
- $\{\mu_i\}_{i=1}^{25}$ have all but the first 2 coordinates at 0. The first two coordinates are evenly spaced at a square with lower corner at $(-10,-10)$ and upper corner at $(10,10)$,
- Each Σ_i is of the type $U_i^T D U_i$ where U_i is obtained as one of the orthonormal matrices from the SVD of a random Gaussian matrix $\tilde{U}_i$ and D is a diagonal matrices with entries $(1, 1/2, \cdots, 1/25)$.

We used a single draw of the aforementioned random numbers for all the experiments.

I.2. Sensitivity to initialization

Gaussian targets. This experiment isolates the effect of a controlled perturbation injected at an intermediate diffusion time $t_{\mathrm{bias}} \in (0,T]$ on the final generated sample, and studies how this effect varies with the perturbation time and the perturbation magnitude.

Fix $t_{\mathrm{bias}} \in (0,T]$ and consider the backward reverse-time transition represented by the Markov kernel $\mathrm{Q}_{T|t_{\mathrm{bias}}}$, together with its numerical approximation $\bar{\mathrm{Q}}_{T|t_{\mathrm{bias}}}$ obtained by time-discretization. We investigate whether we observe a robustness behavior for the *discretized* reverse chain with respect to the true kernel by monitoring the discrepancy

$$\mathcal{W}_2\big(p_{t_{\mathrm{bias}}}\mathrm{Q}_{T|t_{\mathrm{bias}}}, \tilde{p}_{t_{\mathrm{bias}}}\bar{\mathrm{Q}}_{T|t_{\mathrm{bias}}}\big) = \mathcal{W}_2\big(\pi_{\mathrm{data}}, \tilde{p}_{t_{\mathrm{bias}}}\bar{\mathrm{Q}}_{T|t_{\mathrm{bias}}}\big) \ .$$

We define a biased initialization at time $t_{\mathrm{bias}}$by applying a shift in a random unit direction u:

$$\mathbf{x}^{(i)}_{t_{\mathrm{bias}},\lambda} = \mathbf{x}^{(i)}_{t_{\mathrm{bias}}} + \lambda u, \qquad \|u\| = 1,$$

where u is drawn once per replication and $\lambda \geq 0$ is chosen to control the bias magnitude. For each $(t_{\mathrm{bias}}, \lambda)$, we run an Euler–Maruyama discretization of the reverse-time dynamics from time t_{bias} down to T and obtain samples $\hat{\mathbf{x}}^{(i)}_T$. We fix the step-size to h throughout the experiment. A pseudocode of the protocol is given in Algorithm 1. On Gaussian targets, the *exact* reverse semigroup admits closed-form expressions and satisfies a strict contraction in $\mathcal{W}_2$, which provides a clean reference behavior. The goal here is not to re-derive these formulas, but to assess whether the *numerical* reverse chain exhibits the same qualitative robustness to perturbations.

Since the target is Gaussian, we quantify errors by the closed-form Gaussian $\mathcal{W}_2$ (Bures) distance between the empirical output (via its sample mean/covariance) and $\pi_{\mathrm{data}} = \mathcal{N}(\mu, \Sigma)$. Each curve in Figure 5 is obtained by averaging over 5 independent replications and 30 000 points are generated for each bias at each perturbation time. We use for constant step size and linear (non optimized) schedule with maximal value $\beta_{\max} = 20$ and minimal value $\beta_{\min} = 0.1$ in both the VP and VE case. We use the forward-time convention: $t = 0$ corresponds to π_{data} and $t = T$ to the noisiest level used to initialize the reverse sampler. Across all covariance structures and for both VE and VP dynamics, we observe a clear forgetting trend: perturbations injected at larger diffusion times (*i.e.*, early along the reverse trajectory) are substantially discounted. This behavior is consistent with the contraction/robustness mechanism established in the main text: the reverse-time dynamics progressively contracts discrepancies as it evolves toward the data distribution.

Algorithm 1 Initialization-bias robustness at time $t_{\text{bias}} \in \mathcal{T}_{\text{perturb}}$.

Target distribution π_{data}, forward SDE, reverse SDE discretization (Euler–Maruyama) with step size h, horizon T, number of points to generate M, perturbation times $\mathcal{T}_{\text{perturb}} \subset (0, T]$, bias magnitudes $\Lambda \subset \mathbb{R}_+$, number of replications rep.
for $r = 1, \dots, \text{rep}$ **do**
  Sample a unit vector $u^{(r)} \sim \text{Unif}(\mathbb{S}^{d-1})$
  for each $t \in \mathcal{T}_{\text{perturb}}$ **do**
    for each $\lambda \in \Lambda$ **do**
      Sample $\mathbf{x}_t^{(i)} = \mathrm{m}_{t|0}\, \mathbf{x}_0^{(i)} + \sigma_{t|0} G^{(i)}$ for $i = 1, \dots, M$
      Set biased initialization $\mathbf{x}_{t,\lambda}^{\text{bias}(i)} \leftarrow \mathbf{x}_t^{(i)} + \lambda u^{(r)}$ for $i = 1, \dots, M$
      Run discretization of the reverse dynamics from time t to T with initial points $\{\mathbf{x}_{t,\lambda}^{\text{bias}(i)}\}_{i=1}^M$
      Obtain $\{\hat{\mathbf{x}}_{T,\lambda}^{(i)}\}_{i=1}^M$ and empirical law $\widehat{\pi}_{T,\lambda}^{\text{bias}}$
      Compute $\hat{\mathcal{W}}_2(t, \lambda) \leftarrow \mathcal{W}_2\left(\widehat{\pi}_{T,\lambda}^{\text{bias}}, \pi_{\text{data}}\right)$
    end for
  end for
end for
Plot: for each $\lambda \in \Lambda$, plot $t \mapsto \frac{1}{\text{rep}} \sum_{r=1}^{\text{rep}} \hat{\mathcal{W}}_2^{(r)}(t, \lambda)$ (with $\pm$ one std).

Mixture of Gaussian targets. The experiment is similar to the one for Gaussien targets. However, as there is no analytical formula for the Wasserstein distance in this case, we employ the Maximum sliced Wasserstein distance (Max SW) as in Kolouri et al. (2019, Section 4.3) using a total of 10^6 samples. The optimization procedure employs the Adam optimization algorithm with learning rate 10^{-3} and is stop when either the last optimization update is smaller than 10^{-7} or the optimization has reached a total of 10^5 iterations.

In this setting, we consider exclusively the *Variance Exploding* framework and define for $\lambda \in \mathbb{R}_+$ an initialization perturbation of the type

$$\mathbf{x}_{t_{\text{bias}},\lambda}^{(i)} = \mathbf{x}_{t_{\text{bias}}}^{(i)} + \sigma_{t_{\text{bias}}|0}^2 \lambda u_{\max}$$

where $u_{\max} \in \mathbb{S}_{49}$ is obtained as the solution of the optimization problem for the Max SW distance for two sets of 2×10^5 independent samples of p_{data}. All the calculations have been replicated 20 times using different seeds and the reported results consist of the mean and the standard deviations of those 20 replicates. For the scheduling, we used the scheduler from Karras et al. (2022b, Equation 5) with $N = 100$, $\sigma_{\min} = 0.002$, $\sigma_{\max} = 80$ and $\rho = 3$. Results are reported in Figure 3 and Figure 1.

I.3. Sensitivity to approximation errors

This section describes a second numerical protocol designed to analyse the *local-to-global error propagation* predicted by the stability results in the main text. Instead of modifying the initialization, we introduce a *single local perturbation* in the reverse-time dynamics: at one prescribed time $t_{\text{err}} \in [0, T]$, we perturb the score term used by the Euler–Maruyama discretization, and we measure how the resulting output error depends on the location t_{err} and magnitude of this perturbation.

Fix $n_{\text{err}} \in \{0, \cdots, N-1\}$ and consider the Euler–Maruyama discretization of the reverse-time SDE on the uniform grid. We run the reverse chains starting from the same terminal initialization $\mathbf{x}_0^{(i)} \sim \pi_\infty$ with the exact score function except at the unique step corresponding to t_{err} (*i.e.*, $t_{n_{\text{err}}}$), where we replace the score by a perturbed version only at the step $t_{n_{\text{err}}}$.

Gaussian targets. We consider the following perturbation with u a random unit direction

$$\tilde{\mathrm{S}}_{t_{\text{err}}}(\mathbf{x}) = \mathrm{S}_{t_{\text{err}}}(\mathbf{x}) + \frac{\lambda}{\sigma_{t_{\text{err}}|0}^2} u\,,$$

A pseudocode of the protocol is given in Algorithm 2. Once again, because the target distribution is Gaussian, we quantify the output error using the closed-form Gaussian $\mathcal{W}_2$ (Bures) distance between the empirical output distribution (via its sample mean and covariance) and π_{data}. Repeating the experiment over independent replications (fresh draws of the terminal samples and Brownian increments) yields mean and standard deviation curves as functions of t_{err} and λ in Figure 6.

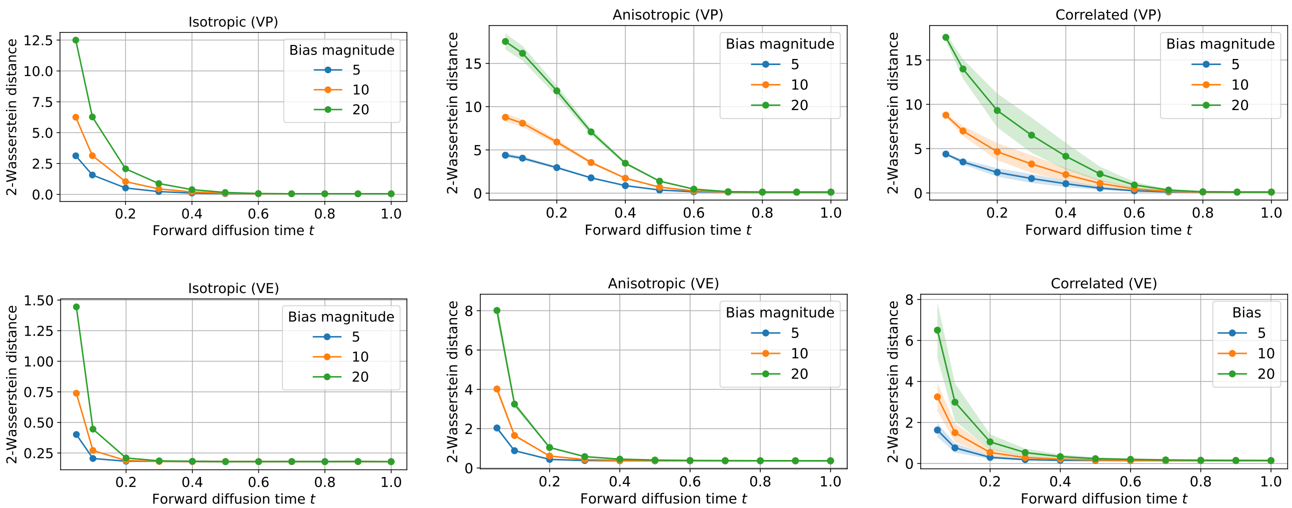

Figure 5. Sensitivity to initialization for isotropic, heteroscedastic, and correlated Gaussian targets. Curves report the mean over 5 independent replications of Algorithm 1 ($T = 1$, $M = 30{,}000$, $h = 2.5 \times 10^{-3}$); shaded regions indicate ± 1 standard deviation across replications. The horizontal axis is the forward diffusion time t. Each dotted point on the lines correspond to a perturbation time and the bias magnitude corresponds to the choice of λ.

Mixture of Gaussian targets. Once again, the experiment is simlar to the one for Gaussien targets. But with the use of the Maximum sliced Wasserstein distance (Max SW). The rest of the conditions is similar to the previous experiment on initialization bias. Also, in this setting, we consider exclusively the *Variance Exploding* framework and define for $\lambda \in \mathbb{R}_+$ and a score perturbation of the type

$$\tilde{\mathrm{S}}_{t_{\mathrm{err}}}(\mathbf{x}) = \mathrm{S}_{t_{\mathrm{err}}}(\mathbf{x}) + \frac{\lambda}{\sigma^2_{t_{\mathrm{err}}|0}} u_{\max}\,, \tag{59}$$

where $u_{\max} \in \mathbb{S}_{49}$ is obtained as the solution of the optimization problem for the Max SW distance for two sets of 2×10^5 independent samples of p_{data}. All the calculations have been replicated 20 times using different seeds and the reported results consist of the mean and the standard deviations of those 20 replicates. Results are reported in Figure 4 and Figure 2.

Algorithm 2 Local score-perturbation study (one-step score error at index n_{err}).

Inputs: target π_{data}; reverse-time Euler–Maruyama sampler with step size h and grid $\{t_n\}_{n=0}^{N-1}$; number of points to generate M; error indices $\mathcal{N}_{\text{err}} \subset \{0, \dots, N-1\}$; magnitudes $\Lambda \subset \mathbb{R}_+$; replications rep.
for $r = 1, \dots, \text{rep}$ **do**
  Sample $\{\mathbf{x}_0^{(i)}\}_{i=1}^M \sim \pi_\infty$.
  Sample $u^{(r)} \sim \text{Unif}(\mathbb{S}^{d-1})$.
  for each $n_{\text{err}} \in \mathcal{N}_{\text{err}}$ **do**
    for each $\lambda \in \Lambda$ **do**
      Initialize $\mathbf{x}^{(i)} \leftarrow \mathbf{x}_T^{(i)}$ for $i = 1, \dots, M$.
      for $n = 0, \dots, N-1$ **do**
        Euler–Maruyama update:

$$\mathbf{x}_{n+1}^{(i)} \leftarrow \mathbf{x}_n^{(i)} + \left(\alpha \bar{\beta}_{t_k} \mathbf{x}_n^{(i)} + 2\bar{\beta}_{t_n} \tilde{s}_{T-t_n}^{(i)}(\mathbf{x}_n^{(i)})\right) h + \sqrt{2\bar{\beta}_{t_n}} \sqrt{h} \xi_n^{(i)} \,, \qquad \xi_n^{(i)} \sim \mathcal{N}(0, \mathbf{I}_d) \,,$$

        with

$$\tilde{s}_{T-t_n}^{(i)}(\mathbf{x}_n^{(i)}) \leftarrow \mathrm{S}_{T-t_n}\left(\mathbf{x}_n^{(i)}\right) + \mathbf{1}_{\{n=n_{\text{err}}\}} \frac{\lambda}{\sigma_{t_n|0}^2} u^{(r)} \qquad \text{for} i = 1, \dots, M \,.$$

      end for
      Let $\hat{\pi}_{N,\lambda,n_{\text{err}}}$ be the empirical law of $\{\mathbf{x}_N^{(i)}\}_{i=1}^M$.
      Compute $\hat{\mathcal{W}}_2(\lambda, n_{\text{err}}) \leftarrow \mathcal{W}_2\left(\hat{\pi}_{N,\lambda,n_{\text{err}}}, \pi_{\text{data}}\right)$
    end for
  end for
end for
Plot: for each $\lambda \in \Lambda$, plot $t \mapsto \frac{1}{\text{rep}} \sum_{r=1}^{\text{rep}} \hat{\mathcal{W}}_2^{(r)}(\lambda, n_{\text{err}})$ (with $\pm$ one std).

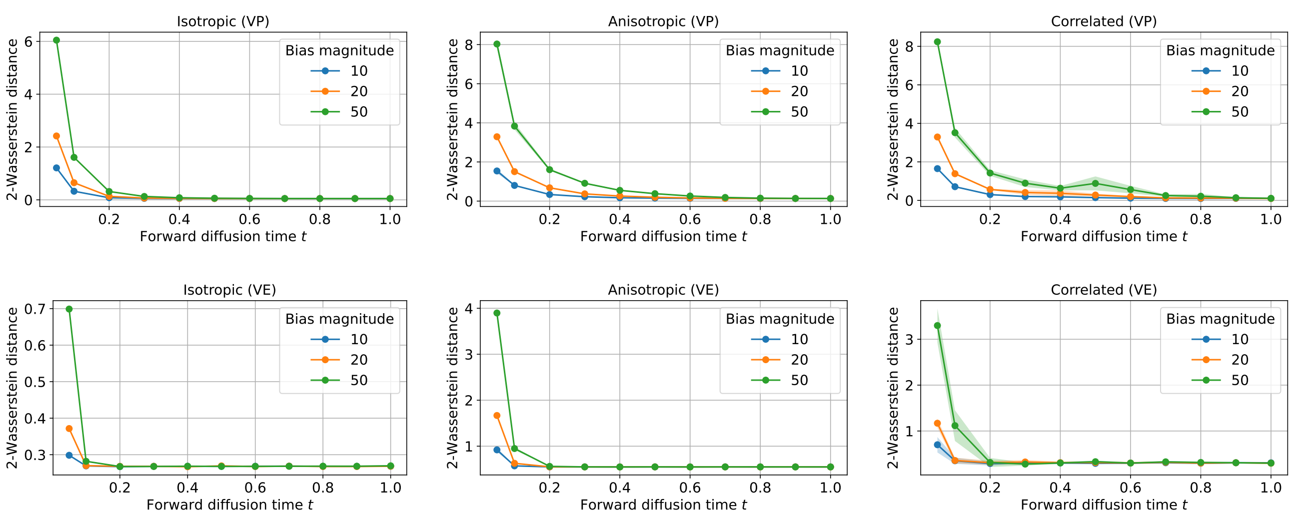

Figure 6. Sensitivity to local score error for isotropic, heteroscedastic, and correlated Gaussian targets. Curves report the mean over 5 independent replications of Algorithm 1 ($T = 1$, $M = 30{,}000$, $h = 5 \times 10^{-3}$); shaded regions indicate ± 1 standard deviation across replications. The horizontal axis is the forward diffusion time t. Each dotted point on the lines correspond to a perturbation time and the bias magnitude corresponds to the choice of λ.