Journal of Machine Learning Research 26 (2025) 1-48 Submitted 3/25; Revised 9/25; Published 11/25

Optimal Denoising in Score-Based Generative Models: The Role of Data Regularity

Eliot Beyler ELIOT.BEYLER@INRIA.FR
INRIA, Ecole Normale Supérieure
PSL Research University
Paris, France

Francis Bach FRANCIS.BACH@INRIA.FR
INRIA, Ecole Normale Supérieure
PSL Research University
Paris, France

Editor: Jianfeng Lu

Abstract

Score-based generative models achieve state-of-the-art sampling performance by denoising a distribution perturbed by Gaussian noise. In this paper, we focus on a single deterministic denoising step, and compare the optimal denoiser for the quadratic loss, we name "full-denoising", to the alternative "half-denoising" introduced by Hyvärinen (2025). We show that looking at the performance in terms of distance between distributions tells a more nuanced story, with different assumptions on the data leading to very different conclusions. We prove that half-denoising is better than full-denoising for regular enough densities, while full-denoising is better for singular densities such as mixtures of Dirac measures or densities supported on a low-dimensional subspace. In the latter case, we prove that full-denoising can alleviate the curse of dimensionality under a *linear manifold hypothesis*.

Keywords: Score-based generative modeling, denoising, Wasserstein distance, diffusion models

1. Introduction

Score-based generative models, or diffusion models (Sohl-Dickstein et al., 2015; Saremi and Hyvärinen, 2019; Ho et al., 2020; Song et al., 2021b), achieve state-of-the-art sampling performance by denoising a distribution perturbed by Gaussian noise. This denoising is made in several steps by removing each time a fraction of the noise, which can be seen as the discretization of a stochastic differential equation.

Here we focus on a single denoising step. This setting enables a more in-depth study, that could be generalized to multiple steps in future works, but is also relevant in practice for several reasons. Firstly, it has been used in recent work by Saremi et al. (2023), who propose an alternative formulation of diffusion models in which each step corresponds to log-concave sampling. The method first reduces the noise level by averaging multiple measurements, then tries to approximate the data distribution with a one-step denoising. Moreover, in the framework of *stochastic localization* (Montanari, 2023), one simulates a process that localizes to the distribution of interest, but as this process is simulated in finite time, the

author specifies that an additional step is needed, for example by taking the conditional expectation, hence denoising. This final denoising step is also present in denoising diffusion models even if not stated explicitly. Indeed, when one lacks regularity assumptions on the data distribution, proofs of convergence for diffusion models fail, and to overcome this, Chen et al. (2023) use early stopping of the denoising process, as keeping a little bit of noise offsets the missing regularity. But then should we keep this extra noise or do an additional denoising step? This last step is examined here, and relies crucially on assumptions about the data distribution.

Formally, given a random variable $X \in \mathbb{R}^d$ with distribution μ_X, we will define $Y = X + \varepsilon$ with X and ε independent and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. The "optimal" denoiser, in the sense of the measurable function that minimizes $\mathbb{E}[\|X - f(Y)\|^2]$, is the conditional expectation $\varphi(y) = \mathbb{E}[X|Y = y]$.[1] This value is related to the score, i.e., the gradient of the logarithm of the density p_Y of Y, through Tweedie's formula (Robbins, 1956; Miyasawa, 1961; Efron, 2011):

$$\varphi(y) = \mathbb{E}[X|Y = y] = y + \sigma^2 \nabla \log p_Y(y). \tag{1}$$

But does it minimize the distance between the distribution μ_X and $\mu_{f(Y)}$ (in Wasserstein or any other distance between probability distributions)?

Of course, we want to limit the space of functions f to ones we can compute.[2] Knowing that in practice, $\nabla \log p_Y$ can be estimated by *denoising score matching* (Vincent, 2011), it is reasonable to look for an expression involving $\nabla \log p_Y$. In a recent paper, Hyvärinen (2025) introduces half-denoising

$$\varphi_{1/2}(y) = \frac{y + \mathbb{E}[X|Y = y]}{2} = y + \frac{\sigma^2}{2} \nabla \log p_Y(y),$$

and uses it to generate samples from X with a modified Langevin algorithm. Our aim in this paper is to study in more details the performance of this half-denoising step and to compare it to full-denoising.

The general philosophy here is that the hypotheses made on the data distribution are essential to assess the performance of the different denoising processes. We distinguish between two kinds of assumptions made in the literature. The first kind is to assume that the distribution of X is regular enough, i.e., that it admits a density p_X with respect to the Lebesgue measure, and that this density is smooth, for example that $x \mapsto \nabla \log p_X(x)$ is L-Lipschitz (see, e.g., Chen et al., 2023). The second kind, incompatible with the first, is to assume that the distribution has some kind of singularities, i.e., that it is concentrated on low dimensional manifolds, with pockets of mass separated by areas of low densities. This framework is known as the *manifold hypothesis* (see, e.g., Tenenbaum et al., 2000; Bengio et al., 2013; Fefferman et al., 2016).

Contributions. In this work, we make the following contributions:

- We show that half-denoising is better for regular enough densities, in the sense that the distance (in MMD – maximum mean discrepancy – and in Wasserstein-2 distance)

1. For this conditional expectation to exist, we will always assume that X is integrable, i.e., $\mathbb{E}[\|X\|] < \infty$.
2. In fact, any distance between μ_X and $\mu_{f(Y)}$ can be made zero by taking f a transport map, but it will not be computable in general.

between the original distribution and the denoised distribution is of order $O(\sigma^4)$, compared to $O(\sigma^2)$ for full-denoising. We thus formalize and extend the scaling in $O(\sigma^4)$ obtained by Hyvärinen (2025). On the contrary, full-denoising is better for singular distributions such as Dirac measures or Gaussian with small variance compared to the additional noise (section 3).

- When the variable is supported on a lower-dimensional subspace, we show that there is a trade-off between full-denoising which reduces the Wasserstein distance by ensuring that the output belongs to the subspace, and half-denoising that reduces the Wasserstein distance on the lower-dimensional subspace (assuming a regular enough density). Moreover, in the case where the subspace is of small enough dimension compared to the full space, we show that full-denoising is adaptive to this low dimensional structure and thus alleviates the curse of dimensionality as the Wasserstein distance only depends on the distance between distributions on the lower-dimensional subspace (section 4).
- We show that the denoising performance for a mixture of distributions with disjoint compact supports behaves as if we were denoising each variable independently, plus an exponentially decreasing term (section 5).
- Finally, combining these results, we show that for a linear version of the manifold hypothesis, where the data distribution is supported on disjoint compact sets, each of these belonging to a (different) linear subspace of low dimension, full-denoising can alleviate the curse of dimensionality even if the support of the distribution itself spans the whole space as it adapts to the local linear structure of the distribution.

2. Notations

We introduce the following notations:

- For $\alpha \in \mathbb{R}$, we denote $\varphi_\alpha(y) = y + \alpha\sigma^2 \nabla \log p_Y(y)$, such that $\alpha = 1/2$ corresponds to half-denoising, $\alpha = 1$ to full-denoising (and $\alpha = 0$ to no denoising at all).
- $\|\cdot\|$ the euclidean norm on $\mathbb{R}^d$, $B(x, r)$ the Euclidean ball of center x and radius r.
- $\mathcal{N}(\mu, \Sigma)$ the multivariate Gaussian distribution of mean μ and covariance matrix Σ.
- For a random variable $Z \in \mathbb{R}^n$, we write $\mathcal{L}(Z)$ its law, and, when it exists, p_Z its density with respect to the Lebesgue measure. We write $Z \sim \mu$ if $\mathcal{L}(Z) = \mu$, $Z_1 \sim Z_2$ if $\mathcal{L}(Z_1) = \mathcal{L}(Z_2)$ and $Z_1 \perp Z_2$ if Z_1 and Z_2 are independent. We also denote, for $\xi \in \mathbb{R}^n$, $\hat{p}_Z(\xi) = \mathbb{E}[e^{i\xi \cdot Z}]$ the characteristic function of Z.
- For $p \geq 1$, we define the p-Wasserstein distance between $\mathcal{L}(Z_1)$ and $\mathcal{L}(Z_2)$ as

$$W_p(\mathcal{L}(Z_1), \mathcal{L}(Z_2)) = \left(\inf_{\Gamma : \Gamma_{z_1} = \mathcal{L}(Z_1), \Gamma_{z_2} = \mathcal{L}(Z_2)} \int \|z_1 - z_2\|^p d\Gamma(z_1, z_2) \right)^{1/p},$$

where $\{\Gamma : \Gamma_{z_1} = \mathcal{L}(Z_1), \Gamma_{z_2} = \mathcal{L}(Z_2)\}$ is the set of distributions on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals $\mathcal{L}(Z_1)$ and $\mathcal{L}(Z_2)$ (see, e.g., Peyré and Cuturi, 2019).

- For a reproducing kernel Hilbert space $\mathcal{H}$, with kernel $k : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$, we define the maximum mean discrepancy (MMD) (Gretton et al., 2012) between $\mathcal{L}(Z_1)$ and $\mathcal{L}(Z_2)$ as

$$\mathrm{MMD}_k(\mathcal{L}(Z_1), \mathcal{L}(Z_2)) = \sup_{f \in \mathcal{H}, \|f\|_{\mathcal{H}} \leq 1} \left(\mathbb{E}[f(Z_1)] - \mathbb{E}[f(Z_2)]\right).$$

- For $k \in \{0, 1, ..\} \cup \{\infty\}$, d_1, d_2 integers, we write $\mathcal{C}^k(\mathbb{R}^{d_1}, \mathbb{R}^{d_2})$ the functions from $\mathbb{R}^{d_1}$ to $\mathbb{R}^{d_2}$ with k continuous derivatives. If $d_2 = 1$, we simply write $\mathcal{C}^k(\mathbb{R}^{d_1})$.

- For a linear operator, $A : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$, we write $\|A\|_{\mathrm{op}}$ the operator norm of A, defined by $\|A\|_{\mathrm{op}} = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}$.

- ∇ the gradient operator, ∇^2 the Hessian operator, $\nabla \cdot$ the divergence operator and Δ the Laplacian, that will always be taken with respect to the space variable $x \in \mathbb{R}^d$.

3. Half-Denoising is Better for Regular Densities

In this section, we show that half-denoising is better for regular enough densities. We start by studying Gaussian variables, for which we have closed-form expressions of the various Wasserstein distances (section 3.1). It provides insights into the behavior of the denoiser φ_α, but will also show that the bound we prove in the following section are tight in the dependence with respect to σ. Then in section 3.2, inspired by Hyvärinen (2025), we prove a bound on the distance between the characteristic functions, $\|\hat{p}_X(\xi) - \hat{p}_{\varphi_\alpha(Y)}(\xi)\|$, that translates to bounds in MMD between the initial and the denoised distribution, under regularity assumptions on p_X. Finally in section 3.3, we prove similar bounds in Wasserstein distance, by making a link between half-denoising and one-step discretization of the diffusion ODE (Song et al., 2021b).

3.1 Gaussian Variables

If X is a multivariate Gaussian distribution, we can diagonalize the covariance matrix of X so that up to a rotation and a translation, $X \sim \mathcal{N}(0, \mathrm{diag}(\tau_1^2, \ldots, \tau_n^2))$. Both the denoising and the W_2-distance can then be calculated coordinate by coordinate, so we can focus on studying Gaussian variables in $\mathbb{R}$.

If $X \sim \mathcal{N}(0, \tau^2)$, then $Y \sim \mathcal{N}(0, \tau^2 + \sigma^2)$ and we can compute $\nabla \log p_Y(y) = \frac{-y}{\tau^2 + \sigma^2}$. It leads to

$$\varphi_\alpha(y) = \frac{\tau^2 + (1-\alpha)\sigma^2}{\tau^2 + \sigma^2} y,$$

in particular,

$$\varphi_1(y) = \frac{\tau^2}{\tau^2 + \sigma^2} y, \quad \varphi_{1/2}(y) = \frac{\tau^2 + (1/2)\sigma^2}{\tau^2 + \sigma^2} y.$$

The denoisers are linear transformations of Y, and their laws are Gaussian, given by

$$\varphi_\alpha(Y) \sim \mathcal{N}\left(0, \frac{(\tau^2 + (1-\alpha)\sigma^2)^2}{\tau^2+\sigma^2}\right),$$
$$\varphi_1(Y) \sim \mathcal{N}\left(0, \frac{\tau^4}{\tau^2+\sigma^2}\right),$$
$$\varphi_{1/2}(Y) \sim \mathcal{N}\left(0, \frac{(\tau^2 + (1/2)\sigma^2)^2}{\tau^2+\sigma^2}\right).$$

For two Gaussian variables $Z_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Z_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, the Wasserstein distance is given by (see, e.g., Peyré and Cuturi, 2019)

$$W_2^2(\mathcal{L}(Z_1), \mathcal{L}(Z_2)) = (\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2.$$

We can therefore compute directly the Wasserstein distances,

$$\begin{aligned}
W_2(\mathcal{L}(X), \mathcal{L}(\varphi_\alpha(Y))) &= \left|\tau - \tfrac{\tau^2+(1-\alpha)\sigma^2}{\sqrt{\tau^2+\sigma^2}}\right| &&= \tau\left|1 - \tfrac{1+(1-\alpha)\left(\frac{\sigma}{\tau}\right)^2}{\sqrt{1+\left(\frac{\sigma}{\tau}\right)^2}}\right|,\\
W_2(\mathcal{L}(X), \mathcal{L}(\varphi_1(Y))) &= \left|\tau - \tfrac{\tau^2}{\sqrt{\tau^2+\sigma^2}}\right| &&= \tau\left|1 - \tfrac{1}{\sqrt{1+\left(\frac{\sigma}{\tau}\right)^2}}\right|,\\
W_2(\mathcal{L}(X), \mathcal{L}(\varphi_{1/2}(Y))) &= \left|\tau - \tfrac{\tau^2+\frac{1}{2}\sigma^2}{\sqrt{\tau^2+\sigma^2}}\right| &&= \tau\left|1 - \tfrac{1+\frac{1}{2}\left(\frac{\sigma}{\tau}\right)^2}{\sqrt{1+\left(\frac{\sigma}{\tau}\right)^2}}\right|.
\end{aligned} \tag{2}$$

In particular for $\frac{\sigma}{\tau} \ll 1$ (small noise), with an expansion in $\frac{\sigma}{\tau}$, we get that

$$W_2(\mathcal{L}(X), \mathcal{L}(\varphi_1(Y))) \sim \frac{1}{2\tau}\sigma^2 \text{ and } W_2(\mathcal{L}(X), \mathcal{L}(\varphi_{1/2}(Y))) \sim \frac{1}{8\tau^3}\sigma^4,$$

showing that half-denoising beats full-denoising in Wasserstein distance for small noises. In fact, when dividing by τ the expression in (2), we see that they only depend on the ratio $\frac{\sigma}{\tau}$. These quantities are plotted in Figure 1, where we observe the behavior in $\left(\frac{\sigma}{\tau}\right)^2$ for full-denoisng and in $\left(\frac{\sigma}{\tau}\right)^4$ for half-denoising, making half-denoising better for small noises, and we remark that it stays better up to $\frac{\sigma}{\tau} = \sqrt{8} \approx 2.83$.

Note, moreover, that for $\alpha = 0$ (no denoising), we have

$$W_2(\mathcal{L}(X), \mathcal{L}(\varphi_0(Y))) = W_2(\mathcal{L}(X), \mathcal{L}(Y)) = \tau\left|1 - \sqrt{1+\left(\frac{\sigma}{\tau}\right)^2}\right| \approx \frac{1}{2\tau}\sigma^2,$$

that is, full-denoising is not better than no denoising at all!

On the contrary, for $\frac{\sigma}{\tau} \gg 1$ (large noise), we have

$$W_2(\mathcal{L}(X), \mathcal{L}(\varphi_1(Y))) \approx \tau \text{ and } W_2(\mathcal{L}(X), \mathcal{L}(\varphi_{1/2}(Y))) \approx \frac{1}{2}\sigma,$$

meaning that $W_2(\mathcal{L}(X), \mathcal{L}(\varphi_1(Y))) \ll W_2(\mathcal{L}(X), \mathcal{L}(\varphi_{1/2}(Y)))$. In fact $W_2(\mathcal{L}(X), \mathcal{L}(\varphi_1(Y))) = 0$ for a Dirac measure ($\tau = 0$) whereas $W_2(\mathcal{L}(X), \mathcal{L}(\varphi_{1/2}(Y))) = \frac{1}{2}\sigma$.

Finally, note that we can compute the optimal α for any τ, as $\alpha = 1 + \frac{\tau^2}{\sigma^2} - \frac{\tau}{\sigma}(1 + \frac{\tau^2}{\sigma^2})^{1/2} = \frac{\tau^2}{\sigma^2}(1 + \frac{\sigma^2}{\tau^2} - (1+\frac{\sigma^2}{\tau^2})^{1/2})$. For $\frac{\sigma}{\tau} \ll 1$, we get $\alpha = 1/2 + O((\frac{\sigma}{\tau})^2)$ and for $\frac{\sigma}{\tau} \gg 1$, $\alpha = 1 - \frac{\tau}{\sigma} + O((\frac{\tau}{\sigma})^2)$. We see that the first order terms do not depend on τ, and corresponds either to half-denoising or to full-denoising.

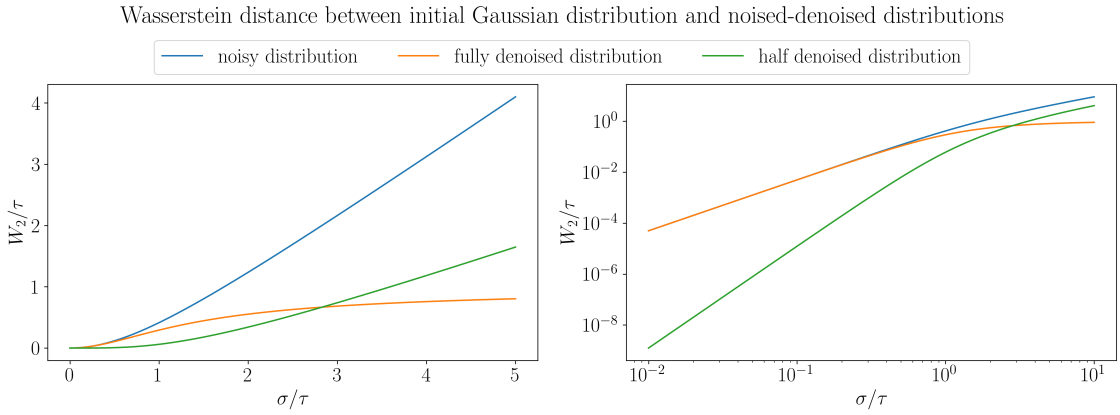


Figure 1: Wasserstein distances for Gaussian distributions at different levels of noise. Left: linear scale, right: logarithmic scale.

3.2 Half-Denoising is Better in MMD for Variables with Smooth Densities

Hyvärinen (2025) shows that $\hat{p}_{\varphi_{1/2}(Y)}(\xi) = \hat{p}_X(\xi) + O(\sigma^4)$. The proof relies on an uniform bound of the score function $\sup_{x\in\mathbb{R}^d} \|\nabla \log p_X(x)\| < +\infty$ and does not keep track of all constants. Based on the same idea, we propose a more complete statement, under a more general L_2 assumption, and an application to MMD.

Proposition 1 *Assume that* $\mathbb{E}[\|\nabla \log p_X(X)\|^2] \leq C$. *Then for all* $\alpha \in \mathbb{R}$,

$$\forall \xi \in \mathbb{R}^n, \|\hat{p}_X(\xi) - \hat{p}_{\varphi_\alpha(Y)}(\xi)\| \leq \frac{\sigma^2(2\alpha\sqrt{C}\|\xi\| + \|\xi\|^2)}{2},$$

and furthermore, for $\alpha = \frac{1}{2}$,

$$\forall \xi \in \mathbb{R}^n, \|\hat{p}_X(\xi) - \hat{p}_{\varphi_{1/2}(Y)}(\xi)\| \leq \frac{\sigma^4(C\|\xi\|^2 + \|\xi\|^4)}{8}.$$

(All proofs can be found in Appendix B.)

Remark: The hypothesis on the score ($\mathbb{E}[\|\nabla \log p_X(X)\|^2] \leq C$) [3] is natural in the framework of *denoising score matching*, where we use a neural network to learn the score $\log p_Y$. Indeed, Vincent (2011) showed the learning the score with an L_2-error $\min_\theta \mathbb{E}[\|s_\theta(Y) - \nabla \log p_Y(Y)\|^2]$ is equivalent to the denoising objective $\mathbb{E}[\|X - f_\theta(Y)\|^2]$, the latter being used in practice to learn the score. Therefore, as we are learning with an L_2-error, it is natural to ask for the L_2-bound $\mathbb{E}[\|\nabla \log p_Y(Y)\|^2] \leq C$. Imposing the bound on p_X, $\mathbb{E}[\|\nabla \log p_X(X)\|^2] \leq C$, allows us to have the bound on p_Y regardless of the level of

3. We can found a similar hypothesis in Assumption 2.5 of Albergo et al. (2023). Note that it's also nearly identical to the hypothesis H2 of Conforti et al. (2025), the difference being that the authors take the density with respect to the Gaussian measure rather than the Lebesgue measure. H2 implies $\mathbb{E}[\|\nabla \log p_X(X)\|^2] \leq C$, but the converse is not true in general.

the added noise σ as $\mathbb{E}[\|\nabla \log p_X(X)\|^2] \leq \mathbb{E}[\|\nabla \log p_Y(Y)\|^2]$ (Lemma 10 in Appendix B). It can also be deduced from other hypotheses made in the literature. For example:

- If $X = Z + \varepsilon_0$, with $Z \perp \varepsilon_0$ and $\varepsilon_0 \sim \mathcal{N}(0, \tau^2)$, and $\mathbb{E}[\|Z\|^2] \leq R^2$ (in particular if $\mathrm{supp}(Z) \subset B(0, R)$ as assumed in Theorem 1 of Saremi et al. (2023)), we have $\mathbb{E}[\|\nabla \log p_X(X)\|^2] \leq \frac{3(2R^2+d\tau^2)}{\tau^4}$. Indeed, with Tweedie's formula (1),

$$\begin{aligned}
\mathbb{E}[\|\nabla \log p_X(X)\|^2] &= \mathbb{E}\left[\left\|\frac{1}{\tau^2}(\mathbb{E}[Z|X] - X)\right\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{\tau^2}(\mathbb{E}[Z|X] - Z - \varepsilon_0)\right\|^2\right] \\
&\leq \frac{3}{\tau^4}(\mathbb{E}[\|\mathbb{E}[Z|X]\|^2] + \mathbb{E}[\|Z\|^2] + \mathbb{E}[\|\varepsilon_0\|^2]) \quad \text{(Jensen's inequality)} \\
&\leq \frac{3}{\tau^4}(\mathbb{E}[\mathbb{E}[\|Z\|^2|X]] + \mathbb{E}[\|Z\|^2] + d\tau^2) \\
&\quad \text{(Jensen's inequality on conditional expectation)} \\
&= \frac{3}{\tau^4}(2\mathbb{E}[\|Z\|^2] + d\tau^2) \leq \frac{3}{\tau^4}(2R^2 + d\tau^2).
\end{aligned}$$

- If X is such that $x \mapsto \nabla \log p_X(x)$ is L-Lipschitz and $\mathbb{E}[\|X\|^2] = m_2 < \infty$ (cf. assumptions A1 and A2 of Chen et al. (2023)), we have $\mathbb{E}[\|\nabla \log p_X(X)\|^2] \leq 2(L^2 m_2 + \|\nabla \log p_X(0)\|^2)$.

These bounds on the characteristic functions lead to bounds in MMD (Gretton et al., 2012) for a translation-invariant kernel. We assume that we are given a kernel k to compute a MMD, and that k is translation-invariant, i.e., $k(x, y) = \psi(x - y)$, with $\psi : \mathbb{R}^n \to \mathbb{R}$ a bounded, continuous positive definite function. From Bochner's theorem, there is a unique finite nonnegative Borel measure Λ on $\mathbb{R}^n$ such that

$$\psi(x) = \int e^{-i\xi \cdot x} d\Lambda(\xi).$$

Then, for two random variables X and Y, we have (Sriperumbudur et al., 2010, Corollary 4)

$$\mathrm{MMD}_k(\mathcal{L}(X), \mathcal{L}(Y)) = \left(\int |\hat{p}_X(\xi) - \hat{p}_Y(\xi)|^2 d\Lambda(\xi)\right)^{1/2}.$$

Corollary 2 *Assume that,* $\mathbb{E}[\|\nabla \log p_X(X)\|^2] \leq C$, *and,*

$$C_2 = \int \|\xi\|^2 d\Lambda(\xi) < \infty, \qquad C_4 = \int \|\xi\|^4 d\Lambda(\xi) < \infty, \quad \textit{and} \quad C_8 = \int \|\xi\|^8 d\Lambda(\xi) < \infty.$$

Then, for all $\alpha \in \mathbb{R}$,

$$\mathrm{MMD}_k(\mathcal{L}(X), \mathcal{L}(\varphi_\alpha(Y))) \leq K_1 \sigma^2,$$

with $K_1 = \frac{\sqrt{4\alpha^2 C C_2 + C_4}}{\sqrt{2}}$.

Furthermore, for $\alpha = \frac{1}{2}$,

$$\mathrm{MMD}_k(\mathcal{L}(X), \mathcal{L}(\varphi_{1/2}(Y))) \leq K_2 \sigma^4,$$

with $K_2 = \frac{\sqrt{C^2 C_4 + C_8}}{4\sqrt{2}}$.

The first result applies for $\alpha = 1$, hence we can compare the bound that we get for full-denoising (first result) to the one we get for half-denoising (second result). It shows that for regular enough densities ($\mathbb{E}[\|\nabla \log p_X(X)\|^2] \leq C$), we have $\mathrm{MMD}_k(\mathcal{L}(X), \mathcal{L}(\varphi_{1/2}(Y))) = O(\sigma^4)$, whereas $\mathrm{MMD}_k(\mathcal{L}(X), \mathcal{L}(\varphi_1(Y))) = O(\sigma^2)$ and hence the bound on $\mathrm{MMD}_k(\mathcal{L}(X), \mathcal{L}(\varphi_{1/2}(Y)))$ is negligible compared to the bound on $\mathrm{MMD}_k(\mathcal{L}(X), \mathcal{L}(\varphi_1(Y)))$ for small σ, therefore extending the result seen above for Gaussian distributions. Moreover, the bound $\mathrm{MMD}_k(\mathcal{L}(X), \mathcal{L}(\varphi_\alpha(Y))) = O(\sigma^2)$ also applies for $\alpha = 0$, hence as in the Gaussian case, full-denoising does not do better than no-denoising for small σ.

3.3 Half-Denoising is Better in Wasserstein-2 for Variables with Smooth Densities

We now prove similar bounds in Wasserstein distance. To do so, we introduce a continuous diffusion process, progressively adding Gaussian noise to X with a Brownian motion, and the diffusion ODE, which generates the same marginals with a deterministic process. This deterministic process, sometimes referred to as the probability flow ODE (Song et al., 2021b), has been used as a way to have deterministic generation with diffusion models. Here, we will use the fact that half-denoising can be seen as a one step discretization of this ODE. We give a complete proof of the construction of this ODE in Appendix A. Here is a brief overview.

We define a process $X_t = X + B_t$, with B_t a Brownian motion, such that we have $X = X_0$ and $Y = X_{\sigma^2}$. We also denote $p_t = p_{X_t}$ the density of X_t with respect to the Lebesgue measure. p_t verifies the Fokker-Planck equation $\partial p_t = \Delta p_t$, which can be rewritten as $\partial p_t = -\nabla \cdot (-p_t \nabla \log p_t)$. We deduce that we can then defined the following ODE:

$$\begin{cases} \frac{dx_t}{dt} &= -\frac{1}{2}\nabla \log p_t(x_t) \quad \text{for } t > 0 \\ x_{t^*} &= X_{t^*} \qquad\qquad\qquad\quad \text{for some } t^* > 0, \end{cases} \tag{3}$$

and that it has the same marginals as X_t: $\forall t \in [0, +\infty[, x_t \sim X_t$, and verifies, for all $t, s \geq 0$,

$$x_t - x_s = -\frac{1}{2}\int_s^t \nabla \log p_u(x_u) du.$$

Note here that $\nabla \log p_t$ is not, in general, Lipschitz-continuous near $t = 0$.[4] That's why we take an initial condition[5] at $t^* > 0$. However, we verify that the trajectory can in fact be integrated up to $t = 0$ (see Appendix A for more details).

As Gentiloni-Silveri and Ocello (2025),[6] we use the continuous time process x_t, and its one step discretization, to get natural couplings between distribution of X and $\varphi_\alpha(Y)$ and compute Wasserstein distances. We have $X \sim x_0$ and $\varphi_\alpha(Y) \sim \hat{x}_0$, with $\hat{x}_0 = x_t +$

4. This should not come as a surprise, as if we take μ_X to be a Dirac at 0, then all trajectories will coincide at $t = 0$, which is prohibited by Cauchy-Lipschitz Theorem.
5. The choice of t^* does not matter as for any $t^* > 0$, we will have the same marginals for $(x_t)_{t \geq 0}$, and in particular it will give a path between $x_{\sigma^2} \sim Y$ and $x_0 \sim X$.
6. Gentiloni-Silveri and Ocello (2025) compute Wasserstein distance for the diffusion SDE with a multiple step discretization.

$\alpha t \nabla \log p_t(x_t)$ and $t = \sigma^2$. This leads to the bound $W_2^2(\mathcal{L}(X), \mathcal{L}(\varphi_\alpha(Y))) \leq \mathbb{E}[\|x_0 - \hat{x}_0\|^2]$, with

$$x_0 - \hat{x}_0 = \frac{1}{2}\int_0^t \nabla \log p_s(x_s) ds - \alpha t \nabla \log p_t(x_t).$$

To conclude, we need to be able to bound $\nabla \log p_s(x_s)$ for $s \in [0, t]$, and to do that, we only need to have a bound on $\nabla \log p_X(X)$. Formally, we have the following result:

Proposition 3 *Assume that* $\mathbb{E}[\|\nabla \log p_X(X)\|^2] \leq C$. *Then*

$$W_2(\mathcal{L}(X), \mathcal{L}(\varphi_\alpha(Y))) \leq \sqrt{\frac{(1+4\alpha^2)C}{2}}\sigma^2.$$

For $\alpha = 1$, this bound is already better (in σ) than the bound given by Proposition 2 of Saremi et al. (2023) which is $W_2(\mathcal{L}(X), \mathcal{L}(\varphi_1(Y))) = O(\sigma)$. Moreover this bound is tight (in σ) as for a Gaussian variable with variance τ^2, we have $W_2(\mathcal{L}(X), \mathcal{L}(\varphi_1(Y))) \sim \frac{1}{2\tau}\sigma^2$ for small enough σ. Note that if we remove the assumption (i.e., $C = +\infty$), then we fall back on the bound by Saremi et al. (2023) (take for example X a Dirac measure, for which we have seen that $W_2(\mathcal{L}(X), \mathcal{L}(\varphi_{1/2}(Y))) = \frac{1}{2}\sigma$).

But for $\alpha = \frac{1}{2}$ we can hope to have a better bound, in σ^4, as it was the case with the MMD. Indeed, we have

$$\begin{aligned} x_0 - \hat{x}_0 &= \frac{1}{2}\int_0^t \nabla \log p_s(x_s) ds - \frac{1}{2} t \nabla \log p_t(x_t) \\ &= \frac{1}{2}\int_0^t \left(\nabla \log p_s(x_s) - \nabla \log p_t(x_t)\right) ds \\ &= \frac{1}{2}\int_0^t \int_s^t \frac{d}{du}\left(\nabla \log p_u(x_u)\right) du ds. \end{aligned}$$

If we can control $\mathbb{E}[\|\frac{d}{dt}\nabla \log p_t(x_t)\|^2] \leq C$, we will get $\mathbb{E}[\|x_0 - x_t\|^2] = O(t^4)$ hence $W_2(\mathcal{L}(X), \mathcal{L}(\varphi_{1/2}(Y))) = O(\sigma^4)$. Formally, we need the following lemma:

Lemma 4 *Assume that the variable* X *of density* p_X *satisfies:*

- $\log p_X \in \mathcal{C}^3(\mathbb{R}^d)$.
- $C_1 = \mathbb{E}[\|\nabla \log p_X(X)\|^6] < \infty$.
- $C_2 = \mathbb{E}[\|\nabla^2 \log p_X(X)\|_{\text{op}}^3] < \infty$, *where* $\|A\|_{\text{op}}$ *is defined for any matrix* A *as* $\|A\|_{\text{op}} = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}$.
- $C_3 = \mathbb{E}[\|\nabla \Delta \log p_X(X)\|^2] < \infty$.

Then,

$$\mathbb{E}\left[\left\|\frac{d}{dt}\nabla \log p_t(x_t)\right\|^2\right] \leq C,$$

with $C = \frac{9}{4}(4C_1 + (2d^2+5)C_1^{1/3}C_2^{2/3} + C_3)$.

Using this lemma, we have the following proposition:

Proposition 5 *Under the assumptions of Lemma 4, we have*

$$W_2(\mathcal{L}(X), \mathcal{L}(\varphi_{1/2}(Y))) \leq K\sigma^4.$$

with $K = \frac{\sqrt{3}}{4}\sqrt{4C_1 + (2d^2+5)C_1^{1/3}C_2^{2/3} + C_3}$.

Remarks:

- In Appendix C, we show that if $X = Z + \varepsilon_0$, with $\mathbb{E}[\|Z\|^6] < \infty$, $Z \perp \varepsilon_0$ and $\varepsilon_0 \sim \mathcal{N}(0, \tau^2)$, then X verifies the assumptions of Lemma 4. In particular, it applies to the case of $\text{supp}(Z) \subset B(0, R)$, assumed in Theorem 1 of Saremi et al. (2023), as then $\mathbb{E}[\|Z\|^6] \leq R^6$. It also proves that a mixture of Gaussian distributions verifies the hypothesis (take τ the smallest eigenvalue of all the covariances matrices of the Gaussian distributions in the mixture).

- The assumptions of Lemma 4 and Proposition 5 control the regularity of the density p_X. The fact that we need to bound derivative up to order 3 comes directly from the Fokker-Planck equation $\partial_t p_t = \Delta p_t$, which can be interpreted heuristically as "one derivative in t equals two derivatives in x", hence to control $\frac{d}{dt}\nabla \log p_t(x_t)$, we need to control derivatives in x up to order 3.

- The control of the Hessian $\|\nabla^2 \log p_X(x)\|_{\text{op}}$ corresponds to controlling the Lipschitz constant of the function $x \mapsto \nabla \log p_X(x)$ and is an assumption usually done in the literature (see, e.g., Chen et al., 2023, assumption A1). Having an uniform bound is a strong assumption, and in particular it implies that the distribution has full support on $\mathbb{R}^d$. But here we only need to control this quantity in expectation under the law of X. As a consequence, this result can apply to distribution such that $x \mapsto \nabla \log p_X(x)$ is not globally Lipschitz-continuous, and moreover it does not even have to be defined everywhere. Take for example

$$p_X(x) = \begin{cases} \frac{1}{Z} e^{-\frac{1}{1-x^2}} & \text{if } |x| \leq 1 \\ 0 & \text{otherwise,} \end{cases}$$

  where Z is a normalizing constant (Fig. 2). Then $\log p_X(x) = -\frac{1}{1-x^2}$ is only defined on the interval $(-1, 1)$ and we have

$$\nabla \log p_X(x) = -\frac{2x}{(1-x^2)^2},$$
$$\nabla^2 \log p_X(x) = -\frac{2\left(3x^2+1\right)}{(1-x^2)^3},$$
$$\nabla\Delta \log p_X(x) = -\frac{24x\left(x^2+1\right)}{(1-x^2)^4},$$

  with none of these quantities being bounded on $(-1, 1)$. However, our assumptions only require the bounds in expectation, and as the density p_X decreases exponentially

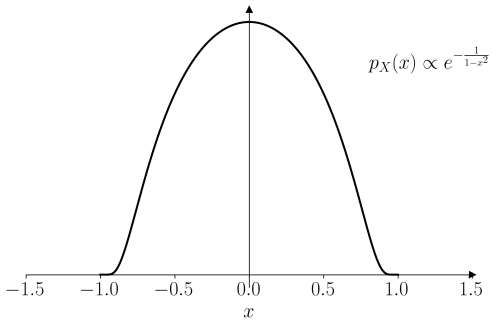


Figure 2: Example of a smooth density with a compact support.

fast when $|x| \to 1$, overcoming the polynomial growth of the derivatives, all three constants C_1, C_2 and C_3 are finite.

Note that as Propositions 3 and 5 can be applied to distributions with compact support, they can be combined with Proposition 7 of section 5.

- The assumptions are not verified for a very singular distribution, e.g., a Dirac $\mu_X = \delta_0$, for which $\log p_X$ is not even defined. In this case, we have $x_t \sim \mathcal{N}(0, tI)$ hence $\log p_t(x) = -\frac{\|x\|^2}{2t}$, $\nabla \log p_t(x) = -\frac{x}{t}$ and $\mathbb{E}[\|\nabla \log p_t(X)\|^2_{\text{op}}] = \frac{d}{t}$, which is not bounded, and not integrable near 0. The result of Proposition 5 don't apply, and indeed we have (cf. section 3.1) $W_2(\mathcal{L}(X), \mathcal{L}(\varphi_{1/2}(Y))) = \frac{\sigma}{2}$ whereas $W_2(\mathcal{L}(X), \mathcal{L}(\varphi_1(Y))) = 0$.
- All results from this subsection can be extended to Wasserstein-p distance for any $p \geq 1$ (see Appendix D).

4. Variable with Support on a Subspace: Balancing Between the Impacts of Singularity and Regularity of the Distribution

In many practical applications, the data distribution is supported on a lower dimensional manifold, a case known as the *manifold hypothesis* (see, e.g., Tenenbaum et al., 2000; Bengio et al., 2013; Fefferman et al., 2016). Locally, this manifold will look like a linear space, therefore, we take a look at the idealized case when the variable is supported on a *linear* lower-dimensional subspace.

Proposition 6 *Assume that X is supported on a linear subspace H of dimension m, with $m \leq d$. Write $X_1 = p_H(X)$, with p_H the orthogonal projection on H, and $Y_1 = X_1 + \varepsilon_1 \in H$ with $\varepsilon_1 \sim \mathcal{N}(0, \sigma^2 I_m) \in H$. Then:*

$$W_2^2(\mathcal{L}(X), \mathcal{L}(\varphi_\alpha(Y))) = W_2^2(\mathcal{L}(X_1), \mathcal{L}(\varphi_\alpha(Y_1))) + (d-m)(1-\alpha)^2\sigma^2.$$

We can interpret this result as trade-off between $\alpha = 1$ (full-denoising) which cancels the second term, as it ensures that $\varphi_\alpha(Y)$ belongs to the subspace H (see the proof in

Appendix B for more details), and $\alpha = 1/2$ (half-denoising) that will reduce the first term if p_{X_1} is regular enough to apply the results of the previous section. More precisely, under the assumption that the density on the subspace is regular enough, $W_2^2(\mathcal{L}(X_1), \mathcal{L}(\varphi_\alpha(Y_1)))$ is in $O(\sigma^4)$ for full-denoising (Proposition 3), and in $O(\sigma^8)$ for half-denoising (Proposition 5). Half-denoising is better on the subspace for σ small enough, but the term $\frac{(d-m)\sigma^2}{4}$ dominates as σ goes to zero. Depending on the ratio between the dimension m of the subspace and d of the whole space, there may or may not be a sweet spot for half-denoising where the gain obtained by reducing the first term outweighs the increase in the second term.

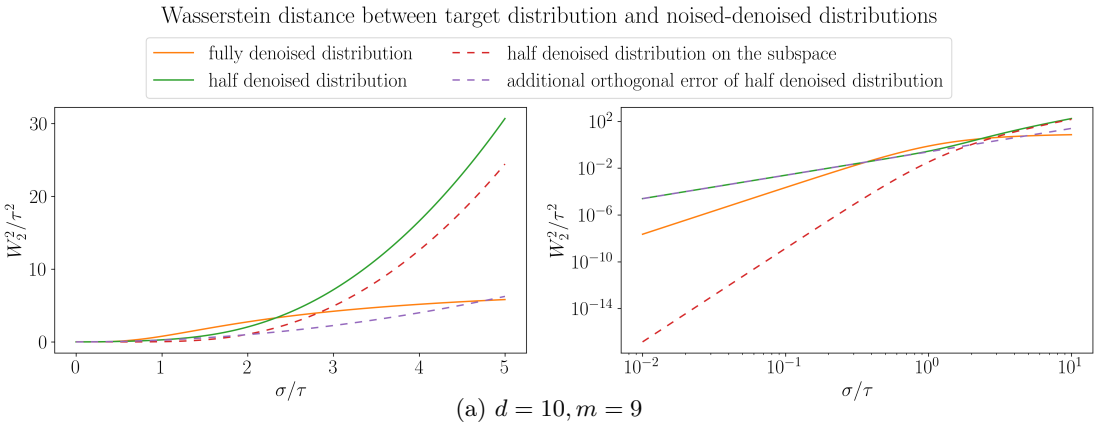
We illustrated this in Figure 3, where the target distribution is a Gaussian $\mathcal{N}(0, \tau^2 I_m)$ supported on the subspace $\mathbb{R}^m \times \{0\}^{d-m}$ (we use the closed-form formulas from Section 3.1). For full-denoising, the only term that remains is $W_2^2(\mathcal{L}(X_1), \mathcal{L}(\varphi_\alpha(Y_1)))$, that is plotted in orange. Half-denoising is plotted in green, and is the sum of the term $W_2^2(\mathcal{L}(X_1), \mathcal{L}(\varphi_\alpha(Y_1)))$ (red dotted line) and the term $\frac{(d-m)\sigma^2}{4}$ (purple dotted line). The term corresponding to half-denoising on the subspace (red dotted line) is smaller that the Wasserstein distance for full-denoising (orange line) for $\sigma \lesssim 2.83\tau$. However, we observe that as σ goes to zero, the term $\frac{(d-m)\sigma^2}{4}$ (purple dotted line) dominates hence the Wasserstein distance for half-denoising (green line) is greater that the error for full-denoising. For $d = 10, m = 9$ (a), there is a sweet spot in which the trade-off is in favor of half-denoising, while for $d = 10, m = 5$ (b), full-denoising is better at all noise levels.

Proposition 6 also shows that full-denoising is adaptive to the low dimensional structure of the data, as the Wasserstein distance only depends on the distance between distributions on the lower-dimensional subspace. Therefore, in the case where $m \ll d$, it alleviates the curse of dimensionality.

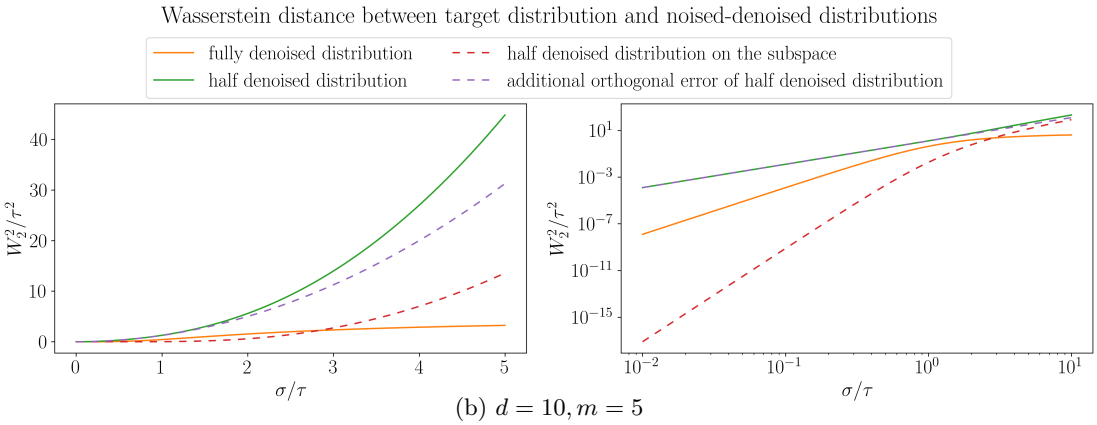
5. Mixtures of Distributions with Disjoint Compact Supports Behave like Independent Variables

In the case of the *manifold hypothesis*, it is also common to suppose that the distribution is made of small pockets with high density of probability, representing different classes of objects, separated by regions of low density. We model this case be saying that the distribution of X is a mixture of distributions with disjoint compact supports. In this case, we show that the denoising performance for a mixture of distributions with disjoint compact supports behaves as if we were denoising each variable independently, plus an exponentially decreasing term.

More formally, let $X \sim \mu_X = \sum_{i=1}^N \pi_i \mu_i$ (with $\sum_{i=1}^N \pi_i = 1, \pi_i \geq 0$) a mixture of distribution μ_i with compact support S_i such that $D = \min_{i \neq j} d(S_i, S_j) > 0$ (with $d(S_i, S_j) = \min_{x_i \in S_i, x_j \in S_j} \|x_i - x_j\|$). We denote $Y = X + \varepsilon$ with $X \perp \varepsilon$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, for $\sigma > 0$. For $\alpha \in \mathbb{R}$, we denote $\varphi_\alpha(y) = y + \alpha\sigma^2 \nabla \log p_Y(y)$, ν the law of Y and μ_α the law of $\varphi_\alpha(Y)$. Similarly, for $X_i \sim \mu_i$, and $Y_i = X_i + \varepsilon$ with $X_i \perp \varepsilon$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, we denote $\varphi_{i,\alpha}(y) = y + \alpha\sigma^2 \nabla \log p_{Y_i}(y)$, ν_i the law of Y_i and $\mu_{i,\alpha}$ the law of $\varphi_{i,\alpha}(Y_i)$. We have the following proposition:



(a) $d = 10, m = 9$



(b) $d = 10, m = 5$

Figure 3: Wasserstein distances for Gaussian distribution supported on the subspace $\mathbb{R}^m \times \{0\}^{d-m}$ as target distribution and different noise levels σ.

Proposition 7 *We have, under the above assumptions,*

$$W_2^2(\mu_X, \mu_\alpha) \leq 2 \sum_i \pi_i W_2^2(\mu_i, \mu_{i,\alpha}) + O\left(\frac{1}{\sigma^{\max(d-2,8)}} \exp\left(-\frac{K}{\sigma^2}\right)\right),$$

where K is a constant that depends only on D.

The O hides a constant depending on α, d, N and R such that $\text{supp}(X) \subset B(0, R)$ (see the proof in Appendix B for more details).

In the case where the μ_i's are Dirac measures, and for $\alpha = 1$, we have $W_2^2(\mu_i, \mu_{i,\alpha}) = 0$, hence

$$W_2^2(\mu_X, \mu_\alpha) = O\left(\frac{1}{\sigma^{\max(d-2,8)}} \exp\left(-\frac{K}{\sigma^2}\right)\right),$$

which is way better than polynomial rates in σ seen above.

In Figure 4, we illustrate this result for a mixture of two Dirac measures in 1D, with $X = \frac{\delta_{-\mu} + \delta_\mu}{2}$ for some $\mu > 0$. In this case, we can derive integral expressions for

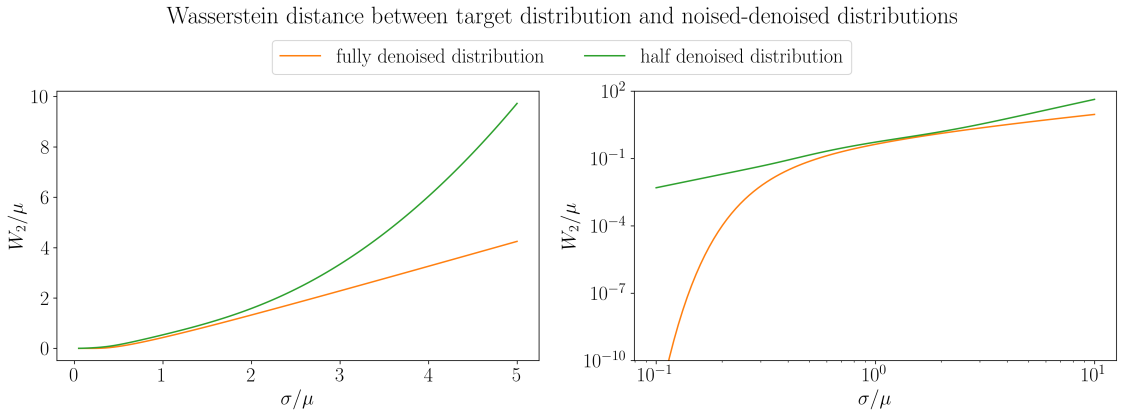


Figure 4: Wasserstein distances for a mixture of two Dirac measures $\frac{\delta_{-\mu}+\delta_{\mu}}{2}$ as target distribution and different noise levels σ.

$\frac{W_2(\mathcal{L}(X),\mathcal{L}(\varphi_\alpha(Y)))}{\mu}$ as functions of $\frac{\sigma}{\mu}$ that can be evaluated numerically (see Appendix E). For full-denoising, the only remaining term is the $O\left(\frac{1}{\sigma^{\max(d-2,8)}}\exp\left(-\frac{K}{\sigma^2}\right)\right)$, which decreases much faster than the error for half-denoising.

6. Consequences

In this section, we examine the consequences of our results for methods based on one-step denoising as well as for multi-step diffusion models.

6.1 Linear Manifold Hypothesis

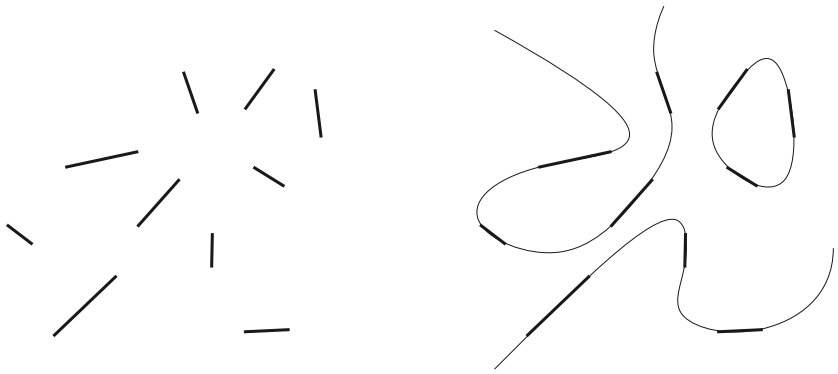
We can combine Propositions 6 and 7 to tackle what we call the *linear manifold hypothesis*. We defined the linear manifold hypothesis as a simplified version of the manifold hypothesis, where the data distribution is supported on disjoint compact sets, each of these belonging to a (different) linear subspace of low dimension, as illustrated in Figure 5. Then applying Proposition 7 allows to bound the Wasserstein distance between the original distribution and the fully denoised distribution by the sum of the Wasserstein distances between the distributions on each compact sets (plus on exponentially decreasing term). As each compact set belongs to a low-dimensional subspace, we can apply Proposition 6, which tells that for full-denoising, the Wasserstein distance depends only on the distribution on the sub-space. In this case, we see that full-denoising can alleviate the curse of dimensionality even if the support of the mixture distribution itself spans the whole space as it adapts to the local linear structure of the distribution. More generally, understanding the performance of score-based generative models for data distribution supported on a low dimensional manifold is an active area of research (see, e.g., Tang and Yang, 2024; Azangulov et al., 2025).

In particular, a result from Azangulov et al. (2025) also gives a bound for full-denoising that only depends on the manifold dimension under the manifold hypothesis. For a smooth manifold M of dimension $d_M < d$ and a regular enough density p_X (see the original paper

for all assumptions), we can deduce from their Theorem 12 that

$$W_2(\mathcal{L}(X), \mathcal{L}(\varphi_1(Y))) \leq \sigma\sqrt{8(40d_M \log_+ \sigma^{-1} + 8d_M C_{\log} + 3)} \tag{4}$$

where $\log_+ : x \mapsto \max(\log x, 1)$ and $C_{\log}$ is a constant that controls the regularity of p_X, in particular, it must verify $\forall x \in M, e^{-d_M C_{\log}} < p_X(x) < e^{d_M C_{\log}}$. This bound shows that the Wasserstein distance between the fully denoised distribution and the original distribution only depends on the subspace dimension. Compared to Propositions 6 and 7, it allows to tackle the more complete case of a smooth manifold, however it relies on stronger smoothness assumptions, with in particular the need for a lower and upper bounded density $e^{-d_M C_{\log}} < p_X(x) < e^{d_M C_{\log}}$, while our results do not not need any regularity assumptions on p_X. Moreover, we see that (4) scales as $O(\sigma)$, whereas Proposition 3 gives a scaling in $O(\sigma^2)$ is the density is regular enough. Future work could be dedicated to seeing whether we can draw inspiration from their approach and ours to achieve a more general result with a scaling in $O(\sigma^2)$.



(a) (b)

Figure 5: Illustration of the linear manifold hypothesis (a – thick line), and the manifold hypothesis (b – fine line).

6.2 Half vs Full for One-Step Denoising

Our results give insights to choose between half and full denoising in algorithms that use one step denoising. This includes the algorithm proposed by Hyvärinen (2025), which can be viewed (for $\mu = \frac{\sigma^2}{2}$) as a regular Langevin on the noisy variable Y followed by one step of half-denoising, *walk-jump sampling* (Saremi and Hyvärinen, 2019), which is a regular Langevin on the noisy variable Y followed one step of full-denoising, or also *Sequential multimeasurement walk-jump sampling* (Saremi et al., 2023), where noise is first reduced by averaging multiple noisy measurements then one step of full-denoising is applied. While in the algorithms listed here, the choice of half or full denoising is arbitrary, our results provide guidance on choosing one over the other depending on practical cases. If the density is regular, then half-denoising is preferable, whereas if it is singular, for example in the case of the manifold hypothesis, full-denoising is preferable.

6.3 Denoising Diffusion Models

We will briefly discuss the first insights that our results on one step denoising give about multi-step diffusion models, leaving a further analysis for future work. We will focus on diffusion models with a deterministic denoising process, such as the probability flow ODE by Song et al. (2021b) and DDIM by Song et al. (2021a). We start from the more general formulation of the probability flow ODE given by Karras et al. (2022):

$$\frac{dx_t}{dt} = \frac{\dot{s}(t)}{s(t)} x_t - s(t)\dot{\sigma}(t)\sigma(t)\nabla \log p\left(\frac{x_t}{s(t)}; \sigma(t)^2\right), \tag{5}$$

where $p\left(x;\sigma^2\right)$ is the density [7] of the variable $X + \mathcal{N}(0,\sigma^2)$, and x_t as marginal $x_t \sim s(t)\cdot(X + \mathcal{N}(0,\sigma(t)^2))$.

From this equation, we get algorithms by choosing some time T, sampling $\hat{X}_0 \sim \mathcal{N}(0, s(t)^2\sigma(t)^2) \approx x_T$, and discretizing the process given by (5) with N (possibly non uniform) steps $t_0 = T > t_1 > \cdots > t_N = 0$.

There are multiple ways to discretize (5), leading to different algorithms (see Appendix F for more details on the derivations). The DDIM algorithm (Song et al., 2021a) corresponds to the update

$$\hat{X}_{k+1} = \frac{s(t_{k+1})}{s(t_k)}\hat{X}_k + s(t_{k+1})(\sigma(t_k)^2 - \sigma(t_k)\sigma(t_{k+1}))\nabla \log p\left(\frac{\hat{X}_k}{s(t_k)}; \sigma(t_k)^2\right).$$

Equivalently,[8] using the rescaled variable $\widetilde{X}_k = \hat{X}_k / s(t_k)$, we have

$$\widetilde{X}_{k+1} = \widetilde{X}_k + \alpha_k\left(\sigma(t_k)^2 - \sigma(t_{k+1})^2\right)\nabla \log p\left(\widetilde{X}_k; \sigma(t_k)^2\right),$$

with $\alpha_k = \frac{\sigma(t_k)}{\sigma(t_{k+1})+\sigma(t_k)}$. This update corresponds to α_k-denoising, between noise levels $\sigma(t_k)$ and $\sigma(t_{k+1})$.

On the other hand, the Euler discretization of (5) gives

$$\hat{X}_{k+1} = \left(1 + \frac{\dot{s}(t_k)(t_{k+1}-t_k)}{s(t_k)}\right)\hat{X}_k - s(t_k)(t_{k+1}-t_k)\dot{\sigma}(t_k)\sigma(t_k)\nabla \log p\left(\frac{\hat{X}_k}{s(t_k)}; \sigma(t_k)^2\right),$$

and with the rescaled variable $\widetilde{X}_k = \hat{X}_k / s(t_k)$, we have

$$\widetilde{X}_{k+1} = \frac{s(t_k) + \dot{s}(t_k)(t_{k+1}-t_k)}{s(t_{k+1})}\widetilde{X}_k - \frac{s(t_k)}{s(t_{k+1})}(t_{k+1}-t_k)\dot{\sigma}(t_k)\sigma(t_k)\nabla \log p\left(\widetilde{X}_k; \sigma(t_k)^2\right). \tag{6}$$

7. We differ slightly from the notation of Karras et al. (2022) as we parametrize this density by σ^2 rather than σ, to be more coherent with the usual notation for Gaussian distributions and such that we have $p_t(x) = p(x;t)$ for the process of (3).

8. From a theoretical perspective, looking at $\hat{X}_k$ and $\widetilde{X}_k$ is strictly equivalent as we can go from one to the other simply by multiplying by a known quantity. Note however that this scaling can have consequences on the training and numerical stability of the algorithms (but those effects can be dissociated by adding multiplicative constants to modify how our neural network and training loss are parametrized, as done by Karras et al. (2022)).

This cannot be directly interpreted as α-denoising, but for $s(t) = 1$ constant, then the update becomes

$$\widetilde{X}_{k+1} = \widetilde{X}_k + \alpha_k \left(\sigma(t_k)^2 - \sigma(t_{k+1})^2\right) \nabla \log p\left(\widetilde{X}_k; \sigma(t_k)^2\right),$$

with $\alpha_k = \frac{-(t_{k+1}-t_k)\dot{\sigma}(t_k)\sigma(t_k)}{\sigma(t_k)^2-\sigma(t_{k+1})^2}$. Here, we have steps of α-denoising, but the coefficient α_k does not only depend on the noise levels $(\sigma(t_k))_k$, but also on the specific choice of noise parametrization $t \mapsto \sigma(t)$. Indeed, taking $\sigma(t) = \sqrt{t}$ (which corresponds to the model introduced in Section 3.3) gives

$$\alpha_k = \frac{-(\sigma(t_{k+1})^2 - \sigma(t_k)^2)\frac{1}{2\sigma(t_k)}\sigma(t_k)}{\sigma(t_k)^2 - \sigma(t_{k+1})^2} = \frac{1}{2},$$

while taking $\sigma(t) = t$ (as Karras et al., 2022), leads to

$$\alpha_k = \frac{-(\sigma(t_{k+1}) - \sigma(t_k))\sigma(t_k)}{\sigma(t_k)^2 - \sigma(t_{k+1})^2} = \frac{\sigma(t_k)}{\sigma(t_{k+1}) + \sigma(t_k)},$$

which is exactly the same as the DDIM update.

For the parametrization $s(t) = 1 - t$ and $\sigma(t) = \frac{t}{1-t}$ used in flow matching (Liu et al., 2023; Lipman et al., 2023; Albergo et al., 2023), the update (6) can also be interpreted as α_k-denoising. Indeed, we have

$$\frac{s(t_k) + \dot{s}(t_k)(t_{k+1} - t_k)}{s(t_{k+1})} = \frac{1 - t_k - (t_{k+1} - t_k)}{1 - t_{k+1}} = 1,$$

and with the identities $t = \frac{\sigma(t)}{1+\sigma(t)}$, $s(t) = \frac{1}{1+\sigma(t)}$, and $\dot{\sigma}(t) = \frac{1}{(1-t)^2} = (1+\sigma(t))^2$, we get that

$$\begin{aligned}
\alpha_k &= \frac{s(t_k)}{s(t_{k+1})} \frac{-(t_{k+1} - t_k)\dot{\sigma}(t_k)\sigma(t_k)}{\sigma(t_k)^2 - \sigma(t_{k+1})^2} \\
&= \frac{1 + \sigma(t_{k+1})}{1 + \sigma(t_k)} \frac{\left(\frac{\sigma(t_k)}{1+\sigma(t_k)} - \frac{\sigma(t_{k+1})}{1+\sigma(t_{k+1})}\right)(1 + \sigma(t_k))^2\sigma(t_k)}{(\sigma(t_k) + \sigma(t_{k+1}))(\sigma(t_k) - \sigma(t_{k+1}))} \\
&= \frac{\sigma(t_k)}{\sigma(t_{k+1}) + \sigma(t_k)},
\end{aligned}$$

which is exactly the same as the DDIM update.

We can interpret these different coefficients α_k in the light of our theoretical results. At the beginning of the reverse process, one is going from a noisy distribution to a sightly less noisy one. As noising regularizes the density, our finding shows that half-denoising is better. For the Euler discretization with noise schedule $\sigma(t) = \sqrt{t}$, we do have $\alpha_k = \frac{1}{2}$. For DDIM—and equivalently the Euler discretization with parametrization $s(t) = 1$, $\sigma(t) = t$ or $s(t) = 1 - t$, $\sigma(t) = t/(1-t)$—if we assume that $\sigma(t_{k+1}) = \sigma(t_k) - \Delta\sigma$ with $\Delta\sigma \ll \sigma(t_k)$, then we have $\alpha_k \approx \frac{1}{2}$.

On the contrary, at the end of the reverse process, one is going from a noisy distribution to the target distribution, therefore the choice of the coefficient α should depend on what we know about the target density. For the Euler discretization with noise schedule $\sigma(t) =$

$\sqrt{t}$, as the coefficients α_k are constant equal to $\frac{1}{2}$, our theoretical results suggest that this choice is best for a regular target density. For DDIM, or Euler with $s(t) = 1$, $\sigma(t) = t$ or $s(t) = 1 - t$, $\sigma(t) = t/(1 - t)$, then α_k gradually shifts to

$$\alpha_{N-1} = \frac{\sigma(t_{N-1})}{\sigma(t_{N-1}) + \sigma(t_N)} = \frac{\sigma(t_{N-1})}{\sigma(t_{N-1}) + 0} = 1,$$

which corresponds to full-denoising.[9] Our results suggest that this choice is best suited for a singular density, for example under the *manifold hypothesis*.

When interpreting each step as α-denoising, it appears that these design choices lead to different algorithms that may be more or less suited to certain target distributions. We believe that it would be interesting in future works to study whether it is possible to choose directly different time schedules for the coefficient α. More generally, as the optimal α depends on the properties of the data distribution, it would be interesting to see if it could be fine-tuned in a data-dependent way.

7. Conclusion

We have shown that half-denoising is better than full-denoising for regular enough densities, while full-denoising is better for singular densities such as mixtures of Dirac measures or Gaussian with small variance compare to the additional noise. Moreover, the performance of the denoisers can be further accessed with additional assumptions on the data distribution, that occur naturally in real-world data, for example with images under the *manifold hypothesis*.

When the variable is supported on a lower-dimensional subspace, we have shown that there is a trade-off between full-denoising which reduces the Wasserstein distance by ensuring that the output belongs to the subspace, and half-denoising that reduces the Wasserstein distance on the lower-dimensional subspace assuming a regular enough density. In the case where the subspace is of small enough dimension compared to the full space, full-denoising alleviates the curse of dimensionality as the Wasserstein distance only depends on the distance between distributions on the lower-dimensional subspace. Moreover, we have shown that the denoising performance for a mixture of distributions with disjoint compact supports behaves as if we were denoising each variable independently, plus an exponentially decreasing term. This led to a case we called *linear manifold hypothesis*, where the data distribution is supported on disjoint compact sets, each of these belonging to a (different) linear subspace of low dimension, and for which full-denoising can alleviate the curse of dimensionality even if the support of the distribution itself spans the whole space, as it adapts to the local linear structure of the distribution.

For algorithms using one step denoising (Saremi and Hyvärinen, 2019; Saremi et al., 2023; Hyvärinen, 2025), our results provide guidance on choosing between half-denoising (regular density) and full-denoising (singular target density, manifold hypothesis). Moreover, for multiple steps denoising models, we have shown that each step can be seen as α-denoising, with different α's depending on design choices. Our theoretical results therefore offer new insights into design choices based on assumptions about the data distribution.

9. Note that this is coherent with the fact that DDIM is exact if the data distribution is a Dirac (Nakkiran et al., 2024), a distribution for which full-denoising is also exact.

There are several avenues to explore to extend this work. We could try to extend our results for the *linear manifold* to a more general low-dimensional manifold, drawing inspiration from Azangulov et al. (2025) to get a result in their setting, while improving the dependency in σ and relaxing some assumptions. It would also be interesting to explore further around diffusion models, for which multiple denoising steps are repeated, and compare to existing theoretical result on the performance of the models (see, e.g., Bortoli, 2022; Chen et al., 2023; Conforti et al., 2025; Gentiloni-Silveri and Ocello, 2025). We believe that the idea of tuning the coefficient α of the denoising steps depending on the target distribution could be further pursued. We also hope that the techniques developed here could lead to a better understanding of the Wasserstein error of diffusion models, both for deterministic sampling (discretization of the diffusion ODE (3)) and stochastic sampling (discretization of a reverse time SDE). To do so, one should also take into account the error at initialization (approximating the noisy distribution by a Gaussian), the impact of the (possibly varying) step size of the noise schedule (Strasman et al., 2024) and the error in learning the score. For the latter, it would be interesting to include results on the ability of neural networks to learn the score for distributions supported on a lower-dimensional manifold, as done by Tang and Yang (2024) and by Azangulov et al. (2025). We tried to answer some of these questions related to the Wasserstein convergence guarantees of the multiple step diffusion models in a subsequent paper (Beyler and Bach, 2025).

Acknowledgments

We thank the Action Editor and the referees for their valuable feedback. We thank Saeed Saremi and Dario Shariatian for insightful discussions related to this work. This work has received support from the French government, managed by the National Research Agency, under the France 2030 program with the reference "PR[AI]RIE-PSAI" (ANR-23-IACL-0008).

Appendices

Appendix A. Fokker-Planck Equation and the Diffusion ODE

In this section, we give a short proof of the Fokker-Planck equation, and use it to define the ODE (3) presented in section 3.3. There are many references about Fokker-Planck equations (see, e.g., Risken, 1996; Bogachev et al., 2015), but here we try to give a proof as simple and self-contain as possible. For general results on SDEs, we refer the reader to Le Gall (2016). We will always assume, but note explicitly write, that we have a probability space, a filtration and a Brownian motion such that all randoms variables are well defined.

We study the process $X_t = X + B_t$, in particular, $X_t = X + tZ$ with $X \perp Z$, $Z \sim \mathcal{N}(0, I)$. From that we deduce that X_t admit a density p_t with respect to the Lebesgue measure that verifies, for $t > 0$,

$$p_t(x) = \int \frac{\exp\left(-\|x-u\|^2/2t\right)}{(2\pi t)^{d/2}} d\mu_X(u) > 0,$$

and that $(t, x) \mapsto p_t(x)$ is $\mathcal{C}^\infty$.

The evolution of the marginal p_t is dictated by a partial differential equation, the Fokker-Planck equation.

Proposition 8 (Fokker-Planck equation) *Let $f \in \mathcal{C}^1(\mathbb{R} \times \mathbb{R}^d, \mathbb{R}^d)$, $g \in \mathcal{C}^0(\mathbb{R}, \mathbb{R})$, and p_t the density of a process X_t that verifies the following SDE*

$$dX_t = f(t, X_t)dt + g(t)dB_t, \tag{7}$$

and assume that $(t, x) \mapsto p_t(x)$ is $\mathcal{C}^2$ on $\mathbb{R}_+^* \times \mathbb{R}^d$. Then for all $t > 0$, we have

$$\partial_t p_t = -\nabla \cdot (f p_t) + \frac{1}{2} g^2 \Delta p_t.$$

Proof Let $\varphi \in \mathcal{C}^\infty(\mathbb{R}^d, \mathbb{R})$ with compact support. Then for $t > 0$,

$$\begin{aligned}
\int \varphi(x) \partial_t p_t(x) dx &= \int \varphi(x) \lim_{h \to 0} \frac{p_{t+h}(x) - p_t(x)}{h} dx \\
&= \lim_{h \to 0} \frac{1}{h} \left(\int \varphi(x) p_{t+h}(x) dx - \int \varphi(x) p_t(x) dx \right) \\
&= \lim_{h \to 0} \frac{1}{h} \mathbb{E}[\varphi(X_{t+h}) - \varphi(X_t)].
\end{aligned}$$

By Itō's formula,

$$\begin{aligned}
\varphi(X_{t+h}) - \varphi(X_t) &= \int_t^{t+h} \nabla \varphi(X_s) \cdot dX_s + \frac{1}{2} \int_t^{t+h} \Delta \varphi(X_s) d\langle X, X \rangle_s \\
&= \int_t^{t+h} \nabla \varphi(X_s) \cdot f(s, X_s) ds + \int_t^{t+h} g(s) \nabla \varphi(X_s) \cdot dB_s \\
&\quad + \frac{1}{2} \int_t^{t+h} \Delta \varphi(X_s) g(s)^2 ds.
\end{aligned}$$

Taking the expectation, we get that,

$$\mathbb{E}[\varphi(X_{t+h}) - \varphi(X_t)] = \int_t^{t+h} \mathbb{E}[\nabla\varphi(X_s) \cdot f(s, X_s)]ds + 0 + \frac{1}{2}\int_t^{t+h} E[\Delta\varphi(X_s)]g(s)^2 ds.$$

As φ has compact support, integration by part gives:

$$\mathbb{E}[\nabla\varphi(X_s) \cdot f(s, X_s)] = \int \nabla\varphi(x) \cdot f(s, x)p_s(x)dx = -\int \varphi(x)\nabla \cdot (f(s, x)p_s(x))dx,$$

and

$$E[\Delta\varphi(X_s)] = \int \Delta\varphi(x)p_s(x)dx = \int \varphi(x)\Delta p_s(x)dx.$$

It follows that,

$$\begin{aligned}
&\int \varphi(x)\partial_t p_t(x)dx \\
&\quad = \lim_{h\to 0} \frac{1}{h}\left(-\int_t^{t+h}\int \varphi(x)\nabla \cdot (f(s, x)p_s(x))dxds + \frac{1}{2}\int_t^{t+h}\int \varphi(x)g(s)^2\Delta p_s(x)dxds\right) \\
&\quad = -\int \varphi(x)\nabla \cdot (f(t, x)p_t(x))dx + \frac{1}{2}\int \varphi(x)g(t)^2\Delta p_t(x)dx,
\end{aligned}$$

as both integrands are continuous with respect to time, and we can rewrite,

$$\int \varphi(x)\left(\partial_t p_t(x) + \nabla \cdot (f(t, x)p_t(x)) - \frac{1}{2}g(t)^2\Delta p_t(x)\right)dx = 0,$$

for all $\varphi \in \mathcal{C}^\infty(\mathbb{R}^d, R)$ with compact support. It follows that for all $t > 0$, almost surely in x,

$$\partial_t p_t(x) = -\nabla \cdot (f(t, x)p_t(x)) + \frac{1}{2}g(t)^2\Delta p_t(x),$$

which gives the desired result as these quantities are continuous in x. ■

For the process $X_t = X + B_t$, we have $dX_t = dB_t$, which is (7) with $f = 0$ and $g = 1$. The Fokker-Planck equation is therefore

$$\partial_t p_t = \frac{1}{2}\Delta p_t.$$

As $\Delta p_t = \nabla \cdot \nabla p_t = \nabla \cdot \frac{p_t}{p_t}\nabla p_t = \nabla \cdot p_t \nabla \log p_t$, the equation can be rewritten as

$$\partial_t p_t = -\nabla \cdot \left(-\frac{1}{2}\nabla \log p_t\right)p_t,$$

which correspond to the Fokker-Planck equation with drift term $f(t, x) = -\frac{1}{2}\nabla \log p_t(x)$. This allows use to define the diffusion ODE, used in section 3.3.

Proposition 9 *Let $t^* > 0$. We can define a process $(x_t)_{t\geq 0}$ by*

$$\begin{cases} \frac{dx_t}{dt} & = -\frac{1}{2}\nabla \log p_t(x_t) \quad \text{for } t > 0 \\ x_{t^*} & = X_{t^*}. \end{cases} \tag{3}$$

This process has the same marginals as X_t: $\forall t \in [0, +\infty[, x_t \sim X_t$, and verifies, for all $t, s \geq 0$,

$$x_t - x_s = -\frac{1}{2}\int_s^t \nabla \log p_u(x_u) du.$$

Proof We know that for $t > 0$, $(t, x) \mapsto p_t(x)$ is $\mathcal{C}^\infty$. In particular, the ODE defined by (3) can be solved for all time $t > 0$, and we have for $s, t > 0$:

$$x_t - x_s = -\frac{1}{2}\int_s^t \nabla \log p_u(x_u) du. \tag{8}$$

For now, we will write $\tilde{p}_t$ the density of the marginal of x_t. As x_t is defined as the solution of an ODE, we introduce the resolvent $R(s, t, x)$ that gives the solution at time t of the ODE

$$y' = -\frac{1}{2}\nabla \log p_t(y),$$

with initial condition x at time s. R is invertible and verifies $R(s, t, x)^{-1} = R(t, s, x)$. Moreover, as $(t, x) \mapsto \nabla \log p_t(x)$ is C^∞, R is also $\mathcal{C}^\infty$ (see, e.g. Paulin, 2009, Théorème 7.21).

By construction, $x_t = R(t^*, t, X_{t^*})$, leading to:

$$\tilde{p}_t(x) = |\det \nabla R(t, t^*, x)| p_{t^*}(R(t, t^*, x)),$$

in particular, $(t, x) \mapsto \tilde{p}_t(x)$ is $\mathcal{C}^\infty$ on $\mathbb{R}_+^* \times \mathbb{R}^d$. We can apply proposition 8, to get that

$$\partial_t \tilde{p}_t = \nabla \cdot \left(-\frac{1}{2}\nabla \log p_t\right) = -\frac{1}{2}\Delta \log p_t = \partial_t p_t,$$

and as $\tilde{p}_{t^*} = p_{t^*}$, it leads to $\tilde{p}_t = p_t$ for all $t > 0$.

We finally need to verify that we can extend (8) up to time $s = 0$, i.e., that the trajectories can be integrated up to time $t = 0$. For $t > 0$, Tweedie's formula (1) gives:

$$\nabla \log p_t(x) = \frac{1}{t}(x - \mathbb{E}[X | X_t = x]).$$

In particular, as x_t and X_t have the same distribution,

$$\mathbb{E}[\|\nabla \log p_t(x_t)\|] = \frac{1}{t}\mathbb{E}[\|X_t - \mathbb{E}[X|X_t]\|] = \frac{1}{t}\mathbb{E}[\|\mathbb{E}[X_t - X|X_t]\|].$$

Jensen's inequality gives

$$\begin{aligned} \mathbb{E}[\|\mathbb{E}[X_t - X|X_t]\|] &\leq \mathbb{E}[\mathbb{E}[\|X_t - X\||X_t]] = \mathbb{E}[\|X_t - X\|] \\ &\leq \sqrt{\mathbb{E}[\|X - X_t\|^2]} = \sqrt{\mathbb{E}[\|B_t\|^2]} = \sqrt{td}, \end{aligned}$$

therefore,

$$\mathbb{E}[\|\nabla \log p_t(x_t)\|] \leq \sqrt{\frac{d}{t}},$$

which is integrable near 0. We get that

$$\mathbb{E}\left[\int_0^t \|\nabla \log p_u(x_u)\| du\right] < \infty,$$

hence

$$\int_0^t \|\nabla \log p_u(x_u)\| du < \infty,$$

almost everywhere, proving that we can extend (8) up to time $s = 0$ (almost everywhere). ■

Appendix B. Proofs

B.1 Proofs of Proposition 1 and Corollary 2

We will first state the following lemma:

Lemma 10 *Let X be a random variable with density p_X such that $\mathbb{E}[\|\nabla \log p_X(X)\|^p] \leq C$ for some constant $C < \infty$ and $p \geq 1$. Then for all $\sigma > 0$, defining $Y = X + \varepsilon$ with $X \perp \varepsilon$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, we have,*

$$\mathbb{E}[\|\nabla \log p_Y(Y)\|^p] \leq C.$$

Proof From (B.2) of Saremi et al. (2023), $\nabla \log p_Y(y) = \mathbb{E}[\nabla \log p_X(X)|Y = y]$ hence:

$$\begin{aligned}
\mathbb{E}[\|\nabla \log p_Y(Y)\|^p] &= \mathbb{E}[\|\mathbb{E}[\nabla \log p_X(X)|Y]\|^p] \\
\text{(Jensen's inequality)} &\leq \mathbb{E}[\mathbb{E}[\|\nabla \log p_X(X)\|^p|Y]] \\
&= \mathbb{E}[\|\nabla \log p_X(X)\|^p] \leq C.
\end{aligned}$$

■

We can now prove Proposition 1.

Proof (Proposition 1)

We follow Hyvärinen (2025) and keep track of the different constants.

For $\alpha = \frac{1}{2}$: Let $\hat{X} = \varphi_{1/2}(Y)$. As $Y = X + \varepsilon$ with $X \perp \varepsilon$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, we have $\hat{p}_Y(\xi) = \hat{p}_X(\xi) e^{-\frac{\sigma^2}{2}\|\xi\|^2}$.

The proof uses the equality:

$$\mathbb{E}[e^{i\xi \cdot Y} i \frac{\sigma^2}{2} \xi \cdot \nabla \log p_Y(Y)] = \frac{\sigma^2}{2} \|\xi\|^2 \hat{p}_Y(\xi),$$

that comes from an integration by parts ((19) of Hyvärinen (2025)). Then the idea is to write $\hat{p}_{\hat{X}}(\xi) = \mathbb{E}[e^{i\xi \cdot \hat{X}}] = \mathbb{E}[e^{i\xi \cdot Y} e^{i\frac{\sigma^2}{2}\xi \cdot \nabla \log p_Y(Y)}] = \mathbb{E}[e^{i\xi \cdot Y}(1 + i\frac{\sigma^2}{2}\xi \cdot \nabla \log p_Y(Y) + O(\sigma^4)] =$

$\mathbb{E}[e^{i\xi\cdot Y}]+\mathbb{E}[e^{i\xi\cdot Y}i\frac{\sigma^2}{2}\xi\cdot\nabla\log p_Y(Y)]+O(\sigma^4)=\hat{p}_Y(\xi)\left(1+\frac{\sigma^2}{2}\|\xi\|^2\right)+O(\sigma^4)=\hat{p}_Y(\xi)e^{\frac{\sigma^2}{2}\|\xi\|^2}+O(\sigma^4)=\hat{p}_X(\xi)+O(\sigma^4)$.

To make it quantitative, we start by noticing that for any $z\in\mathbb{C}$, $d\in\mathbb{N}$,

$$\left|e^z-\sum_{k=0}^{d}\frac{z^k}{k!}\right|\leq\frac{|z|^{d+1}\max(1,e^{\mathfrak{Re}(z)})}{(d+1)!}.$$

Then

$$\begin{aligned}
|\hat{p}_{\hat{X}}(\xi)-(1+1/2\sigma^2\|\xi\|^2)\hat{p}_Y(\xi)| &\leq \mathbb{E}[|e^{i\xi\cdot Y}|\cdot|e^{i\frac{1}{2}\sigma^2\xi\cdot\nabla\log p_Y(Y)}-(1+i\frac{1}{2}\sigma^2\xi\cdot\nabla\log p_Y(Y))|]\\
&\leq \mathbb{E}[\frac{1}{2}(\frac{1}{2}\sigma^2\xi\cdot\nabla\log p_Y(Y))^2]\\
&\leq \frac{\sigma^4\|\xi\|^2}{8}\mathbb{E}[\|\nabla\log p_Y(Y)\|^2]\\
&\leq \frac{\sigma^4\|\xi\|^2C}{8}\quad\text{(Lemma 10)},
\end{aligned}$$

and,

$$\begin{aligned}
|\hat{p}_X(\xi)-(1+1/2\sigma^2\|\xi\|^2)\hat{p}_Y(\xi)| &= |\hat{p}_X(\xi)||1-(1+1/2\sigma^2\|\xi\|^2)e^{-\frac{1}{2}\sigma^2\|\xi\|^2}|\\
&\leq |1-(1+1/2\sigma^2\|\xi\|^2)e^{-\frac{1}{2}\sigma^2\|\xi\|^2}|\\
&= e^{-\frac{1}{2}\sigma^2\|\xi\|^2}|e^{\frac{1}{2}\sigma^2\|\xi\|^2}-(1+\frac{1}{2}\sigma^2\|\xi\|^2)|\\
&\leq e^{-\frac{1}{2}\sigma^2\|\xi\|^2}e^{\frac{1}{2}\sigma^2\|\xi\|^2}\frac{1}{2}(\frac{1}{2}\sigma^2\|\xi\|^2)^2\\
&= \frac{\sigma^4\|\xi\|^4}{8}.
\end{aligned}$$

Finally we get the result by combining the two inequalities.

For $\alpha\neq\frac{1}{2}$: Let $\hat{X}=\varphi_\alpha(Y)$. We can still write $\hat{p}_{\hat{X}}(\xi)=\mathbb{E}[e^{i\xi\cdot\hat{X}}]=\mathbb{E}[e^{i\xi\cdot Y}e^{i\alpha\sigma^2\xi\cdot\nabla\log p_Y(Y)}]=\mathbb{E}[e^{i\xi\cdot Y}(1+i\alpha\sigma^2\xi\cdot\nabla\log p_Y(Y)+O(\sigma^4)]=\mathbb{E}[e^{i\xi\cdot Y}]+\mathbb{E}[e^{i\xi\cdot Y}i\alpha\sigma^2\xi\cdot\nabla\log p_Y(Y)]+O(\sigma^4)=\hat{p}_Y(\xi)\left(1+\alpha\sigma^2\|\xi\|^2\right)+O(\sigma^4)=\hat{p}_Y(\xi)e^{\alpha\sigma^2\|\xi\|^2}+O(\sigma^4)=\hat{p}_X(\xi)e^{\left(\alpha-\frac{1}{2}\right)\sigma^2\|\xi\|^2}+O(\sigma^4)$. But here the term $\left(\alpha-\frac{1}{2}\right)$ does not cancel, hence we do not have $\hat{p}_{\hat{X}}(\xi)=\hat{p}_X(\xi)+O(\sigma^4)$. However, we can still write that $e^{\left(\alpha-\frac{1}{2}\right)\sigma^2\|\xi\|^2}=1+O(\sigma^2)$ to have the equality, but to a smaller order in σ. Quantitatively:

$$\begin{aligned}
|\hat{p}_{\hat{X}}(\xi)-\hat{p}_Y(\xi)| &\leq \mathbb{E}[|e^{i\xi\cdot Y}|\cdot|e^{i\alpha\sigma^2\xi\cdot\nabla\log p_Y(Y)}-1|]\\
&\leq \mathbb{E}[|\alpha\sigma^2\xi\cdot\nabla\log p_Y(Y)|]\\
&\leq \alpha\sigma^2\|\xi\|\mathbb{E}[\|\nabla\log p_Y(Y)\|]\\
&\leq \alpha\sigma^2\|\xi\|\sqrt{\mathbb{E}[\|\nabla\log p_Y(Y)\|^2]}\\
&\leq \alpha\sigma^2\|\xi\|\sqrt{C}\quad\text{(Lemma 10)},
\end{aligned}$$

and,

$$\begin{aligned}
|\hat{p}_X(\xi) - \hat{p}_Y(\xi)| &= |\hat{p}_X(\xi)| \cdot |1 - 1e^{-\frac{1}{2}\sigma^2\|\xi\|^2}| \\
&\leq |1 - e^{-\frac{1}{2}\sigma^2\|\xi\|^2}| \\
&= e^{-\frac{1}{2}\sigma^2\|\xi\|^2}|e^{\frac{1}{2}\sigma^2\|\xi\|^2} - 1| \\
&\leq e^{-\frac{1}{2}\sigma^2\|\xi\|^2}e^{\frac{1}{2}\sigma^2\|\xi\|^2}(\frac{1}{2}\sigma^2\|\xi\|^2) \\
&= \frac{\sigma^2\|\xi\|^2}{2},
\end{aligned}$$

which gives the result by combining the two inequalities. ■

Proof (Corollary 2)

For $\alpha = \frac{1}{2}$:

$$\begin{aligned}
\mathrm{MMD}_k(\mathcal{L}(X), \mathcal{L}(\varphi_{1/2}(Y))) &= \left(\int |\hat{p}_X(\xi) - \hat{p}_{\varphi_{1/2}(Y)}(\xi)|^2 d\Lambda(\xi)\right)^{1/2} \\
&\leq \left(\int \left(\frac{\sigma^4(C\|\xi\|^2 + \|\xi\|^4)}{8}\right)^2 d\Lambda(\xi)\right)^{1/2} \\
&\leq \frac{\sigma^4}{8}\left(2C^2\int \|\xi\|^4 d\Lambda(\xi) + 2\int \|\xi\|^8 d\Lambda(\xi)\right)^{1/2} \\
&= \sigma^4\frac{\sqrt{C^2C_4 + C_8}}{4\sqrt{2}}.
\end{aligned}$$

For $\alpha \neq \frac{1}{2}$:

$$\begin{aligned}
\mathrm{MMD}_k(\mathcal{L}(X), \mathcal{L}(\varphi_\alpha(Y))) &= \left(\int |\hat{p}_X(\xi) - \hat{p}_{\varphi_\alpha(Y)}(\xi)|^2 d\Lambda(\xi)\right)^{1/2} \\
&\leq \left(\int \left(\frac{\sigma^2(2\alpha\sqrt{C}\|\xi\| + \|\xi\|^2)}{2}\right)^2 d\Lambda(\xi)\right)^{1/2} \\
&\leq \frac{\sigma^2}{2}\left(8\alpha^2 C\int \|\xi\|^2 d\Lambda(\xi) + 2\int \|\xi\|^4 d\Lambda(\xi)\right)^{1/2} \\
&= \sigma^2\frac{\sqrt{4\alpha^2CC_2 + C_4}}{\sqrt{2}}.
\end{aligned}$$

■

B.2 Proofs of Proposition 3, Lemma 4 and Proposition 5

Proof (Proposition 3)

We have

$$x_0 = x_t + \frac{1}{2}\int_0^t \nabla \log p_s(x_s) ds,$$

and

$$\hat{x}_0 = x_t + \alpha t \nabla \log p_t(x_t).$$

Therefore,

$$x_0 - \hat{x}_0 = \frac{1}{2}\int_0^t \nabla \log p_s(x_s) ds - \alpha t \nabla \log p_t(x_t).$$

Using Jensen's inequality two times,

$$\begin{aligned}\|x_0 - \hat{x}_0\|^2 &\leq 2\left(\left\|\frac{1}{2}\int_0^t \nabla \log p_s(x_s) ds\right\|^2 + \|\alpha t \nabla \log p_t(x_t)\|^2\right) \\ &\leq 2\left(\frac{t}{4}\int_0^t \|\nabla \log p_s(x_s)\|^2 ds + \alpha^2 t^2 \|\nabla \log p_t(x_t)\|^2\right),\end{aligned}$$

and it follows that,

$$W_2^2(\mathcal{L}(X), \mathcal{L}(\varphi_\alpha(X_t))) \leq \mathbb{E}[\|x_0 - x_t\|^2] \leq 2\left(\frac{t}{4}\int_0^t \mathbb{E}[\|\nabla \log p_s(x_s)\|^2] ds + \alpha^2 t^2 \mathbb{E}[\|\nabla \log p_t(x_t)\|^2]\right).$$

As stated in Lemma 10, assuming $\mathbb{E}[\|\nabla \log p_X(X)\|^2] \leq C$ leads to $\mathbb{E}[\|\nabla \log p_s(x_s)\|^2] \leq C$ for all $s \geq 0$, so finally:

$$W_2^2(\mathcal{L}(X), \mathcal{L}(\varphi_\alpha(X_t))) \leq \frac{(1+4\alpha^2)C}{2} t^2,$$

which gives for $t = \sigma^2$:

$$W_2(\mathcal{L}(X), \mathcal{L}(\varphi_\alpha(Y))) \leq \sqrt{\frac{(1+4\alpha^2)C}{2}}\sigma^2.$$

■

Proof (Lemma 4) Using the chain rule and equation (3), we have

$$\begin{aligned}\frac{d}{dt}\nabla \log p_t(x_t) &= \nabla^2 \log p_t(x_t) \cdot \frac{d}{dt} x_t + [\partial_t \nabla \log p_t](x_t) \\ &= -\frac{1}{2}\nabla^2 \log p_t(x_t) \cdot \nabla \log p_t(x_t) + [\nabla \partial_t \log p_t](x_t).\end{aligned}$$

Moreover, the Fokker-Planck equation for p_t is

$$\partial_t p_t = \frac{1}{2}\Delta p_t = \frac{1}{2}\nabla \cdot \nabla p_t = \frac{1}{2}\nabla \cdot (p_t \nabla \log p_t) = \frac{1}{2}\nabla p_t \cdot \nabla \log p_t + \frac{1}{2} p_t \Delta \log p_t,$$

hence,

$$\partial_t \log p_t = \frac{1}{2}\|\nabla \log p_t\|^2 + \frac{1}{2}\Delta \log p_t.$$

Taking the gradient in x gives

$$\nabla \partial_t \log p_t = \nabla^2 \log p_t \cdot \nabla \log p_t + \frac{1}{2} \nabla \Delta \log p_t.$$

This finally leads to

$$\frac{d}{dt} \nabla \log p_t(x_t) = \frac{1}{2} \nabla^2 \log p_t(x_t) \cdot \nabla \log p_t(x_t) + \frac{1}{2} \nabla \Delta \log p_t(x_t).$$

We can express $\nabla \log p_t$, $\nabla^2 \log p_t$ and $\nabla \Delta \log p_t$ in term of $\nabla \log p_X$, $\nabla^2 \log p_X$ and $\nabla \Delta \log p_X$. With $Y = X + \varepsilon$ with $X \perp \varepsilon$ and $\varepsilon \sim \mathcal{N}(0, tI)$, (B.2) and (B.4) of Saremi et al. (2023) give

$$\nabla \log p_t(x) = \mathbb{E}[\nabla \log p_X(X) | Y = x],$$

and

$$\begin{aligned}
\nabla^2 \log p_t(x) = \mathbb{E}[&\nabla^2 \log p_X(X) | Y = x] \\
&+ \mathbb{E}[\nabla \log p_X(X) \nabla \log p_X(X)^\top | Y = x] \\
&- \mathbb{E}[\nabla \log p_X(X) | Y = x] \mathbb{E}[\nabla \log p_X(X) | Y = x]^\top \\
= \mathbb{E}[&\nabla^2 \log p_X(X) | Y = x] + \operatorname{cov}(\nabla \log p_X(X) | Y = x).
\end{aligned}$$

Similarly, we can compute,

$$\begin{aligned}
\nabla \Delta \log p_t(x) = \mathbb{E}[&\nabla \Delta \log p_X(X) | Y = x] \\
&+ \mathbb{E}[\Delta \log p_X(X) \nabla \log p_X(X) | Y = x] \\
&- \mathbb{E}[\Delta \log p_X(X) | Y = x] \mathbb{E}[\nabla \log p_X(X) | Y = x] \\
&+ 2\mathbb{E}[\nabla^2 \log p_X(X) \cdot \nabla \log p_X(X) | Y = x] \\
&+ \mathbb{E}[\|\nabla \log p_X(X)\|^2 \nabla \log p_X(X) | Y = x] \\
&- \mathbb{E}[\|\nabla \log p_X(X)\|^2 | Y = x] \mathbb{E}[\nabla \log p_X(X) | Y = x] \\
&- 2\mathbb{E}[\nabla^2 \log p_X(X) | Y = x] \cdot \mathbb{E}[\nabla \log p_X(X) | Y = x] \\
&- 2\mathbb{E}[\nabla \log p_X(X) \nabla \log p_X(X)^\top | Y = x] \cdot \mathbb{E}[\nabla \log p_X(X) | Y = x] \\
&+ 2\|\mathbb{E}[\nabla \log p_X(X) | Y = x]\|^2 \mathbb{E}[\nabla \log p_X(X) | Y = x].
\end{aligned}$$

Combining the expressions above, we get:

$$\begin{aligned}
\frac{d}{dt}\nabla \log p_t(x_t) = &\frac{1}{2}\mathbb{E}[\nabla\Delta \log p_X(X)|Y = x_t] && (9)\\
&+ \frac{1}{2}\mathbb{E}[\Delta \log p_X(X)\nabla \log p_X(X)|Y = x_t] && (10)\\
&- \frac{1}{2}\mathbb{E}[\Delta \log p_X(X)|Y = x_t]\mathbb{E}[\nabla \log p_X(X)|Y = x_t] && (11)\\
&+ \mathbb{E}[\nabla^2 \log p_X(X) \cdot \nabla \log p_X(X)|Y = x_t] && (12)\\
&+ \frac{1}{2}\mathbb{E}[\|\nabla \log p_X(X)\|^2\nabla \log p_X(X)|Y = x_t] && (13)\\
&- \frac{1}{2}\mathbb{E}[\|\nabla \log p_X(X)\|^2|Y = x_t]\mathbb{E}[\nabla \log p_X(X)|Y = x_t] && (14)\\
&- \frac{1}{2}\mathbb{E}[\nabla^2 \log p_X(X)|Y = x_t] \cdot \mathbb{E}[\nabla \log p_X(X)|Y = x_t] && (15)\\
&- \frac{1}{2}\mathbb{E}[\nabla \log p_X(X)\nabla \log p_X(X)^\top|Y = x_t] \cdot \mathbb{E}[\nabla \log p_X(X)|Y = x_t] && (16)\\
&+ \frac{1}{2}\|\mathbb{E}[\nabla \log p_X(X)|Y = x_t]\|^2\mathbb{E}[\nabla \log p_X(X)|Y = x_t]. && (17)
\end{aligned}$$

We want to control all these terms in L_2 norm. First note that for all $x \in \mathbb{R}^d$, $|\Delta \log p_X(x)| = |\operatorname{tr}(\nabla^2 \log p_X(x))| \leq \sqrt{d}\|\nabla^2 \log p_X(x))\|_{\text{fro}} \leq d\|\nabla^2 \log p_X(x)\|_{\text{op}}$, where $\|A\|_{\text{fro}}$ is defined for any matrix A as $\|A\|_{\text{fro}} = \sqrt{\operatorname{tr}(AA^\top)}$ and verifies $\|A\|_{\text{fro}} \leq \sqrt{d}\|A\|_{\text{op}}$. In particular, $\mathbb{E}[|\Delta \log p_X(X)|^3] \leq d^3C_2$. Also note that x_t has the same distribution as Y. We can now control in expectation all the the terms in $\frac{d}{dt}\nabla \log p_t(x_t)$.

Term (9):

$$\begin{aligned}
\mathbb{E}[\|\mathbb{E}[\nabla\Delta \log p_X(X)|Y]\|^2] &\leq \mathbb{E}[\mathbb{E}[\|\nabla\Delta \log p_X(X)\|^2|Y]] = \mathbb{E}[\|\nabla\Delta \log p_X(X)\|^2]\\
&\quad \text{(Jensen's inequality on the conditional expectation)}\\
&= C_3.
\end{aligned}$$

Term (10):

$$\begin{aligned}
\mathbb{E}[\|\mathbb{E}[\Delta \log p_X(X)\nabla \log p_X(X)|Y]\|^2] &\leq \mathbb{E}[|\Delta \log p_X(X)|^2\|\nabla \log p_X(X)\|^2]\\
&\quad \text{(Jensen's inequality on the conditional expectation)}\\
&\leq \mathbb{E}[|\Delta \log p_X(X)|^3]^{2/3}\mathbb{E}[\|\nabla \log p_X(X)\|^6]^{1/3}\\
&\quad \text{(Hölder's inequality)}\\
&\leq d^2C_2^{2/3}C_1^{1/3}.
\end{aligned}$$

Term (11):

$$\begin{aligned}
&\mathbb{E}[\|\mathbb{E}[\Delta \log p_X(X)|Y]\mathbb{E}[\nabla \log p_X(X)|Y]\|^2] \\
&\qquad = \mathbb{E}[|\mathbb{E}[\Delta \log p_X(X)|Y]|^2\|\mathbb{E}[\nabla \log p_X(X)|Y]\|^2] \\
&\qquad \le \mathbb{E}[|\mathbb{E}[\Delta \log p_X(X)|Y]|^3]^{2/3}\mathbb{E}[\|\mathbb{E}[\nabla \log p_X(X)|Y]\|^6]^{1/3} \\
&\qquad\quad \text{(Hölder's inequality)} \\
&\qquad \le \mathbb{E}[|\Delta \log p_X(X)|^3]^{2/3}\mathbb{E}[\|\nabla \log p_X(X)\|^6]^{1/3} \\
&\qquad\quad \text{(Jensen's inequality on the conditional expectation)} \\
&\qquad \le d^2 C_2^{2/3} C_1^{1/3}.
\end{aligned}$$

Term (12):

$$\begin{aligned}
&\mathbb{E}[\|\mathbb{E}[\nabla^2 \log p_X(X) \cdot \nabla \log p_X(X)|Y]\|^2] \\
&\qquad \le \mathbb{E}[\|\nabla^2 \log p_X(X) \cdot \nabla \log p_X(X)\|^2] \\
&\qquad\quad \text{(Jensen's inequality on the conditional expectation)} \\
&\qquad \le \mathbb{E}[\|\nabla^2 \log p_X(X)\|_{\mathrm{op}}^2 \|\nabla \log p_X(X)\|^2] \\
&\qquad \le \mathbb{E}[\|\nabla^2 \log p_X(X)\|_{\mathrm{op}}^3]^{2/3}\mathbb{E}[\|\nabla \log p_X(X)\|^6]^{1/3} \\
&\qquad\quad \text{(Hölder's inequality)} \\
&\qquad = C_2^{2/3} C_1^{1/3}.
\end{aligned}$$

Term (13):

$$\begin{aligned}
&\mathbb{E}[\|\mathbb{E}[\|\nabla \log p_X(X)\|^2 \nabla \log p_X(X)|Y]\|^2] \\
&\qquad \le \mathbb{E}[\|\nabla \log p_X(X)\|^6] \\
&\qquad\quad \text{(Jensen's inequality on the conditional expectation)} \\
&\qquad = C_1.
\end{aligned}$$

Term (14):

$$\begin{aligned}
&\mathbb{E}[\|\mathbb{E}[\|\nabla \log p_X(X)\|^2|Y]\mathbb{E}[\nabla \log p_X(X)|Y]\|^2] \\
&\qquad = \mathbb{E}[(\mathbb{E}[\|\nabla \log p_X(X)\|^2|Y])^2\|\mathbb{E}[\nabla \log p_X(X)|Y]\|^2] \\
&\qquad \le \mathbb{E}[(\mathbb{E}[\|\nabla \log p_X(X)\|^2|Y])^2\mathbb{E}[\|\nabla \log p_X(X)\|^2|Y]] \\
&\qquad\quad \text{(Jensen's inequality on the conditional expectation)} \\
&\qquad = \mathbb{E}[(\mathbb{E}[\|\nabla \log p_X(X)\|^2|Y])^3] \\
&\qquad \le \mathbb{E}[\|\nabla \log p_X(X)\|^6] \\
&\qquad\quad \text{(Jensen's inequality on the conditional expectation)} \\
&\qquad = C_1.
\end{aligned}$$

Term (15):

$$\begin{aligned}
&\mathbb{E}[\|\mathbb{E}[\nabla^2 \log p_X(X)|Y] \cdot \mathbb{E}[\nabla \log p_X(X)|Y]\|^2] \\
&\qquad \leq \mathbb{E}[\|\mathbb{E}[\nabla^2 \log p_X(X)|Y]\|_{\text{op}}^2 \|\mathbb{E}[\nabla \log p_X(X)|Y]\|^2] \\
&\qquad \leq \mathbb{E}[\|\mathbb{E}[\nabla^2 \log p_X(X)|Y]\|_{\text{op}}^3]^{2/3} \mathbb{E}\|\mathbb{E}[\nabla \log p_X(X)|Y]\|^6]^{1/3} \\
&\qquad\quad \text{(Hölder's inequality)} \\
&\qquad \leq \mathbb{E}[\|\nabla^2 \log p_X(X)\|_{\text{op}}^3]^{2/3} \mathbb{E}\|\nabla \log p_X(X)\|^6]^{1/3} \\
&\qquad\quad \text{(Jensen's inequality on the conditional expectation)} \\
&\qquad = C_2^{2/3} C_1^{1/3}.
\end{aligned}$$

Term (16):

$$\begin{aligned}
&\mathbb{E}[\|\mathbb{E}[\nabla \log p_X(X) \nabla \log p_X(X)^\top |Y] \cdot \mathbb{E}[\nabla \log p_X(X)|Y]\|^2] \\
&\qquad \leq \mathbb{E}[\|\mathbb{E}[\nabla \log p_X(X) \nabla \log p_X(X)^\top |Y]\|_{\text{op}}^2 \|\mathbb{E}[\nabla \log p_X(X)|Y]\|^2] \\
&\qquad \leq \mathbb{E}[(\mathbb{E}[\|\nabla \log p_X(X) \nabla \log p_X(X)^\top\|_{\text{op}}|Y])^2 \mathbb{E}[\|\nabla \log p_X(X)\|^2|Y]] \\
&\qquad\quad \text{(Jensen's inequality on the conditional expectation)} \\
&\qquad = \mathbb{E}[(\mathbb{E}[\|\nabla \log p_X(X)\|^2|Y])^2 \mathbb{E}[\|\nabla \log p_X(X)\|^2|Y]] \\
&\qquad = \mathbb{E}[(\mathbb{E}[\|\nabla \log p_X(X)\|^2|Y])^3] \\
&\qquad \leq \mathbb{E}[\|\nabla \log p_X(X)\|^6] \\
&\qquad\quad \text{(Jensen's inequality on the conditional expectation)} \\
&\qquad = C_1.
\end{aligned}$$

Term (17):

$$\begin{aligned}
&\mathbb{E}[\|\|\mathbb{E}[\nabla \log p_X(X)|Y]\|^2 \mathbb{E}[\nabla \log p_X(X)|Y]\|^2] \\
&\qquad = \mathbb{E}[\|\mathbb{E}[\nabla \log p_X(X)|Y]\|^6] \\
&\qquad \leq \mathbb{E}[\|\nabla \log p_X(X)\|^6] \\
&\qquad\quad \text{(Jensen's inequality on the conditional expectation)} \\
&\qquad = C_1.
\end{aligned}$$

To combine this bounds, we use Jensen's inequality on $x \mapsto x^2$, that gives for all $x_1, \ldots, x_k \in \mathbb{R}$,

$$\left(\sum_{i=1}^k x_i \right)^2 \leq k \sum_{i=1}^k x_i^2,$$

and we finally have,

$$\mathbb{E}\left[\left\| \frac{d}{dt} \nabla \log p_t(x_t) \right\|^2 \right] \leq \frac{9}{4}(4C_1 + (2d^2 + 5) C_1^{1/3} C_2^{2/3} + C_3) = C.$$

■

Proof (Proposition 5) We have,

$$\begin{aligned} x_0 - \hat{x}_0 &= \frac{1}{2}\int_0^t \nabla \log p_s(x_s) ds - \frac{1}{2} t \nabla \log p_t(x_t) \\ &= \frac{1}{2}\int_0^t \left(\nabla \log p_s(x_s) - \nabla \log p_t(x_t)\right) ds \\ &= \frac{1}{2}\int_0^t \int_s^t \frac{d}{du}\left(\nabla \log p_u(x_u)\right) du ds, \end{aligned}$$

hence, with Jensen's inequality,

$$\mathbb{E}[\|x_0 - x_t\|^2] \leq \frac{t}{4}\int_0^t (t-s) \int_s^t \mathbb{E}\left[\left\| \frac{d}{du}\left(\nabla \log p_u(x_u)\right)\right\|^2\right] du ds.$$

Using Lemma 4, it leads to $\mathbb{E}[\|x_0 - x_t\|^2] \leq \frac{C}{12} t^4$. In particular for $t = \sigma^2$, we get:

$$W_2(\mathcal{L}(X), \mathcal{L}(\varphi_{1/2}(Y))) \leq (\mathbb{E}[\|x_0 - x_{\sigma^2}\|^2])^{1/2} \leq K\sigma^4,$$

with $K = \sqrt{\frac{C}{12}} = \frac{\sqrt{3}}{4}\sqrt{4C_1 + (2d^2+5)C_1^{1/3}C_2^{2/3} + C_3}$. ■

B.3 Proof of Proposition 6

Proof As the added noise ε is isotropic, it is invariant by rotation, thus, we can limit ourselves to the setting where $H = \mathbb{R}^m \times \{0\}^{d-m}$ for which the variable X can be written $X = (X_1, 0)$ with $X_1 \in \mathbb{R}^m$.

Write $\varepsilon = (\varepsilon_1, \varepsilon_2)$, with $\varepsilon_1 \in \mathbb{R}^m \sim \mathcal{N}(0, \sigma^2 I_m), \varepsilon_2 \in \mathbb{R}^{(d-m)} \sim \mathcal{N}(0, \sigma^2 I_{d-m})$ and $Y = (Y_1, Y_2) = (X_1 + \varepsilon_1, \varepsilon_2)$, with $Y_1 \in \mathbb{R}^m, Y_2 \in \mathbb{R}^{(d-m)}$ and $Y_1 \perp Y_2$. Hence, we have that

$$\nabla \log p_Y(y) = (\nabla \log p_{Y_1}(y_1), \nabla \log p_{Y_2}(y_2)).$$

For the Gaussian $Y_2 = \varepsilon_2 \sim \mathcal{N}(0, \sigma^2 I)$, $\nabla \log p_{Y_2}(y_2) = \frac{-y_2}{\sigma^2}$.

Therefore we have $\varphi_\alpha(Y) = (\varphi_\alpha(Y_1), (1-\alpha)\varepsilon_2)$ and we can use the fact that for distributions $\mu_1, \mu_2, \nu_1, \nu_2$ $W_2^2(\mu_1 \otimes \mu_2, \nu_1 \otimes \nu_2) = W_2^2(\mu_1, \nu_1) + W_2^2(\mu_2, \nu_2)$, leading to

$$W_2^2(\mathcal{L}(X), \mathcal{L}(\varphi_\alpha(Y))) = W_2^2(\mathcal{L}(X_1), \mathcal{L}(\varphi_\alpha(Y_1))) + W_2^2(0, (1-\alpha)\varepsilon_2),$$

with $W_2^2(\mathcal{L}(0), \mathcal{L}((1-\alpha)\varepsilon_2)) = (d-m)(1-\alpha)^2\sigma^2$.

Note in particular that for $\alpha = 1$,

$$\varphi_\alpha(Y) = (\varphi_\alpha(Y_1), (1-\alpha)\varepsilon_2) = (\varphi_\alpha(Y_1), 0) \in H = \mathbb{R}^m \times \{0\}^{d-m}.$$

Full-denoising hence ensures that the denoising variable $\varphi_\alpha(Y)$ belongs to the subspace H as it remove all the orthogonal noise ε_2. ■

B.4 Proof of Proposition 7

Proof Let $\mu = \sum_{i=1}^{N} \pi_i \mu_i$ (with $\sum_{i=1}^{N} \pi_i = 1, \pi_i \geq 0$) a mixture of distribution μ_i with compact support S_i such that $D = \min_{i \neq j} d(S_i, S_j) > 0$ ($d(S_i, S_j) = \min_{x_i \in S_i, x_j \in S_j} \|x_i - x_j\|$).

We denote $X \sim \mu$, $Y = X + \varepsilon$ with $X \perp \varepsilon$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, for $\sigma > 0$. For $\alpha \in \mathbb{R}$, we denote $\varphi_\alpha(y) = y + \alpha\sigma^2 \nabla \log p_Y(y)$, ν the law of Y and μ_α the law of $\varphi_\alpha(Y)$. Similarly, for $X_i \sim \mu_i$, and $Y_i = X_i + \varepsilon$ with $X_i \perp \varepsilon$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, we denote $\varphi_{i,\alpha}(y) = y + \alpha\sigma^2 \nabla \log p_{Y_i}(y)$, ν_i the law of Y_i and $\mu_{i,\alpha}$ the law of $\varphi_{i,\alpha}(Y_i)$. We also denote $\hat{\mu}_{i,\alpha}$ the law of $\varphi_\alpha(Y_i)$.

We denote $R = \sup_{i, x \in S_i} \|x\| < \infty$. We have

$$\varphi_\alpha(y) = y + \alpha\sigma^2 \nabla \log p_Y(y) = (1 - \alpha)y + \alpha \mathbb{E}[X | Y = y],$$

and,

$$\varphi_{i,\alpha}(y) = y + \alpha\sigma^2 \nabla \log p_{Y_i}(y) = (1 - \alpha)y + \alpha \mathbb{E}[X_i | Y_i = y].$$

Note that for all y, $\|\mathbb{E}[X|Y = y]\| \leq R$ and $\|\mathbb{E}[X_i|Y_i = y]\| \leq R$.

By limiting ourselves to the couplings $\Gamma^0(\mu, \mu_\alpha) = \{\gamma : (X, \hat{X}) \sim \gamma \text{ with } X = \sum_i 1_{Z=i} X_i, \hat{X} = \sum 1_{Z=i} \hat{X}_i, Z \text{ such that } P(Z = i) = \pi_i, \text{ and } (X_i, \hat{X}_i) \sim \gamma_i \in \Gamma(\mu_i, \hat{\mu}_{i,\alpha})\}$, we have:

$$\begin{aligned}
W_2^2(\mu, \mu_\alpha) &= \inf_{\gamma \in \Gamma(\mu, \mu_\alpha)} \int \|x - \hat{x}\|^2 d\gamma(x, \hat{x}) \\
&\leq \inf_{\gamma \in \Gamma^0(\mu, \mu_\alpha)} \int \|x - \hat{x}\|^2 d\gamma(x, \hat{x}) \\
&= \sum_i \pi_i \inf_{\gamma \in \Gamma(\mu_i, \hat{\mu}_{i,\alpha})} \int \|x - \hat{x}\|^2 d\gamma(x, \hat{x}) \\
&= \sum_i \pi_i W_2^2(\mu_i, \hat{\mu}_{i,\alpha}).
\end{aligned}$$

Then for $i \in \{1, \ldots, N\}$ we have:

$$\begin{aligned}
W_2^2(\mu_i, \hat{\mu}_{i,\alpha}) &= \inf_{\gamma \in \Gamma(\mu_i, \hat{\mu}_{i,\alpha})} \int \|x - \hat{x}\|^2 d\gamma(x, \hat{x}) \\
&= \inf_{\gamma \in \Gamma(\mu_i, \nu_i)} \int \|x - \varphi_\alpha(y)\|^2 d\gamma(x, y) \\
&\leq 2 \inf_{\gamma \in \Gamma(\mu_i, \nu_i)} \int \|x - \varphi_{i,\alpha}(y)\|^2 d\gamma(x, y) + \int \|\varphi_{i,\alpha}(y) - \varphi_\alpha(y)\|^2 d\gamma(x, y) \\
&= 2W_2^2(\mu_i, \mu_{i,\alpha}) + \int \|\varphi_{i,\alpha}(y) - \varphi_\alpha(y)\|^2 d\nu_i(y).
\end{aligned}$$

To conclude, it is sufficient to prove that $\int \|\varphi_{i,\alpha}(y) - \varphi_\alpha(y)\|^2 d\nu_i(y) = O\left(\frac{1}{\sigma^{d-2}} \exp\left(-\frac{K}{\sigma^2}\right)\right)$ with $K = K(D)$ a constant. We fix $\delta_1 > 0$ such that $D - \delta_1 > 0$.

We have,

$$\int \|\varphi_{i,\alpha}(y) - \varphi_\alpha(y)\|^2 d\nu_i(y) = \int_{y \in S_i + B(0,\delta_1)} \|\varphi_{i,\alpha}(y) - \varphi_\alpha(y)\|^2 d\nu_i(y) + \int_{y \notin S_i + B(0,\delta_1)} \|\varphi_{i,\alpha}(y) - \varphi_\alpha(y)\|^2 d\nu_i(y).$$

We start by bounding the term $\int_{y \notin S_i + B(0,\delta_1)} \|\varphi_{i,\alpha}(y) - \varphi_\alpha(y)\|^2 d\nu_i(y)$.

$$\begin{aligned}\int_{y \notin S_i + B(0,\delta_1)} \|\varphi_{i,\alpha}(y) - \varphi_\alpha(y)\|^2 d\nu_i(y) &= \mathbb{E}[\|\varphi_{i,\alpha}(Y_i) - \varphi_\alpha(Y_i)\|^2 1_{Y_i \notin S_i + B(0,\delta_1)}] \\ &= \alpha^2 \mathbb{E}[\|\mathbb{E}[X_i|Y_i] - \mathbb{E}[X|Y = Y_i]\|^2 1_{Y_i \notin S_i + B(0,\delta_1)}] \\ &\leq 4R^2\alpha^2 P(Y_i \notin S_i + B(0,\delta_1)).\end{aligned}$$

We have $Y_i = X_i + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and $\frac{\|\varepsilon\|^2}{\sigma^2} \sim \chi^2$ (chi-squared distribution). Moreover, as $\mathrm{supp}(X_i) \subset S_i$, $\{Y_i \notin S_i + B(0,\delta_1)\} \subset \{\varepsilon \notin B(0,\delta_1)\} = \left\{\frac{\|\varepsilon\|^2}{\sigma^2} > \frac{\delta_1^2}{\sigma^2}\right\}$. Therefore:

$$\begin{aligned}P(Y_i \notin S_i + B(0,\delta_1)) &\leq P\left(\frac{\|\varepsilon\|^2}{\sigma^2} > \frac{\delta_1^2}{\sigma^2}\right) \\ &= \frac{(1/2)^{d/2}}{\Gamma(d/2)} \int_{\frac{\delta_1^2}{\sigma^2}}^{\infty} t^{d/2-1} e^{-t/2} dt \\ &\leq \frac{4(1/2)^{d/2}\delta_1^{d-2}}{\Gamma(d/2)} \frac{1}{\sigma^{d-2}} \exp\left(-\frac{\delta_1^2}{2\sigma^2}\right) \\ &= O\left(\frac{1}{\sigma^{d-2}} \exp\left(-\frac{\delta_1^2}{2\sigma^2}\right)\right) \text{ when } \sigma \to 0,\end{aligned}$$

where Γ is the Gamma function defined by $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt$, for $\mathfrak{Re}(z) > 0$. This leads to

$$\begin{aligned}\int_{y \notin S_i + B(0,\delta_1)} \|\varphi_{i,\alpha}(y) - \varphi_\alpha(y)\|^2 d\nu_i(y) &\leq \frac{16R^2\alpha^2(1/2)^{d/2}\delta_1^{d-2}}{\Gamma(d/2)} \frac{1}{\sigma^{d-2}} \exp\left(-\frac{\delta_1^2}{2\sigma^2}\right) \\ &= O\left(\frac{1}{\sigma^{d-2}} \exp\left(-\frac{\delta_1^2}{2\sigma^2}\right)\right) \text{ when } \sigma \to 0.\end{aligned}$$

We now turn to the term $\int_{y \in S_i + B(0,\delta_1)} \|\varphi_{i,\alpha}(y) - \varphi_\alpha(y)\|^2 d\nu_i(y)$. We fix $\delta_2 > 0$ such that $(D - \delta_1)^2 - \delta_2 > 0$ and we denote $A = \{p_{Y_i}(y) \leq \frac{1}{\pi_i \sigma^{d-2}} \exp\left(-\frac{(D-\delta_1)^2 - \delta_2}{2\sigma^2}\right)\}$ and

$B = \{p_{Y_i}(y) \geq \frac{1}{\pi_i \sigma^{d-2}} \exp\left(-\frac{(D-\delta_1)^2-\delta_2}{2\sigma^2}\right)\}$. On A, we have

$$\begin{aligned}
&\int_{y\in(S_i+B(0,\delta_1))\cap A} \|\varphi_{i,\alpha}(y) - \varphi_\alpha(y)\|^2 d\nu_i(y) \\
&\qquad = \int_{y\in(S_i+B(0,\delta_1))\cap A} \|\varphi_{i,\alpha}(y) - \varphi_\alpha(y)\|^2 p_{Y_i}(y) dy \\
&\qquad = \alpha^2 \int_{y\in(S_i+B(0,\delta_1))\cap A} \|\mathbb{E}[X_i|Y_i = y] + \mathbb{E}[X|Y = y]\|^2 d\nu_i(y) \\
&\qquad \leq \alpha^2 \int_{y\in(S_i+B(0,\delta_1))\cap A} \frac{4R^2}{\pi_i\sigma^{d-2}} \exp\left(\frac{(D-\delta_1)^2-\delta_2}{2\sigma^2}\right) dy \\
&\qquad \leq \frac{4\alpha^2R^2}{\pi_i\sigma^{d-2}} \exp\left(\frac{(D-\delta_1)^2-\delta_2}{2\sigma^2}\right) Vol(S_i + B(0,\delta_1)) \\
&\qquad \leq \frac{4\alpha^2R^2(R+\delta_1)^d}{\pi_i\sigma^{d-2}} \exp\left(\frac{(D-\delta_1)^2-\delta_2}{2\sigma^2}\right).
\end{aligned}$$

To bound the term on B, we first write $p_Y(y) = \sum_j \pi_j p_{Y_j}(y) = \pi_i p_{Y_i}(y) + f_i(y)$ with $f_i(y) = \sum_{j\neq i} \pi_j p_{Y_j}(y)$, and we notice that for $y \in S_i + B(0,\delta_1)$, we have:

$$\begin{aligned}
f_i(y) &= \sum_{j\neq i} \pi_j \int \frac{1}{C\sigma^d} \exp\left(-\frac{\|y-x\|^2}{2\sigma^2}\right) d\mu_j(x) \\
&\leq \frac{1}{C\sigma^d} \exp\left(\frac{-(D-\delta_1)^2}{2\sigma^2}\right),
\end{aligned}$$

with $C = (2\pi)^{d/2}$, as well as,

$$\begin{aligned}
\|\nabla f_i(y)\| &= \left\| \sum_{j\neq i} \pi_j \int \frac{y-x}{C\sigma^2} \exp\left(-\frac{\|y-x\|^2}{2\sigma^{d+2}}\right) d\mu_j(x) \right\| \\
&\leq \frac{2(R+\delta_1)}{C\sigma^{d+2}} \exp\left(\frac{-(D-\delta_1)^2}{2\sigma^2}\right),
\end{aligned}$$

and,

$$\|\nabla \log p_{Y_i}(y)\| = \left\| \frac{1}{\sigma^2}(\mathbb{E}[X_i|Y_i = y] - y) \right\| \leq \frac{2(R+\delta_1)}{\sigma^2}.$$

Therefore, for $y \in (S_i + B(0,\delta_1)) \cap B$:

$$\begin{aligned}
\|\nabla \log p_Y(y) - \nabla \log p_{Y_i}(y)\| &= \left\| \frac{\pi_i \nabla p_{Y_i}(y) + \nabla f_i(y)}{\pi_i p_{Y_i}(y) + f_i(y)} - \nabla \log p_{Y_i}(y) \right\| \\
&= \left\| \nabla \log p_{Y_i}(y) \left(\frac{1}{1 + \frac{f_i(y)}{\pi_i p_{Y_i}(y)}} - 1 \right) + \frac{\nabla f_i(y)}{\pi_i p_{Y_i}(y) + f_i(y)} \right\| \\
&\leq \frac{f_i(y)}{\pi_i p_{Y_i}(y)} \|\nabla \log p_{Y_i}(y)\| + \frac{\|\nabla f_i(y)\|}{\pi_i p_{Y_i}(y)} \\
&\quad (\text{for } t > 0, \left| \frac{1}{1+t} - 1 \right| = t \left| \frac{1}{1+t} \right| \leq t) \\
&\leq \frac{1}{C\sigma^2} \exp\left(\frac{-(D-\delta_1)^2}{2\sigma^2} + \frac{(D-\delta_1)^2 - \delta_2}{2\sigma^2} \right) \frac{2(R+\delta_1)}{\sigma^2} \\
&\quad + \frac{2(R+\delta_1)}{C\sigma^4} \exp\left(\frac{-(D-\delta_1)^2}{2\sigma^2} + \frac{(D-\delta_1)^2 - \delta_2}{2\sigma^2} \right) \\
&\leq \frac{4(R+\delta_1)}{C\sigma^4} \exp\left(-\frac{\delta_2}{2\sigma^2} \right).
\end{aligned}$$

This leads to:

$$\begin{aligned}
&\int_{y \in (S_i + B(0,\delta_1)) \cap B} \|\varphi_{i,\alpha}(y) - \varphi_\alpha(y)\|^2 d\nu_i(y) \\
&\qquad = \alpha^2 \sigma^4 \int_{y \in (S_i + B(0,\delta_1)) \cap B} \|\nabla \log p_Y(y) - \nabla \log p_{Y_i}(y)\|^2 d\nu_i(y) \\
&\qquad \leq \alpha^2 \sigma^4 \int_{y \in (S_i + B(0,\delta_1)) \cap B} \frac{16(R+\delta_1)^2}{C^2\sigma^8} \exp\left(-\frac{\delta_2}{\sigma^2} \right) d\nu_i(y) \\
&\qquad \leq \frac{16\alpha^2(R+\delta_1)^2}{C^2\sigma^8} \exp\left(-\frac{\delta_2}{\sigma^2} \right).
\end{aligned}$$

Finally, we have:

$$\begin{aligned}
&\int \|\varphi_{i,\alpha}(y) - \varphi_\alpha(y)\|^2 d\nu_i(y) \\
&\qquad \leq \frac{16R^2\alpha^2(1/2)^{d/2}\delta_1^{d-2}}{\Gamma(d/2)} \frac{1}{\sigma^{d-2}} \exp\left(-\frac{\delta_1^2}{2\sigma^2} \right) \\
&\qquad\quad + \frac{4\alpha^2 R^2 (R+\delta_1)^d}{\pi_i \sigma^{d-2}} \exp\left(\frac{(D-\delta_1)^2 - \delta_2}{2\sigma^2} \right) \\
&\qquad\quad + \frac{16\alpha^2(R+\delta_1)^2}{C^2\sigma^8} \exp\left(-\frac{\delta_2}{\sigma^2} \right) \\
&\qquad = O\left(\frac{1}{\sigma^{d-2}} \exp\left(-\frac{\delta_1^2}{2\sigma^2} \right) + \frac{1}{\sigma^{d-2}} \exp\left(-\frac{(D-\delta_1)^2 - \delta_2}{2\sigma^2} \right) + \frac{1}{\sigma^8} \exp\left(-\frac{\delta_2}{\sigma^2} \right) \right) \\
&\qquad = O\left(\frac{1}{\sigma^{\max(d-2,8)}} \exp\left(-\frac{K}{\sigma^2} \right) \right),
\end{aligned}$$

with $K = \min\left(\frac{\delta_1^2}{2}, \frac{(D-\delta_1)^2-\delta_2}{2}, \delta_2\right) > 0$. (For example take $\delta_1 = \frac{D}{2}$ and $\delta_2 = \frac{D^2}{8}$ to get $K = \frac{D^2}{16}$.) ■

Remark: we also have,

$$\begin{aligned} W_2^2(\mu, \mu_\alpha) - 2\sum_i \pi_i W_2^2(\mu_i, \mu_{i,\alpha}) \leq & \frac{32R^2\alpha^2(1/2)^{d/2}\delta_1^{d-2}}{\Gamma(d/2)} \frac{1}{\sigma^{d-2}} \exp\left(-\frac{\delta_1^2}{2\sigma^2}\right) \\ & + \frac{8N\alpha^2R^2(R+\delta_1)^d}{\sigma^{d-2}} \exp\left(\frac{(D-\delta_1)^2-\delta_2}{2\sigma^2}\right) \\ & + \frac{32\alpha^2(R+\delta_1)^2}{C^2\sigma^8} \exp\left(-\frac{\delta_2}{\sigma^2}\right) \\ = & O\left(\frac{1}{\sigma^{\max(d-2,8)}} \exp\left(-\frac{K}{\sigma^2}\right)\right). \end{aligned}$$

Appendix C. Usual distributions Verify the Hypothesis of Lemma 4

Proposition 11 *Assume that $X = Z + \varepsilon_0$, with $\mathbb{E}[\|Z\|^6] < \infty$, $Z \perp \varepsilon_0$ and $\varepsilon_0 \sim \mathcal{N}(0, \tau^2)$, then X verifies the assumptions of Lemma 4 with the constants:*

$$\begin{aligned} C_1 &= \mathbb{E}[\|\nabla \log p_X(X)\|^6] && \leq \tfrac{243}{\tau^{12}}(2\mathbb{E}[\|Z\|^6] + 15d\tau^6), \\ C_2 &= \mathbb{E}[\|\nabla^2 \log p_X(X)\|_{\text{op}}^3] && \leq 9\left(\tfrac{1}{\tau^6} + \tfrac{2\mathbb{E}[\|Z\|^6]}{\tau^{12}}\right), \\ C_3 &= \mathbb{E}[\|\nabla\Delta \log p_X(X)\|^2] && \leq \tfrac{40\mathbb{E}[\|Z\|^6]}{\tau^{12}}. \end{aligned}$$

Proof First note that $p_X : x \mapsto \frac{1}{(2\pi\tau^2)^{d/2}} \int e^{\frac{\|x-z\|^2}{2\tau^2}} d\mu(z) \in \mathcal{C}^\infty(\mathbb{R}^d)$. Then (B.1) and (B.3) of Saremi et al. (2023) give

$$\nabla \log p_X(x) = \frac{1}{\tau^2}(\mathbb{E}[Z|X = x] - x),$$

and

$$\begin{aligned} \nabla^2 \log p_X(x) &= -\frac{1}{\tau^2} I + \frac{1}{\tau^4}\left(\mathbb{E}[ZZ^\top|X = x] - \mathbb{E}[Z|X = x]\mathbb{E}[Z|X = x]^\top\right) \\ &= -\frac{1}{\tau^2} I + \frac{1}{\tau^4}\text{cov}(Z|X = x). \end{aligned}$$

Similarly, we can compute,

$$\begin{aligned} \nabla\Delta \log p_X(x) = & \frac{-1}{\tau^6}\mathbb{E}[\|Z\|^2 Z|X = x] \\ & + \frac{1}{\tau^6}\mathbb{E}[\|Z\|^2|X = x]\mathbb{E}[Z|X = x] \\ & - \frac{2}{\tau^6}\mathbb{E}[ZZ^\top|X = x] \cdot \mathbb{E}[Z|X = x] \\ & + \frac{2}{\tau^6}\|\mathbb{E}[Z|X = x]\|^2\mathbb{E}[Z|X = x]. \end{aligned}$$

We can now bound the constants C_1, C_2 and C_3. Using Jensen's inequality on $x \mapsto x^p$, that gives for all $x_1, \dots, x_k \in \mathbb{R}$,

$$\left(\sum_{i=1}^{k} x_i\right)^p \leq k^{p-1} \sum_{i=1}^{k} x_i^p,$$

and we have,

$$\begin{aligned}
C_1 = \mathbb{E}[\|\nabla \log p_X(X)\|^6] &= \mathbb{E}\left[\left\|\frac{1}{\tau^2}(\mathbb{E}[Z|X] - X)\right\|^6\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{\tau^2}(\mathbb{E}[Z|X] - Z - \varepsilon_0)\right\|^6\right] \\
&\leq \frac{243}{\tau^{12}}(\mathbb{E}[\|\mathbb{E}[Z|X]\|^6] + \mathbb{E}[\|Z\|^6] + \mathbb{E}[\|\varepsilon_0\|^6]) \\
&\leq \frac{243}{\tau^{12}}(\mathbb{E}[\mathbb{E}[\|Z\|^6|X]] + \mathbb{E}[\|Z\|^6] + 15d\tau^6) \\
&\quad \text{(Jensen's inequality on conditional expectation)} \\
&= \frac{243}{\tau^{12}}(2\mathbb{E}[\|Z\|^6] + 15d\tau^6),
\end{aligned}$$

and,

$$\begin{aligned}
C_2 &= \mathbb{E}[\|\nabla \log p_X(X)\|_{\text{op}}^3] \\
&= \mathbb{E}\left[\left\|-\frac{1}{\tau^2}I + \frac{1}{\tau^4}\left(\mathbb{E}[ZZ^\top|X] - \mathbb{E}[Z|X]\mathbb{E}[Z|X]^\top\right)\right\|_{\text{op}}^3\right] \\
&\leq 9\left(\frac{\|I\|_{\text{op}}^3}{\tau^6} + \frac{1}{\tau^{12}}\mathbb{E}[\|\mathbb{E}[ZZ^\top|X]\|_{\text{op}}^3] + \frac{1}{\tau^{12}}\mathbb{E}[\|\mathbb{E}[Z|X]\mathbb{E}[Z|X]^\top\|_{\text{op}}^3]\right) \\
&= 9\left(\frac{1}{\tau^6} + \frac{1}{\tau^{12}}\mathbb{E}[\|\mathbb{E}[ZZ^\top|X]\|_{\text{op}}^3] + \frac{1}{\tau^{12}}\mathbb{E}[\|\mathbb{E}[Z|X]\|^6]\right) \\
&\leq 9\left(\frac{1}{\tau^6} + \frac{1}{\tau^{12}}\mathbb{E}[\|ZZ^\top\|_{\text{op}}^3] + \frac{1}{\tau^{12}}\mathbb{E}[\|Z\|^6]\right) \\
&\quad \text{(Jensen's inequality on conditional expectation)} \\
&=\leq 9\left(\frac{1}{\tau^6} + \frac{2\mathbb{E}[\|Z\|^6]}{\tau^{12}}\right),
\end{aligned}$$

and finally we control the 4 terms in $\nabla\Delta \log p_X$,

$$\begin{aligned}
\mathbb{E}[\|\mathbb{E}[\|Z\|^2 Z|X]\|^2] &\leq \mathbb{E}[\|\|Z\|^2 Z\|^2] \\
&\quad \text{(Jensen's inequality on conditional expectation)} \\
&= \mathbb{E}[\|Z\|^6] \\
\mathbb{E}[\|\mathbb{E}[\|Z\|^2|X]\mathbb{E}[Z|X]\|^2] &= \mathbb{E}[(E[\|Z\|^2|X])^2\|\mathbb{E}[Z|X]\|^2] \\
&\leq \mathbb{E}[(E[\|Z\|^2|X])^2\mathbb{E}[\|Z\|^2|X]] = \mathbb{E}[(E[\|Z\|^2|X])^3] \\
&\quad \text{(Jensen's inequality on conditional expectation)} \\
&\leq \mathbb{E}[\|Z\|^6] \\
&\quad \text{(Jensen's inequality on conditional expectation)} \\
\mathbb{E}[\|\mathbb{E}[ZZ^\top|X] \cdot \mathbb{E}[Z|X]\|^2] &\leq \mathbb{E}[\|\mathbb{E}[ZZ^\top|X]\|_{\text{op}}^2\|\mathbb{E}[Z|X]\|^2] \\
&\leq \mathbb{E}[(\mathbb{E}[\|ZZ^\top\|_{\text{op}}|X])^2\mathbb{E}[\|Z\|^2|X]] \\
&= \mathbb{E}[(E[\|Z\|^2|X])^3] \\
&\leq \mathbb{E}[\|Z\|^6] \\
&\quad \text{(Jensen's inequality on conditional expectation)} \\
\mathbb{E}[\|\|\mathbb{E}[Z|X]\|^2\mathbb{E}[Z|X]\|^2] &= \mathbb{E}[\|\mathbb{E}[Z|X]\|^6] \\
&\leq \mathbb{E}[\|Z\|^6] \\
&\quad \text{(Jensen's inequality on conditional expectation),}
\end{aligned}$$

to get,

$$C_3 \leq \frac{40\mathbb{E}[\|Z\|^6]}{\tau^{12}}.$$

■

Appendix D. Extension of Results from Section 3.3 to any $p \geq 1$

We extend all results from section 3.3 to Wasserstein-p distances for any $p \geq 1$.

Proposition 12 *Assume that* $\mathbb{E}[\|\nabla \log p_X(X)\|^p] \leq C$. *Then*

$$W_p(\mathcal{L}(X), \mathcal{L}(\varphi_\alpha(Y))) \leq \left(\frac{(1+2^p\alpha^p)C}{2}\right)^{1/p} \sigma^2.$$

Proof We have

$$x_0 = x_t + \frac{1}{2}\int_0^t \nabla \log p_s(x_s) ds,$$

and

$$\hat{x}_0 = x_t + \alpha t \nabla \log p_t(x_t).$$

Therefore,

$$x_0 - \hat{x}_0 = \frac{1}{2}\int_0^t \nabla \log p_s(x_s) ds - \alpha t \nabla \log p_t(x_t).$$

Using Jensen's inequality two times,

$$\begin{aligned}\|x_0 - \hat{x}_0\|^p &\leq 2^{p-1}\left(\left\|\frac{1}{2}\int_0^t \nabla \log p_s(x_s)ds\right\|^p + \|\alpha t \nabla \log p_t(x_t)\|^p\right) \\ &\leq 2^{p-1}\left(\frac{t^{p-1}}{2^p}\int_0^t \|\nabla \log p_s(x_s)\|^p ds + \alpha^p t^p \|\nabla \log p_t(x_t)\|^p\right),\end{aligned}$$

and it follows that,

$$\begin{aligned}W_p^p(\mathcal{L}(X), \mathcal{L}(\varphi_\alpha(X_t))) &\leq \mathbb{E}[\|x_0 - x_t\|^p] \\ &\leq 2^{p-1}\left(\frac{t^{p-1}}{2^p}\int_0^t \mathbb{E}[\|\nabla \log p_s(x_s)\|^p]ds + \alpha^p t^p \mathbb{E}[\|\nabla \log p_t(x_t)\|^p]\right).\end{aligned}$$

As stated in Lemma 10, assuming $\mathbb{E}[\|\nabla \log p_X(X)\|^p] \leq C$ leads to $\mathbb{E}[\|\nabla \log p_s(x_s)\|^p] \leq C$ for all $s \geq 0$, so finally:

$$W_p^p(\mathcal{L}(X), \mathcal{L}(\varphi_\alpha(X_t))) \leq \frac{(1 + 2^p\alpha^p)C}{2}t^p,$$

which gives for $t = \sigma^2$:

$$W_p(\mathcal{L}(X), \mathcal{L}(\varphi_\alpha(Y))) \leq \left(\frac{(1 + 2^p\alpha^p)C}{2}\right)^{1/p}\sigma^2.$$

■

Lemma 13 *Assume that the variable $X \sim \mu$ of density p_X verifies that:*

- $\log p_X \in \mathcal{C}^3(\mathbb{R}^d)$.
- $C_1 = \mathbb{E}[\|\nabla \log p_X(X)\|^{3p}] < \infty$.
- $C_2 = \mathbb{E}[\|\nabla^2 \log p_X(X)\|_{\text{op}}^{3p/2}] < \infty$, *where $\|A\|_{\text{op}}$ is defined for any matrix A as $\|A\|_{\text{op}} = \sup_{x\neq 0}\frac{\|Ax\|}{\|x\|}$.*
- $C_3 = \mathbb{E}[\|\nabla \Delta \log p_X(X)\|^p] < \infty$.

Then,

$$\mathbb{E}\left[\left\|\frac{d}{dt}\nabla \log p_t(x_t)\right\|^p\right] \leq C,$$

with $C = \frac{9^{p-1}}{2^p}(3C_1 + 2(d^2+1)C_1^{1/3}C_2^{2/3} + C_3)$.

Proof Once again, we want to control all the terms (9-17), but this time in L_p norm. Recall that for all $x \in \mathbb{R}^d$, $|\Delta \log p_X(x)| \leq d\|\nabla^2 \log p_X(x)\|_{\text{op}}$, hence, $\mathbb{E}[|\Delta \log p_X(X)|^{3p/2}] \leq d^{3p/2}C_2$.

Term (9):

$$\begin{aligned}\mathbb{E}[\|\mathbb{E}[\nabla\Delta\log p_X(X)|Y]\|^p] &\leq \mathbb{E}[\mathbb{E}[\|\nabla\Delta\log p_X(X)\|^p|Y]] = \mathbb{E}[\|\nabla\Delta\log p_X(X)\|^p] \\ &\quad \text{(Jensen's inequality on the conditional expectation)} \\ &= C_3.\end{aligned}$$

Term (10):

$$\begin{aligned}\mathbb{E}[\|\mathbb{E}[\Delta\log p_X(X)\nabla\log p_X(X)|Y]\|^p] &\leq \mathbb{E}[|\Delta\log p_X(X)|^p\|\nabla\log p_X(X)\|^p] \\ &\quad \text{(Jensen's inequality on the conditional expectation)} \\ &\leq \mathbb{E}[|\Delta\log p_X(X)|^{3p/2}]^{2/3}\mathbb{E}[\|\nabla\log p_X(X)\|^{3p}]^{1/3} \\ &\quad \text{(Hölder's inequality)} \\ &\leq d^p C_2^{2/3} C_1^{1/3}.\end{aligned}$$

Term (11):

$$\begin{aligned}&\mathbb{E}[\|\mathbb{E}[\Delta\log p_X(X)|Y]\mathbb{E}[\nabla\log p_X(X)|Y]\|^p] \\ &\quad= \mathbb{E}[|\mathbb{E}[\Delta\log p_X(X)|Y]|^p\|\mathbb{E}[\nabla\log p_X(X)|Y]\|^p] \\ &\quad\leq \mathbb{E}[|\mathbb{E}[\Delta\log p_X(X)|Y]|^{3p/2}]^{2/3}\mathbb{E}[\|\mathbb{E}[\nabla\log p_X(X)|Y]\|^{3p}]^{1/3} \\ &\quad\quad \text{(Hölder's inequality)} \\ &\quad\leq \mathbb{E}[|\Delta\log p_X(X)|^{3p/2}]^{2/3}\mathbb{E}[\|\nabla\log p_X(X)\|^{3p}]^{1/3} \\ &\quad\quad \text{(Jensen's inequality on the conditional expectation)} \\ &\quad\leq d^p C_2^{2/3} C_1^{1/3}.\end{aligned}$$

Term (12):

$$\begin{aligned}&\mathbb{E}[\|\mathbb{E}[\nabla^2\log p_X(X)\cdot\nabla\log p_X(X)|Y]\|^p] \\ &\quad\leq \mathbb{E}[\|\nabla^2\log p_X(X)\cdot\nabla\log p_X(X)\|^p] \\ &\quad\quad \text{(Jensen's inequality on the conditional expectation)} \\ &\quad\leq \mathbb{E}[\|\nabla^2\log p_X(X)\|_{\text{op}}^p\|\nabla\log p_X(X)\|^p] \\ &\quad\leq \mathbb{E}[\|\nabla^2\log p_X(X)\|_{\text{op}}^{3p/2}]^{2/3}\mathbb{E}[\|\nabla\log p_X(X)\|^{3p}]^{1/3} \\ &\quad\quad \text{(Hölder's inequality)} \\ &\quad= C_2^{2/3} C_1^{1/3}.\end{aligned}$$

Term (13):

$$\begin{aligned}&\mathbb{E}[\|\mathbb{E}[\|\nabla\log p_X(X)\|^2\nabla\log p_X(X)|Y]\|^p] \\ &\quad\leq \mathbb{E}[\|\nabla\log p_X(X)\|^{3p}] \\ &\quad\quad \text{(Jensen's inequality on the conditional expectation)} \\ &\quad= C_1.\end{aligned}$$

Term (14):

$$\begin{aligned}
&\mathbb{E}[\|\mathbb{E}[\|\nabla \log p_X(X)\|^2|Y]\mathbb{E}[\nabla \log p_X(X)|Y]\|^p] \\
&\qquad = \mathbb{E}[(\mathbb{E}[\|\nabla \log p_X(X)\|^2|Y])^p\|\mathbb{E}[\nabla \log p_X(X)|Y]\|^p] \\
&\qquad \le \mathbb{E}[(\mathbb{E}[\|\nabla \log p_X(X)\|^2|Y])^p(\mathbb{E}[\|\nabla \log p_X(X)\|^2|Y])^{p/2}] \\
&\qquad\quad \text{(Jensen's inequality on the conditional expectation)} \\
&\qquad = \mathbb{E}[(\mathbb{E}[\|\nabla \log p_X(X)\|^2|Y])^{3p/2}] \\
&\qquad \le \mathbb{E}[\|\nabla \log p_X(X)\|^{3p}] \\
&\qquad\quad \text{(Jensen's inequality on the conditional expectation)} \\
&\qquad = C_1.
\end{aligned}$$

Term (15):

$$\begin{aligned}
&\mathbb{E}[\|\mathbb{E}[\nabla^2 \log p_X(X)|Y] \cdot \mathbb{E}[\nabla \log p_X(X)|Y]\|^p] \\
&\qquad \le \mathbb{E}[\|\mathbb{E}[\nabla^2 \log p_X(X)|Y]\|_{\mathrm{op}}^p\|\mathbb{E}[\nabla \log p_X(X)|Y]\|^p] \\
&\qquad \le \mathbb{E}[\|\mathbb{E}[\nabla^2 \log p_X(X)|Y]\|_{\mathrm{op}}^{3p/2}]^{2/3}\mathbb{E}\|\mathbb{E}[\nabla \log p_X(X)|Y]\|^{3p}]^{1/3} \\
&\qquad\quad \text{(Hölder's inequality)} \\
&\qquad \le \mathbb{E}[\|\nabla^2 \log p_X(X)\|_{\mathrm{op}}^{3p/2}]^{2/3}\mathbb{E}\|\nabla \log p_X(X)\|^{3p}]^{1/3} \\
&\qquad\quad \text{(Jensen's inequality on the conditional expectation)} \\
&\qquad = C_2^{2/3}C_1^{1/3}.
\end{aligned}$$

Term (16):

$$\begin{aligned}
&\mathbb{E}[\|\mathbb{E}[\nabla \log p_X(X)\nabla \log p_X(X)^\top|Y] \cdot \mathbb{E}[\nabla \log p_X(X)|Y]\|^p] \\
&\qquad \le \mathbb{E}[\|\mathbb{E}[\nabla \log p_X(X)\nabla \log p_X(X)^\top|Y]\|_{\mathrm{op}}^p\|\mathbb{E}[\nabla \log p_X(X)|Y]\|^p] \\
&\qquad \le \mathbb{E}[(\mathbb{E}[\|\nabla \log p_X(X)\nabla \log p_X(X)^\top\|_{\mathrm{op}}|Y])^p(\mathbb{E}[\|\nabla \log p_X(X)\|^2|Y])^{p/2}] \\
&\qquad\quad \text{(Jensen's inequality on the conditional expectation)} \\
&\qquad = \mathbb{E}[(\mathbb{E}[\|\nabla \log p_X(X)\|^2|Y])^p\|\mathbb{E}[\nabla \log p_X(X)|Y]\|^p] \\
&\qquad = \mathbb{E}[(\mathbb{E}[\|\nabla \log p_X(X)\|^2|Y])^{3p/2}] \\
&\qquad \le \mathbb{E}[\|\nabla \log p_X(X)\|^{3p}] \\
&\qquad\quad \text{(Jensen's inequality on the conditional expectation)} \\
&\qquad = C_1.
\end{aligned}$$

Term (17):

$$\begin{aligned}
&\mathbb{E}[\|\|\mathbb{E}[\nabla \log p_X(X)|Y]\|^2\mathbb{E}[\nabla \log p_X(X)|Y]\|^p] \\
&\qquad = \mathbb{E}[\|\mathbb{E}[\nabla \log p_X(X)|Y]\|^{3p}] \\
&\qquad \le \mathbb{E}[\|\nabla \log p_X(X)\|^{3p}] \\
&\qquad\quad \text{(Jensen's inequality on the conditional expectation)} \\
&\qquad = C_1.
\end{aligned}$$

To combine this bounds, we use Jensen's inequality on $x \mapsto x^p$, that gives for all $x_1, \dots, x_k \in \mathbb{R}$,

$$\left(\sum_{i=1}^{k} x_i\right)^p \leq k^{p-1} \sum_{i=1}^{k} x_i^p,$$

and we finally have,

$$\mathbb{E}\left[\left\|\frac{d}{dt}\nabla \log p_t(x_t)\right\|^2\right] \leq \frac{9^{p-1}}{2^p}(4C_1 + (2d^2+5)C_1^{1/3}C_2^{2/3} + C_3) = C.$$

■

Proposition 14 *Under the same assumptions of Lemma 4, we have*

$$W_2(\mathcal{L}(X), \mathcal{L}(\varphi_{1/2}(Y))) \leq K\sigma^4,$$

with $K = \frac{9^{(p-1)/p}}{4(p+1)^{1/p}}(4C_1 + (2d^2+5)C_1^{1/3}C_2^{2/3} + C_3)^{1/p}$.

Proof We have,

$$\begin{aligned} x_0 - \hat{x}_0 &= \frac{1}{2}\int_0^t \nabla \log p_s(x_s) ds - \frac{1}{2} t \nabla \log p_t(x_t) \\ &= \frac{1}{2}\int_0^t \left(\nabla \log p_s(x_s) - \nabla \log p_t(x_t)\right) ds \\ &= \frac{1}{2}\int_0^t \int_s^t \frac{d}{du}\left(\nabla \log p_u(x_u)\right) du ds. \end{aligned}$$

hence, with Jensen's inequality,

$$\mathbb{E}[\|x_0 - x_t\|^p] \leq \frac{t^{p-1}}{2^p}\int_0^t (t-s)^{p-1} \int_s^t \mathbb{E}\left[\left\|\frac{d}{du}\left(\nabla \log p_u(x_u)\right)\right\|^p\right] du ds,$$

Using Lemma 4, it leads to $\mathbb{E}[\|x_0 - x_t\|^2] \leq \frac{C}{2^p(p+1)} t^{2p}$. In particular for $t = \sigma^2$, we get:

$$W_p(\mathcal{L}(X), \mathcal{L}(\varphi_{1/2}(Y))) \leq (\mathbb{E}[\|x_0 - x_{\sigma^2}\|^p])^{1/p} \leq K\sigma^4,$$

with $K = \left(\frac{C}{2^p(p+1)}\right)^{1/p} = \frac{9^{(p-1)/p}}{4(p+1)^{1/p}}(4C_1 + (2d^2+5)C_1^{1/3}C_2^{2/3} + C_3)^{1/p}$. ■

Remark: Proposition 7 from section 5 can also be extended for any $p \geq 1$ (tough it will make the proof harder to read). However Proposition 6 from section 4 relies on the fact that $\|(x_1, x_2)\|^2 = \|x_1\|^2 + \|x_2\|^2$ for $\|\cdot\|$ the Euclidean norm on $\mathbb{R}^d$. Therefore it could only by extended to any $p \geq 1$ if we use the norm $\|\cdot\|_p$ on $\mathbb{R}^d$.

Appendix E. Computation of Wasserstein Distance for a Mixture of Two Dirac Distributions

We derive here integral expressions for $W_2(\mathcal{L}(X), \mathcal{L}(\varphi_\alpha(Y)))$ when $X = \frac{\delta_{-\mu}+\delta_\mu}{2}$ on $\mathbb{R}$ for some $\mu > 0$. In this case, we have

$$p_Y(y) = \frac{1}{2\sqrt{2\pi\sigma^2}}\left(e^{-\frac{(y-\mu)^2}{2\sigma^2}} + e^{-\frac{(y+\mu)^2}{2\sigma^2}}\right)$$

leading to,

$$\nabla \log p_Y(y) = \frac{1}{\sigma^2}(-y + \mu \tanh\left(\frac{y\mu}{\sigma^2}\right)),$$

hence

$$\varphi_\alpha(y) = \alpha\mu \tanh\left(\frac{y\mu}{\sigma^2}\right) + (1-\alpha)y$$

The function φ_α is increasing and verifies $\forall y \in \mathbb{R}, \varphi_\alpha(-y) = -\varphi_\alpha(y)$. Moreover, as X is a symmetric random variable ($-X \sim X$), Y is also symmetric, hence $\varphi_\alpha(Y)$ as well. Computing the optimal transport plan is therefore straightforward, as all the mass of $\varphi_\alpha(Y)$ on $\mathbb{R}_-$ should be transported to $-\mu$ and all the mass of $\varphi_\alpha(Y)$ on $\mathbb{R}_+$ should be transported to μ, each part having the same transport cost, leading to the expression

$$\begin{aligned}
W_2^2(\mathcal{L}(X), \mathcal{L}(\varphi_\alpha(Y))) &= 2\mathbb{E}[(\varphi_\alpha(Y)-\mu)^2 \mathbb{1}_{\varphi_\alpha(Y)\geq 0}] \\
&= 2 \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^\infty (\varphi_\alpha(y)-\mu)^2 \frac{e^{-\frac{(y-\mu)^2}{2\sigma^2}} + e^{-\frac{(y+\mu)^2}{2\sigma^2}}}{2} dy \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^\infty (\varphi_\alpha(y)-\mu)^2 e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy \\
&\quad + \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^\infty (\varphi_\alpha(y)-\mu)^2 e^{-\frac{(y+\mu)^2}{2\sigma^2}} dy \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^\infty (\varphi_\alpha(y)-\mu)^2 e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy \\
\text{(change } y \leftarrow -y) \quad &+ \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^0 (\varphi_\alpha(-y)-\mu)^2 e^{-\frac{(-y+\mu)^2}{2\sigma^2}} dy \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^\infty (\varphi_\alpha(|y|)-\mu)^2 e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy
\end{aligned}$$

We can rearrange this expression to write $\frac{W_2(\mathcal{L}(X),\mathcal{L}(\varphi_\alpha(Y)))}{\mu}$ as a function of $\frac{\sigma}{\mu}$. We first define, for $u \in \mathbb{R}$,

$$\varphi_\alpha^\sigma(u) = \alpha \tanh\left(\frac{u}{\sigma^2}\right) + (1-\alpha)u,$$

then taking the square root, dividing by μ and making the change of variable $u = \frac{y}{\mu}$, we get

$$\frac{W_2(\mathcal{L}(X), \mathcal{L}(\varphi_\alpha(Y)))}{\mu} = \left(\frac{1}{\sqrt{2\pi(\sigma/\mu)^2}} \int_{-\infty}^\infty (\varphi_\alpha^{\sigma/\mu}(|u|)-1)^2 e^{-\frac{(y-1)^2}{2(\sigma/\mu)^2}} dy\right)^{1/2}.$$

Appendix F. Derivation of Deterministic Diffusion-Model Samplers

First note that the equation for the probability flow ODE given by Karras et al. (2022) is (with our notations):

$$\frac{dx_t}{dt} = \frac{\dot{s}(t)}{s(t)}x_t - s(t)^2\dot{\sigma}(t)\sigma(t)\nabla_x\left[(x,\sigma^2)\mapsto \log p\left(\frac{x}{s(t)};\sigma^2\right)\right]\left(x_t;\sigma(t)^2\right),$$

which can by rewritten

$$\frac{dx_t}{dt} = \frac{\dot{s}(t)}{s(t)}x_t - s(t)\dot{\sigma}(t)\sigma(t)\nabla_x\left[(x,\sigma^2)\mapsto \log p\left(x;\sigma^2\right)\right]\left(\frac{x_t}{s(t)};\sigma(t)^2\right).$$

Writing $\nabla \log p = \nabla_x\left[(x,\sigma^2)\mapsto \log p\left(x;\sigma^2\right)\right]$ the score function for the density $p\left(x;\sigma^2\right)$ of the normalized variable at noise level σ^2, we finally get (5):

$$\frac{dx_t}{dt} = \frac{\dot{s}(t)}{s(t)}x_t - s(t)\dot{\sigma}(t)\sigma(t)\nabla \log p\left(\frac{x_t}{s(t)};\sigma(t)^2\right). \tag{5}$$

F.1 DDIM Algorithm

There are different ways to derive the DDIM updates (Song et al., 2021a). Firstly, one can consider a specific process (X_t) with the same marginal as (x_t) defined by (5). Consider $Z\sim\mathcal{N}(0,I)$ independent from X and define, for $t\geq 0$,

$$X_t = s(t)(X+\sigma(t)Z)$$

For $t_k, t_{k+1}\in\mathbb{R}$, we can rewrite

$$\begin{aligned}
X_{t_{k+1}} &= s(t_{k+1})(X+\sigma(t_{k+1})Z)\\
&= s(t_{k+1})\left(X+\frac{\sigma(t_{k+1})}{\sigma(t_k)}\left(\frac{X_{t_k}}{s(t_k)}-X\right)\right)\\
&= s(t_{k+1})\left(1-\frac{\sigma(t_{k+1})}{\sigma(t_k)}\right)X + \frac{\sigma(t_{k+1})s(t_{k+1})}{\sigma(t_k)s(t_k)}X_{t_k}
\end{aligned}$$

from which we deduce,

$$\mathbb{E}[X_{t_{k+1}}|X_{t_k}] = s(t_{k+1})\left(1-\frac{\sigma(t_{k+1})}{\sigma(t_k)}\right)\mathbb{E}[X|X_{t_k}] + \frac{\sigma(t_{k+1})s(t_{k+1})}{\sigma(t_k)s(t_k)}X_{t_k}.$$

As $X_{t_k}/s(tk) = X+\sigma(t_k)Z$, Tweedie's formula gives

$$\begin{aligned}
\mathbb{E}[X|X_{t_k}] = \mathbb{E}\left[X\middle|\frac{X_{t_k}}{s(t_k)}\right] &= \frac{X_{t_k}}{s(t_k)} + \sigma(t_k)^2\nabla \log p_{\frac{X_{t_k}}{s(t_k)}}\left(\frac{X_{t_k}}{s(t_k)}\right)\\
&= \frac{X_{t_k}}{s(t_k)} + \sigma(t_k)^2\nabla \log p\left(\frac{X_{t_k}}{s(t_k)};\sigma(t_k)^2\right),
\end{aligned}$$

leading to,

$$\mathbb{E}[X_{t_{k+1}}|X_{t_k}] = \frac{s(t_{k+1})}{s(t_k)}X_{t_k} + s(t_{k+1})(\sigma(t_k)^2-\sigma(t_k)\sigma(t_{k+1}))\nabla \log p\left(\frac{X_{t_k}}{s(t_k)};\sigma(t_k)^2\right).$$

The DDIM update comes from approximating $X_{t_{k+1}}$ by $\mathbb{E}[X_{t_{k+1}}|X_{t_k}]$, leading to

$$\hat{X}_{k+1} = \frac{s(t_{k+1})}{s(t_k)}\hat{X}_k + s(t_{k+1})(\sigma(t_k)^2 - \sigma(t_k)\sigma(t_{k+1}))\nabla \log p\left(\frac{\hat{X}_k}{s(t_k)}; \sigma(t_k)^2\right).$$

Another method to derive the same update is to start from (5), and take an Euler discretization of (5) with the approximations $\dot{s}(t) \approx (s(t_{k+1}) - s(t_k))/(t_{k+1} - t_k)$, $\dot{\sigma}(t) \approx (\sigma(t_{k+1}) - \sigma(t_k))/(t_{k+1} - t_k)$, $s(t) \approx s(t_k)$ for the first occurrence of $s(t)$ in (5) and $s(t) \approx s(t_{k+1})$ for the second, leading to

$$\frac{\hat{X}_{k+1} - \hat{X}_k}{t_{k+1} - t_k} = \frac{1}{s(t_k)}\frac{s(t_{k+1}) - s(t_k)}{t_{k+1} - t_k}\hat{X}_k - s(t_{k+1})\frac{\sigma(t_{k+1}) - \sigma(t_k)}{t_{k+1} - t_k}\sigma(t_k)\nabla \log p\left(\frac{\hat{X}_k}{s(t_k)}; \sigma(t_k)^2\right),$$

that can be rewritten as

$$\hat{X}_{k+1} = \frac{s(t_{k+1})}{s(t_k)}\hat{X}_k + s(t_{k+1})(\sigma(t_k)^2 - \sigma(t_k)\sigma(t_{k+1}))\nabla \log p\left(\frac{\hat{X}_k}{s(t_k)}; \sigma(t_k)^2\right).$$

The procedure proposed by Song et al. (2021a) is recovered by taking $s(t) = \sqrt{\alpha_t}$ and $\sigma(t) = \frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha_t}}$.

F.2 Euler Discretization

The Euler discretization of (5) is

$$\frac{\hat{X}_{k+1} - \hat{X}_k}{t_{k+1} - t_k} = \frac{\dot{s}(t_k)}{s(t_k)}\hat{X}_k - s(t_k)\dot{\sigma}(t_k)\sigma(t_k)\nabla \log p\left(\frac{\hat{X}_k}{s(t_k)}; \sigma(t_k)^2\right),$$

which can by rewritten

$$\hat{X}_{k+1} = \left(1 + \frac{\dot{s}(t_k)(t_{k+1} - t_k)}{s(t_k)}\right)\hat{X}_k - s(t_k)(t_{k+1} - t_k)\dot{\sigma}(t_k)\sigma(t_k)\nabla \log p\left(\frac{\hat{X}_k}{s(t_k)}; \sigma(t_k)^2\right).$$

References

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions, November 2023. arXiv:2303.08797.

Iskander Azangulov, George Deligiannidis, and Judith Rousseau. Convergence of diffusion models under the manifold hypothesis in high-dimensions, April 2025. arXiv:2409.18804.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, August 2013.

Eliot Beyler and Francis Bach. Convergence of deterministic and stochastic diffusion-model samplers: A simple analysis in Wasserstein distance, November 2025. arXiv:2508.03210.

Vladimir Bogachev, Nicolai Krylov, Michael Röckner, and Stanislav Shaposhnikov. *Fokker–Planck–Kolmogorov Equations*, volume 207 of *Mathematical Surveys and Monographs*. American Mathematical Society, December 2015.

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

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: Theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*, 2023.

Giovanni Conforti, Alain Durmus, and Marta Gentiloni-Silveri. KL convergence guarantees for score diffusion models under minimal data assumptions. *SIAM Journal on Mathematics of Data Science*, pages 86–109, March 2025.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, December 2011.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, October 2016.

Marta Gentiloni-Silveri and Antonio Ocello. Beyond log-concavity and score regularity: Improved convergence bounds for score-based generative models in W2-distance. In *International Conference on Machine Learning*, 2025.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, March 2012.

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

Aapo Hyvärinen. A noise-corrected Langevin algorithm and sampling by half-denoising. *Transactions on Machine Learning Research*, 2025.

Tero Karras, Timo Aila, Miika Aittala, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022.

Jean-François Le Gall. *Brownian Motion, Martingales, and Stochastic Calculus*, volume 274 of *Graduate Texts in Mathematics*. Springer International Publishing, Cham, 2016.

Yaron Lipman, Ricky T Q Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Koichi Miyasawa. An empirical Bayes estimator of the mean of a normal population. *Bulletin of the International Statistical Institute*, 38(4):181–188, 1961.

Andrea Montanari. Sampling, diffusions, and stochastic localization, May 2023. arXiv:2305.10690.

Preetum Nakkiran, Arwen Bradley, Hattie Zhou, and Madhu Advani. Step-by-step diffusion: An elementary tutorial, June 2024. arXiv:2406.08929.

Frédéric Paulin. *Topologie, Analyse et Calcul Différentiel*. FIMFA. 2009.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, February 2019.

Hannes Risken. *The Fokker-Planck Equation: Methods of Solution and Applications*. Number 18 in Springer Series in Synergetics. Springer, second edition, 1996.

Herbert Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 157–163, 1956.

Saeed Saremi and Aapo Hyvärinen. Neural empirical Bayes. *Journal of Machine Learning Research*, 20(181):1–23, 2019.

Saeed Saremi, Ji Won Park, and Francis Bach. Chain of log-concave Markov chains. In *International Conference on Learning Representations*, October 2023.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, June 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

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

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Scholkopf, and Gert R G Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

Stanislas Strasman, Antonio Ocello, Claire Boyer, Sylvain Le Corff, and Vincent Lemaire. An analysis of the noise schedule for score-based generative models. *Transactions on Machine Learning Research*, December 2024.

Rong Tang and Yun Yang. Adaptivity of diffusion models to manifold structures. In *International Conference on Artificial Intelligence and Statistics*, April 2024.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, December 2000.

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